Enhancing Adversarial Transferability with Adversarial Weight Tuning

Jiahao Chen[1*], **Zhou Feng**[1*], **Rui Zeng**[1], **Yuwen Pu**[1], **Chunyi Zhou**[1†],
Yi Jiang[1], **Yuyou Gan**[1], **Jinbao Li**[23] **and Shouling Ji**[1]

[1] College of Computer Science and Technology, Zhejiang University
[2] Shandong Artificial Intelligence Institute
[3] School of Mathematics and Statistics, Qilu University of Technology

Abstract

Deep neural networks (DNNs) are vulnerable to adversarial examples (AEs) that mislead the model while appearing benign to human observers. A critical concern is the transferability of AEs, which enables black-box attacks without direct access to the target model. However, many previous attacks have failed to explain the intrinsic mechanism of adversarial transferability. In this paper, we rethink the property of transferable AEs and reformalize the formulation of transferability. Building on insights from this mechanism, we analyze the generalization of AEs across models with different architectures and prove that we can find a local perturbation to mitigate the gap between surrogate and target models. We further establish the inner connections between model smoothness and flat local maxima, both of which contribute to the transferability of AEs. Further, we propose a new adversarial attack algorithm, Adversarial Weight Tuning (AWT), which adaptively adjusts the parameters of the surrogate model using generated AEs to optimize the flat local maxima and model smoothness simultaneously, without the need for extra data. AWT is a data-free tuning method that combines gradient-based and model-based attack methods to enhance the transferability of AEs. Extensive experiments on a variety of models with different architectures on ImageNet demonstrate that AWT yields superior performance over other attacks, with an average increase of nearly 5% and 10% attack success rates on CNN-based and Transformer-based models, respectively, compared to state-of-the-art attacks.

Introduction

In recent years, deep neural networks (DNNs) have achieved remarkable success across a wide range of applications, from computer vision (He et al. 2016) and natural language processing (Radford et al. 2019) to speech recognition (Povey et al. 2011) and autonomous driving (Caesar et al. 2020). The proliferation of large-scale models (Rombach et al. 2022), empowered by advancements in computational resources and the availability of vast datasets, has further accelerated the adoption of DNNs in real-world scenarios. Despite these achievements, DNNs have been shown to be vulnerable to adversarial examples (AEs) (Goodfellow,

---

*These authors contributed equally.
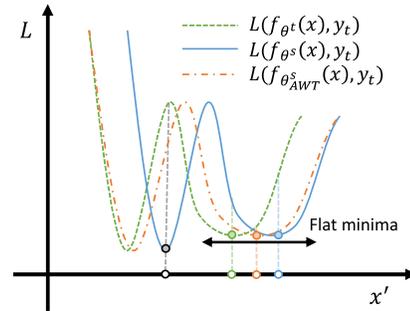†Corresponding author.

Figure 1: The motivation of the proposed AWT that achieves flat local minima via adjusting the surrogate model's parameters θ^s to enable more transferable AEs. This plot illustrates that we can achieve low $\nabla_{x'}\ell(x', y; \theta^s)$ through the input and parameter space. The colored dots represent the generated AEs with different loss value on the target model θ^t.

Shlens, and Szegedy 2014; Madry et al. 2017), which are intentionally crafted to mislead the model while appearing benign to human observers.

From a practical perspective, one particularly concerning aspect of AEs is their ability to transfer between different models (Ma et al. 2023; Zhang et al. 2023), allowing attackers to craft effective perturbations without direct access to the target model's architecture or parameters (Dong et al. 2018; Wang et al. 2021c; Wang and He 2021). Recently, researchers have proposed a variety of transferable attack methods aiming to maximize the success rate of these attacks and avoid over-fitting to the surrogate models (Dong et al. 2018, 2019). These methods leverage different strategies, such as input transformation by applying data augmentation (Xie et al. 2019), model ensemble that exploits the outputs of multiple models in loss or logits (Liu et al. 2016) and gradient rectification with momentum (Dong et al. 2018; Lin et al. 2019), to generate transferable AEs across diverse architectures of the surrogate model. Although these methods have presented better performance, there are no clear explanations for the factors behind transferability.

To explain what factors have resulted in the different transferability of different models, Wu et al. (Wu et al. 2018) explored how adversarial transferability relies on the

model-specific factors empirically and found that local non-smoothness of the loss surface harms the transferability of generated AEs. Next, Yang et al. (Yang et al. 2021) first theoretically linked two key factors, i.e., model smoothness and gradient similarity, with the lower bound of AEs transferability. With this intuition, researchers have proposed to finetune the attacker's surrogate model to enhance its smoothness (Zhu et al. 2022; Wu, Luo, and Wunsch II 2024; Zhang et al. 2024), since AEs generated on smoothed surrogate models have shown better transferability. Nevertheless, these methods all assume that attackers can obtain extra data with a similar distribution of the target models' training data, and time-consuming tuning on these large-scale datasets introduces computation overhead as well. Different from these model-based attacks, recent works (Ge et al. 2023; Qiu et al. 2024) can achieve state-of-the-art (SOTA) performance on transferability by optimizing the flat local maxima (minima) of AEs, as illustrated (from black to blue dot) in Figure 1.

While these methods have made significant strides in improving the transferability, **limitaions** remain. (1) Many of the existing methods rely on empirical evaluations with different models, lacking a unified and representative metric to measure transferability across various architectures. (2) Moreover, model-based methods require access to extra data (the same distribution as that of the target model), which may not always be feasible in real-world scenarios. (3) Model-based methods only optimize smoother landscapes on benign samples, leading to suboptimal performance since AEs are often generated iteratively. (4) Both gradient and model-based methods fail to explain why flat local maxima or model smoothness boosts the generalization of AEs across different models. Therefore, it is necessary to formalize the definition of transferable AEs and find the connection with the aforementioned factors, which will offer more insights into understanding adversarial transferability.

To this end, we rethink the property of transferable AEs and develop a novel formulation from the perspective of generalization to measure transferability. Building upon the insights from this formulation, we first analyze the generalization of AEs across models with different architectures and prove that we can always find a local perturbation to mitigate the gap between surrogate and target models. Further, with these insights, we observe inner connections between model smoothness and flat local maxima, which both contribute to the transferability of AEs. Based on this, we then propose a new adversarial attack algorithm, **A**dversarial **W**eight **T**uning (AWT), which adaptively adjusts the parameters of the surrogate model using generated AEs to optimize the flat local maxima and model smoothness simultaneously, without the need for extra data. As shown in Figure 1, better transferability (orange dot) can be achieved through input and parameter space. We evaluate AWT on a variety of models with traditional experiments and the metric derived from our proposed formulation. Our contributions are as follows:

- We theoretically analyze the transferability of AEs and reveal the inherent factors of the surrogate model and AEs that account for the transferability of AEs, which have been neglected by previous works.

- Based on the analysis above, a data-free tuning method AWT is developed by adaptively adjusting the parameters of the surrogate model while optimizing adversarial perturbations with better transferability. To the best of our knowledge, AWT is the first method that combines gradient-based and model-based attack methods to enhance the transferability of AEs.

- Further, with our formulation as a measurement of adversarial transferability. We conduct comprehensive experiments for evaluation with CNN-based and Tranformer-based models on ImageNet and demonstrate that, AWT yields superior attack success rates and metric values over eight SOTA transferable attacks.

Related Work

Gradient-based Adversarial Attack

Gradient-based methods are among the most common approaches to generating AEs and improving their transferability. These methods basically involve crafting perturbations based on the gradients of the loss function with respect to the input. One of the earliest and most influential gradient-based methods is the Fast Gradient Sign Method (FGSM) (Goodfellow, Shlens, and Szegedy 2014), which adds a small perturbation to the input in the direction of the sign of the gradient. Subsequent improvements, such as the Iterative FGSM (Kurakin, Goodfellow, and Bengio 2016) and the Momentum-based FGSM (Dong et al. 2018; Lin et al. 2019), iteratively refine the perturbation to achieve higher success rates and better transferability. Later, researchers proposed methods to improve the transferability of AEs by leveraging additional gradient information. For example, the Virtual Momentum Iterative FGSM (VMI) (Wang and He 2021) and the Ensemble Momentum Iterative FGSM (EMI) (Wang et al. 2021c) use virtual and ensemble gradients, respectively, to enhance the transferability of AEs. More recently, the connection between flat local maxima and generalization (Keskar et al. 2016; Foret et al. 2020) have inspired researchers to generate AEs at a flat local region of the loss function, making them more likely to generalize across models. For instance, Qin et al. (Qin et al. 2022) seek to minimize the loss of AEs at a region with low loss value by adding reverse perturbations before calculating gradients. Ge et al. and Qiu et al. (Ge et al. 2023; Qiu et al. 2024) introduce approximated Hessian-vector products and aggregate the gradients of randomly sampled data around AEs.

Model-based Adversarial Attack

There are also many works focusing on enhancing the smoothness of the surrogate model since Wu et al. (Wu et al. 2018) found that local non-smoothness of the loss surface harms the transferability of AEs, leading to methods that finetune the surrogate model to be smoother (Zhu et al. 2022; Wu, Luo, and Wunsch II 2024). These model-based methods generate transferable AEs by modifying the surrogate model itself, which often involve pre-processing steps that modify the model's architecture (Li et al. 2020) or training loss function. For example, Distribution-Relevant Attack

(DRA) (Zhu et al. 2022) achieves smoother by minimizing the Hessian-vector products of the surrogate model.

Others

In addition to gradient-based and model-based methods, researchers have explored other techniques to improve the transferability of AEs (Wang et al. 2022, 2021a, 2024a,b). One approach is to apply input transformations, such as data augmentation, to the surrogate model during training. Techniques like Diverse Input (DIM) (Xie et al. 2019) and Translation-invariant method (TIM) (Dong et al. 2019) have been shown to improve the transferability of AEs by making the AEs more robust to data transformations. Another line of research focuses on optimizing AEs with the aggregated gradients from the ensemble of surrogate models (Liu et al. 2016; Gubri et al. 2022). With augmentation of surrogate models, these methods aim to create AEs that are able to fool a wider range of target models with significantly different architectures.

Methodology

Preliminaries

Formally, take an image classification task as an example, let $\mathcal{F} : \mathcal{X} \to \mathcal{Y}$ denotes the set of mapping function with parameters, where input space $\mathcal{X} = \mathbb{R}^{B \times H \times W \times C}$ and output space $\mathcal{Y} = \mathbb{R}^{B \times K}$ (B and K means batch size and the number of the classes, respectively). For a dataset $\mathcal{D} = \{(x_1, y_1), (x_2, y_2), ...(x_i, y_i), ...(x_N, y_N)\}$, where $x_i \in \mathcal{X}$ and $y_i \in \mathcal{Y}$. Under white-box settings, attackers can generate AEs x' by solving the maximization problem with f_θ:

$$\max_{x'} \ell(x', y; \theta), \quad \textbf{s.t.} \|x - x'\|_p \le \epsilon, \tag{1}$$

where $f_\theta \in \mathcal{F}$, $\ell(\cdot, \cdot; \cdot)$ is the loss function, and $\|x - x'\|_p \le \epsilon$ restricts the L_p norm of the discrepancy between x and x', where we adopt $p = \infty$ in this paper.

In black-box scenarios, different from Eq.(1), the attacker aims to generate AEs with high generalization (transferability) across models with different architectures and parameters. To achieve this, the attacker needs to solve the following problem:

$$\max_{x'} \mathbb{E}_{f_{\theta_i} \in \mathcal{F}} \left[\ell(x', y; \theta_i) \right], \tag{2}$$

where f_{θ_i} denotes the target models trained on $\mathcal{D}$ from $\mathcal{F}$. Since the architecture and parameters of the target model θ^t are out of reach, x' can only generated on surrogate model θ^s held by attackers. Given that the optimization problem in Eq.(2) involves unknown models, we first simplify this by assuming that our target models are trained on $\mathcal{D}$ with the same architecture but different parameters. Therefore, we can simplify Eq.(2) as follows:

$$\max_{x'} \mathbb{E}_{\eta_i \in \Omega} \left[\ell(x', y; \theta^s \oplus \eta_i) \right], \tag{3}$$

where $\theta^s \oplus \eta_i$ means that we can add different perturbation η on parameter space θ^s to stand for different target models with the same architecture but different parameters. Ω represents the set of parameter perturbations.

Enhance Transferability with Local (Min)Maxima

With the preliminaries above, we aim to find the inner connection between transferability and corresponding factors like flat local maxima and model smoothness in this subsection. Based this connection, we then solve the problem in Eq.(3) with local maxima (minima) from the perspective of both input and parameter space to achieve transferability.

Achieving Flat Local Maxima from Input Space. From the perspective of generalization, to achieve Eq.(2), recent works (Ge et al. 2023; Qiu et al. 2024) proposed to find a flat local maxima of AEs to achieve better transferability:

$$\max_{x'} \ell(x', y; \theta^s) - \lambda \|\nabla_{x'} \ell(x', y; \theta^s)\|_2. \tag{4}$$

However, the mechanism behind this factor remains to be discussed and explored. Here, we present a theoretical analysis on this factor.

Proposition 1: *Given a model f_θ with parameters θ, and a perturbation η such that $\|\eta\|_p \le \kappa$ ($\kappa \to 0$), we want to prove that $\forall x \in \mathcal{D}$ and $\gamma \to 0$, there exists an perturbation δ such that:*

$$\|f_{\theta+\eta}(x) - f_\theta(x + \delta)\|_p \le \gamma. \tag{5}$$

Proposition 1 states that models with different parameters can be approximated by introducing perturbations in the input space, which suggests that we can bridge the gap between target and surrogate models using input perturbations. Therefore, we can further approximate Eq.(3) as follows:

$$\max_{x'} \mathbb{E}_{\delta_i \in \hat{\Omega}} \left[\ell(x' + \delta_i, y; \theta^s) \right], \tag{6}$$

where $\delta_i \in \hat{\Omega} = \{\delta \,|\, \|f_{\theta+\eta}(x) - f_\theta(x + \delta)\|_p \le \gamma, \eta \in \Omega\}$ represents the perturbation set in the input space to eliminate the differences in parameter space. For the generation of AEs, we can transform Eq.(6) into a bi-level optimization problem by maximizing the lower bound of the expectation item in Eq.(6):

$$\max_{x'} \min_{\delta} \left[\ell(x' + \delta, y; \theta^s) \right]. \tag{7}$$

Subsequently, we further approximate the first-order Taylor expansion of the loss function $\ell(\cdot)$ at the point x':

$$\ell(x' + \delta, y; \theta^s) = \ell(x', y; \theta^s) + \delta^T \cdot \nabla_{x'} \ell(x', y; \theta^s) + \mathcal{O}(\|\delta\|_2^2). \tag{8}$$

Please note that solving the inner optimization of Eq.(7) can be approximated by $-\alpha \nabla_{x'} \ell(x', y; \theta^s)$. This approximation leads to an equation nearly equivalent to Eq.(4), which have actually been proven in NCS (Qiu et al. 2024). Consequently, we can relate the flat local maxima $\|\nabla_{x'} \ell(x', y; \theta^s)\|_2$ to the transferability of AEs across models with the same architecture but different parameters, as outlined in Eq.(6).

However, SOTA attacks such as NCS (Qiu et al. 2024) and PGN (Ge et al. 2023) achieve flat local maxima of x' solely through optimization in the input space. In contrast, we propose achieving this by adjusting the parameters of the surrogate models. Unlike model-based tuning methods, AWT
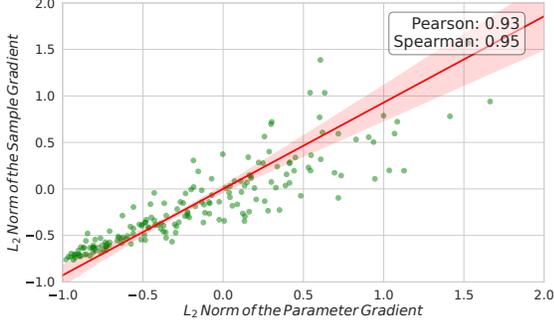
Figure 2: Correlation between the L_2 norms of sample and parameter gradient (with normalization: $g = \frac{g-\mu}{\sigma}$).

leverages the intermediate AEs during the optimization process. Since using $\|\nabla_{x'}\ell(x', y; \theta^s)\|_2$ as loss for model optimization is non-trivial, we use generated x'_t to achieve this:

$$
\begin{aligned}
\ell(x'_t, y; \theta^s) = (x'_t - x'_{t-1})^T \cdot \nabla_{x'_{t-1}}\ell(x'_{t-1}, y; \theta^s) \\
+ \ell(x'_{t-1}, y; \theta^s) + \mathcal{O}(\|\delta\|_2^2),
\end{aligned} \tag{9}
$$

where x'_t denotes the generated AEs in the t-th iteration, and $(x'_{t+1} - x'_t)^T \cdot \nabla_{x'}\ell(x', y; \theta^s)$ approximates $\|\nabla_{x'}\ell(x', y; \theta^s)\|_2$. Consequently, we can also achieve flat local maxima by minimizing $\ell(x'_t, y; \theta^s)$. Additionally, we add the term $\ell(x, y; \theta^s)$ to maintain the stability of the surrogate model. Thus, the overall loss function $\mathcal{L}_{AWT}$ for model tuning at the iteration t is given by:

$$
\mathcal{L}_{AWT} = \ell(x'_t, y; \theta^s) + \ell(x, y; \theta^s). \tag{10}
$$

Achieving Flat Local Minima from Parameter Space. In addition to the flat local maxima of AEs, Proposition 1 suggests that model smoothness also contributes to adversarial transferability.

Proposition 2: *For a model f_θ parameterized by θ and a loss function $\ell(x, y; \theta)$, there is a positive correlation between the L_2 of the input gradient $\|\nabla_x \ell(x, y; \theta)\|_2$ and that of the parameter gradient $\|\nabla_\theta \ell(x, y; \theta)\|_2$.*

Proposition 2 indicates that the L_2 norms of the input gradient and the parameter gradient are positively correlated. To empirically validate this correlation, we visualize $\|\nabla_x \ell(x, y; \theta)\|_2$ and $\|\nabla_\theta \ell(x, y; \theta)\|_2$ for each sample in Figure 2, from which we can observe a strong positive correlation between the L_2 norm of the sample gradient and that of the parameter gradient, with Pearson and Spearman correlation coefficients of 0.93 and 0.95, respectively. Leveraging this correlation, we can instead minimize $\|\nabla_{\theta^s}\ell(x', y; \theta)\|_2$. However, directly incorporating the gradient norm as a loss term introduces significant computational overhead due to the necessity of computing second-order derivatives. To address this challenge, we propose reformulating the problem as a bi-level optimization:

$$
\min_{\theta^s} \max_{\|\eta\|_2 \le \rho} \mathcal{L}_{AWT}(x'_t, x, y; \theta^s \oplus \eta), \tag{11}
$$

where η denotes the l_2 norm bound of parameter perturbations. In this framework, the inner optimization problem implicitly minimizes the gradient norm by seeking the most

---

Algorithm 1: Adversarial Weight Tuning (AWT) attack

Require: Benign input x with ground-truth label y; the loss function ℓ with surrogate model f_{θ^s}; the magnitude of perturbation ϵ; the maximum number of iterations T; the decay factor μ; the balanced coefficient ω; the step size of parameters: β, lr; the number of randomly sampled examples N.

Ensure: An adversarial example x'.

1: $g_0 = 0$, $x'_0 = x$, $\theta_{temp} = \theta^s$
2: **for** $t = 0, 1, \cdots, T - 1$ **do**
3:     Set $\bar{g} = 0$
4:     Compute the perturbed $\hat{\theta}^s$ with $\theta^s + \beta \cdot \nabla_{\theta^s}\mathcal{L}_{AWT}$
5:     Update surrogate θ^s with $\theta^s - lr \cdot \nabla_{\hat{\theta}^s}\mathcal{L}_{AWT}$
6:     *// Adjust the parameters of θ^s to obtain surrogate model with better transferability.*
7:     **for** $i = 0, 1, \cdots, N - 1$ **do**
8:         Randomly sample an example $x^* \sim \mathcal{N}(x'_t, \zeta)$
9:         Calculate the gradient at x^*: $g' = \nabla_{x^*}\ell(x^*, y; \theta^s)$
10:        Compute the predicted point by $x^* = x^* - \alpha \cdot \frac{g'}{\|g'\|_1}$
11:        Calculate the gradient at x^*: $g^* = \nabla_{x^*}\ell(x^*, y; \theta^s)$
12:        Update gradient by $\bar{g} = \bar{g} + \frac{1}{N}[(1-\omega)\cdot g' + \omega \cdot g^*]$
13:     **end for**
14:     $g_{t+1} = \mu \cdot g_t + \frac{\bar{g}}{\|\bar{g}\|_1}$
15:     Update x'_{t+1} via $x'_{t+1} = \Pi_{\mathcal{B}_\epsilon(x)}[x'_t + \alpha \cdot \text{sign}(g_{t+1})]$
16: **end for**
17: $\theta = \theta_{temp}$
18: **return** $x' = x'_T$

---

challenging perturbations under which the model should remain robust. Simultaneously, the outer optimization adjusts the parameters θ^s to minimize the worst-case loss, thereby enhancing model smoothness. This approach preserves the benefits of gradient norm minimization while avoiding prohibitive computational costs.

For practical implementation, we incorporate previous gradient-based techniques (Qiu et al. 2024; Ge et al. 2023) (flat local maxima of input space) to achieve flat local maxima in both input and parameter spaces, thereby improving the generalization of AEs across different models. The complete algorithm for generating AEs is detailed in Algorithm 1, which also ensures a low $\|\nabla_{\theta^s}\ell(x', y; \theta^s)\|_2$ by optimizing samples around x'.

Evaluation

Experimental Settings

Dataset. Evaluation is conducted on the ImageNet-compatible dataset, which is widely utilized in prior work (Qin et al. 2022; Ge et al. 2023; Qiu et al. 2024). This dataset comprises 1,000 images which we resized with a resolution of $224 \times 224 \times 3$ pixels, along with their corresponding ground-truth labels.

Models. To demonstrate the efficacy of our methods, we evaluate attack performance on several popular pre-trained models including **CNN-based** and **Transformer-based** models. For CNN models, we consider: ResNet-18 (RN18), ResNet-50 (RN50), and ResNet-101(RN101) (He

Model	Attack	ViT-B/16	ViT-B/32	Swin-T	Swin-S	ConViT	LeViT	CAiT-S	DeiT-B	PiT-B
	MI	11.7±0.3	11.6±0.2	25.1±0.6	20.6±0.5	11.5±0.3	17.8±0.2	12.3±0.4	10.9±0.5	15.5±0.3
	NI	11.3±0.3	11.3±0.3	25.2±0.1	20.9±0.4	11.2±0.2	17.7±0.2	11.6±0.4	10.7±0.2	14.5±0.1
	VMI	28.6±0.2	21.6±0.3	42.5±0.3	38.1±0.3	27.4±0.4	36.0±0.4	29.4±0.4	28.5±0.9	33.3±0.3
	VNI	23.0±0.7	17.4±0.4	40.8±0.5	35.4±0.3	23.0±0.2	33.5±0.6	24.4±0.4	23.6±0.3	29.3±0.4
RN50	EMI	19.3±0.2	15.6±0.3	39.5±0.6	32.8±0.1	19.8±0.4	30.6±0.4	20.9±0.5	19.2±0.4	25.1±0.3
	RAP	10.8±0.2	10.9±0.2	23.9±0.1	19.9±0.6	10.6±0.1	16.5±0.2	11.4±0.1	9.9±0.2	14.6±0.4
	PGN	49.2±0.4	40.8±0.2	66.0±0.1	59.2±0.2	46.8±0.4	60.4±0.6	50.9±0.2	49.5±0.3	53.0±0.2
	NCS	49.1±0.3	39.7±0.3	63.7±0.1	59.4±0.2	48.0±0.2	58.8±0.5	51.5±0.2	50.6±0.3	54.1±0.5
	AWT	**59.0±0.3**	**46.6±0.5**	**72.5±0.3**	**67.6±0.4**	**56.6±0.4**	**69.0±0.3**	**61.3±0.2**	**60.0±0.4**	**62.7±0.5**
	MI	18.5±0.4	14.4±0.2	31.6±0.6	28.1±0.4	18.7±0.5	25.8±0.9	19.2±0.1	18.3±0.3	23.3±0.4
	NI	17.3±0.3	13.5±0.3	31.2±0.6	27.8±0.9	17.8±0.5	25.5±0.2	18.1±0.4	16.9±0.2	22.0±0.2
	VMI	36.0±0.2	26.2±0.2	48.8±0.3	45.3±0.5	35.8±0.6	44.3±0.7	37.9±0.4	37.0±0.6	41.0±0.9
	VNI	31.0±0.8	21.9±0.7	47.5±0.4	42.2±1.1	31.0±0.6	42.1±0.9	33.5±0.5	32.2±0.7	37.8±1.1
RN101	EMI	30.4±1.0	21.3±0.2	51.5±0.5	44.8±0.5	30.8±0.4	44.8±0.6	33.5±0.5	31.8±0.4	37.7±0.5
	RAP	16.4±0.3	13.6±0.1	29.4±0.3	25.6±0.3	16.7±0.2	23.8±0.1	17.3±0.3	15.6±0.2	20.9±0.2
	PGN	63.0±0.5	51.9±0.3	73.4±0.3	69.4±0.2	60.2±0.5	70.7±0.4	65.9±0.3	63.5±0.1	65.8±0.1
	NCS	58.5±0.1	48.3±0.4	69.7±0.2	66.3±0.2	57.5±0.3	66.0±0.3	62.5±0.2	59.6±0.1	62.6±0.1
	AWT	**68.0±0.4**	**55.1±0.4**	**76.8±0.2**	**73.8±0.2**	**65.9±0.3**	**75.3±0.6**	**71.3±0.1**	**68.9±0.5**	**70.1±0.2**

Table 1: The attack success rates (%±std, over 10 random runs) of nine gradient-based attacks on nine normally trained Transformer-based models. The best and second results are **bold** and underline, respectively.

et al. 2016), Inception-v3 (Inc-v3) (Szegedy et al. 2016), Inception-v4 (Inc-v4), InceptionResNet-v2 (IncRes-v2) (Szegedy et al. 2017), VGG-13, VGG-16 (Simonyan and Zisserman 2014) and DenseNet-121 (Dense121) (Huang et al. 2017). Additionally, we consider Transformer-based models, including ViT-Base/16 (ViT-B/16), ViT-Base/32 (ViT-B/32) (Dosovitskiy et al. 2020), PiT-B (Heo et al. 2021), Swin-T, Swin-S (Liu et al. 2021), ConViT (d'Ascoli et al. 2021), LeViT (Graham et al. 2021), CAiT-S (Touvron et al. 2021b), DeiT-B (Touvron et al. 2021a)

Compared Methods. As for baselines, we employ eight popular gradient-based iterative adversarial attacks: MI (Dong et al. 2018), Nesterov Accelerated FGSM (NI) (Lin et al. 2019), VMI (Wang and He 2021), EMI (Wang et al. 2021c), and Robust Adversarial Perturbations (RAP) (Qin et al. 2022), Probabilistic Gradient Noise (PGN) (Ge et al. 2023) and Neighborhood Conditional Sampling (NCS) (Qiu et al. 2024).

Hyper-parameters. Following the previous work (Ge et al. 2023; Qin et al. 2022; Qiu et al. 2024), we set the maximum perturbation $\epsilon = 16.0/255$, the number of iterations $T = 10$, and the step size $\alpha = 1.6$. For MI and NI, the decay factor $\mu = 1.0$. For VMI, we set the number of sampled examples $N = 20$ and the upper bound of the neighborhood size $\beta = 1.5 \times \epsilon$. For EMI, we set $N = 11$, the sampling interval bound $\eta = 7$, and use linear sampling. For the RAP attack, we set $\alpha = 2.0/255$, $K = 400$, the inner iteration number $T = 10$, the late-start $K_{LS} = 100$, and the size of neighborhoods $\epsilon_n = 16.0/255$. For PGN, NCS and AWT, we set $N = 20$, the balanced coefficient $\delta = 0.5$, and the upper bound $\zeta = 3.0 \times \epsilon$. For AWT alone, we set $\beta = 0.005$ and $lr = 0.002$ for surrogate models used for evaluation.

Note. *Due to the extensive amount of evaluation results, we do not include all the data in the main body of the paper. The complete set of results is available in the appendix.*

Measurement of Adversarial Transferability

With Equation 3, we wonder if it is possible to derive a metric for the measurement of transferability, which can also serve as a metric to measure the transferability of AEs. Specifically, we propose quantifying the difference between the benign model and the perturbed one using AEs. Here, we define the transferability score T as follows:

$$T(\mathcal{A}, \theta^s) = \mathbb{E}_{\substack{x \in \mathcal{D} \\ \eta_i \in \Omega}} \left[\| f_{\theta^s \oplus \eta_i}(\mathcal{A}(x)) - f_{\theta^s}(\mathcal{A}(x)) \|_2 \right], \quad (12)$$

where $\mathcal{A}$ denotes the attack method. The transferability score T measures the expected difference in loss between the surrogate model and the perturbed one when evaluated on AEs. A lower value of T indicates a higher transferability, meaning that the AEs generated on the θ^s are more likely to fool the target models. Considering that obtaining Ω is not trivial, we assume that the l_2 norm of the perturbation in Ω is bounded. Specifically, we sample $\theta \in \mathcal{B}_\varepsilon(\theta)$, where $\mathcal{B}_\varepsilon(\theta)$ denotes the ε-ball around θ, to approximate the target models. The proposed metric can be used to evaluate the effectiveness of various attack strategies in generating AEs that are transferable across different models. For the hyper-parameters of the metric in Eq.(12), we sample 10 perturbations around the parameters of the surrogate models with the distribution $\mathcal{N}(\theta^s, \varepsilon)$ and $\varepsilon \in \{0.001, 0.01, 0.1\}$.

Experimental Results

Evaluation on Transformer-based Models. With the rise of large models in various domains, transformer architecture models have become increasingly prevalent due to their superior performance in handling complex tasks. The success of these models has led to a surge in their adoption across a wide range of applications. Therefore, we mainly focus on evaluating the effectiveness of the AWT method in generating AEs with high transferability on Transformer-based

Model	Attack	Inc-v4	Inc-v3	Dense121	VGG-16
RN50	MI	26.7±0.2	32.8±0.8	39.1±0.7	40.4±0.5
	NI	26.3±0.5	32.8±0.2	40.5±0.5	41.8±0.3
	VMI	46.9±0.5	49.1±0.3	58.1±0.7	56.5±0.4
	VNI	45.5±0.3	47.6±0.6	60.6±1.0	58.2±0.8
	EMI	43.1±0.7	48.1±0.9	64.3±0.7	63.9±0.8
	RAP	25.1±0.2	30.8±0.7	36.4±0.6	38.9±0.5
	PGN	73.2±0.3	75.3±0.3	82.5±0.2	80.2±0.2
	NCS	67.8±0.1	68.5±0.2	76.5±0.2	73.9±0.1
	AWT	**77.0±0.3**	**79.1±0.3**	**86.7±0.2**	**84.6±0.3**
RN101	MI	36.6±0.3	40.5±0.6	49.2±0.4	46.9±0.4
	NI	36.7±0.7	40.2±0.4	50.0±0.9	48.0±0.9
	VMI	55.2±0.3	57.2±0.4	66.0±1.0	60.0±0.2
	VNI	55.4±0.5	56.2±1.0	69.1±0.3	61.7±0.3
	EMI	59.6±0.8	61.3±0.4	76.4±0.3	70.1±0.4
	RAP	34.2±0.4	38.4±0.9	46.0±1.0	44.9±0.7
	PGN	82.1±0.1	81.0±0.2	86.1±0.1	82.1±0.3
	NCS	75.0±0.4	74.2±0.2	80.4±0.1	76.5±0.3
	AWT	**82.4±0.2**	**82.0±0.2**	**87.4±0.3**	**84.9±0.2**

Table 2: The attack success rates (%±std, over 10 random runs) of nine gradient-based attacks on four normally trained CNN-based models. The best and second results are **bold** and underline, respectively.
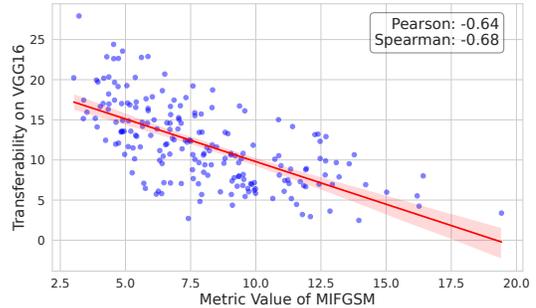

Figure 3: Relationship between the proposed metric and experimental transferability on VGG16 (target model). Each dot represents an adversarial sample. Note that AEs are generated on RN50 and the amplitude ε of the parameter perturbation is set to 0.05.

models. Table 1 presents the attack success rates (ASR) of nine gradient-based attacks on nine pretrained Transformer-based models. The results are averaged over 10 random runs, and the standard deviations are reported alongside the mean values. As shown in Table 1, among the existing methods, PGN and NCS perform particularly well, often achieving the second-highest ASR. However, AWT consistently outperforms these methods, demonstrating the superiority of our method in generating AEs with high transferability. For instance, on the ViT-B/16 model (target), AWT achieves an ASR of 59.0%, surpassing the second-best method, PGN, which has an ASR of 49.2%. Similarly, on the RN101 (surrogate), AWT achieves an ASR of 68.0%, outperforming PGN, which has an ASR of 63.0%.

Evaluation on CNN-based Models. Building on the strong performance of the AWT on Transformer-based models, we now evaluate its effectiveness on CNN-based models. Table 2 presents the ASR of nine gradient-based attacks on four pretrained CNN-based models, including Inc-v4, Inc-v3, DN121, and VGG16. As shown in Table 2, AWT consistently outperforms other attacks, achieving the highest ASRs across all CNN-based models. For instance, on the Inc-v4 model, AWT achieves an ASR of 77.0%, surpassing the second-best method, PGN, which has an ASR of 73.2%. Similarly, on the RN101 model, AWT achieves an ASR of 82.4%, outperforming PGN, which has an ASR of 82.1%. The superior performance of AWT on CNN-based models further supports its effectiveness in generating AEs with high transferability. These results are consistent with those observed on Transformer-based models, indicating that AWT is a robust method capable of enhancing the transferability of AEs across different model architectures.

Integrated with Other Attacks. Having demonstrated

the effectiveness of the AWT (only model tuning) method on both CNN- and Transformer-based models, we now investigate its impact when integrated with other attacks which also follows the approach taken in previous works (Qiu et al. 2024). AWT can further enhance the adversarial transferability of other attacks without any conflict or overlap, effectively complementing each other. Specifically, we integrate AWT with input transformation-based attacks, such as DIM (Xie et al. 2019), TIM (Dong et al. 2019), SIM (Lin et al. 2019), and Admix (Wang et al. 2021b), and evaluate the ASR on a variety of target models. Table 3 presents the ASR of these attacks when integrated with AWT, using RN50 and RN101 as the surrogate models. The best results are highlighted in bold. As shown in Table 3, integrating AWT with existing attack methods consistently improves the ASRs across all target models. For instance, when integrated with DIM, AWT increases the ASR from 43.3% to 48.5% on the Inc-v4 model, and from 55.2% to 58.1% on the Inc-v3 model when using RN50 as the surrogate model. Similarly, when integrated with TIM, AWT improves the ASR from 36.9% to 42.9% on Inc-v3 when using RN101 as the surrogate model. The improvements are substantial across all attack methods. For example, when integrated with Admix, AWT increases the ASR from 62.3% to 66.2% on Inc-v3 when using RN50 as the surrogate model. The integration of AWT with other attacks demonstrates its versatility and ability to enhance the transferability of AEs.

Transferability Measurement. To better understand the relationship between our proposed formulation (metric) (Eq.(12)) and the actual transferability, we first analyze the correlation between the metric values and the transferability obtained on target model (VGG16). Figure 3 depicts the relationship between the proposed metric and the transferability on the VGG16 model with 100 generated samples. For experimental transferability (y-axis), we use $\|f_{\theta^t}^y(x) - f_{\theta^t}^y(x')\|$ empirically, which represents how adversarial perturbations influence the target model from the confidence (without softmax) of the benign label y. The x-axis represents the calculated metric value, while the y-axis shows the corresponding transferability achieved on the VGG16

Attack	Inc-v4	Inc-Res-v2	Inc-v3	DN121	VGG16	ViT-B/16	Swin-T	LeViT	DeiT-B
DIM	43.3	38.1	46.7	55.1	52.3	23.7	37.1	31.4	24.8
DIM+AWT	**48.5**	**42.2**	**50.6**	**60.4**	**59.7**	**27.9**	**43.4**	**34.5**	**28.4**
TIM	31.5	24.2	34.0	41.1	43.7	18.0	27.2	21.3	13.8
TIM+AWT	**39.4**	**31.2**	**41.3**	**50.4**	**51.1**	**21.2**	**33.6**	**26.0**	**17.9**
Admix	49.7	42.1	52.1	65.1	62.2	27.0	44.6	36.8	27.2
Admix+AWT	**61.3**	**52.9**	**61.6**	**75.7**	**72.3**	**32.6**	**54.2**	**46.6**	**32.2**
SIM	39.3	31.1	43.2	53.9	51.6	18.6	36.0	25.8	18.4
SIM+AWT	**58.2**	**49.1**	**58.5**	**71.7**	**66.5**	**31.6**	**51.0**	**42.0**	**31.4**

Table 3: The attack success rates (%) of AWT method, when it is integrated with DIM, TIM, SIM and Admix, respectively. The AEs are generated on RN50. The best and second results are **bold**.

Model	ε	MI	NI	VMI	VNI	EMI	RAP	PGN	NCS	AWT
RN50	0.001	0.193	0.192	0.190	0.192	0.252	0.221	0.224	0.197	**0.184**
	0.010	1.751	1.816	1.687	1.867	2.128	2.136	1.978	1.646	**1.642**
	0.100	15.934	16.240	18.397	19.288	19.043	16.911	16.643	16.351	**15.852**
RN101	0.001	0.154	0.145	0.146	0.147	0.161	0.165	0.159	0.144	**0.142**
	0.010	1.500	1.394	1.374	1.360	1.583	1.471	1.512	1.382	**1.283**
	0.100	15.095	13.017	14.125	15.719	15.506	13.570	14.247	14.480	**12.121**

Table 4: The metric value of different attacks with different amplitude of parameter perturbation on surrogate models. Note that ε denotes the amplitude of the perturbation mentioned in Eq.12 and the lower value means higher transferability.

model. As shown in Figure 3, there exists a strong negative correlation between the metric value and the transferability on VGG16. The Pearson coefficient of -0.64 and the Spearman coefficient of -0.68 indicate a significant inverse relationship between the two variables. This observation suggests that higher metric values correspond to lower transferability, providing evidence that our metric can capture the characteristics of transferable AEs to a certain degree.

Further, with the evidence above, we use our proposed metric to evaluate nine attacks under different parameter perturbations with RN50 and RN101 and the results are shown in Table 4. We can observe that AWT can achieve the lowest metric value among all attacks, suggesting that it generates AEs with better transferability. This finding aligns with the results reported earlier in Table 1 and Table 2. However, we can also notice there are discrepancies between experimental results and our metric. For instance, PGN and NCS tend to perform better transferability than other attacks but they sometimes have larger metric values than attacks than VMI and VNI. This can be attributed to the insufficient sampling of our metric, leading to significant errors. Therefore, it is crucial to acknowledge that the transferability of AEs in real-world settings can be significantly more complex due to the diversity of model architectures. Given these complexities, our proposed metric, while effective in certain contexts, may not fully capture the intricacies of transferability in practice. (**Limitations**) (1) The metric assumes a certain level of similarity between the surrogate model and the target model with the same architecture, which might not always hold in practical applications. (2) Even if two models share the same architecture, differences in their training data or hyperparameter settings can lead to variations in the transferability of AEs (Wu et al. 2018), which have not been considered in the proposed metric. (3) Note that AWT has

introduced extra computational cost with two backward for calculating parameter gradients during each iteration while different samples do need different surrogate parameters to achieve better optimal. In practice, attackers should balance the trade-off between computational cost and effectiveness.

Despite the above limitations, the proposed metric serves as a valuable starting point for evaluating the transferability of AEs. It can be used as a guiding tool to identify potentially effective attacks and to inform the selection of appropriate surrogate models. Besides, further research is needed to develop more comprehensive metrics that can account for the complexities involved in real-world scenarios.

Conclusion

In this work, we focus on transferable adversarial attacks, which pose significant threats to the robustness and security of DNNs in real-world applications. Motivated by the empirical evidence that flat maxima enhance the transferability of AEs, we propose the AWT method, a novel approach aimed at finding flat adversarial regions through a max-min bi-level optimization problem. AWT leverages the interplay between the optimization of adversarial perturbations and the adjustment of the surrogate model's parameters without the requirement for extra data to achieve better transferability. Through extensive experiments on a variety of models with different architectures, we show that AWT generates AEs in flatter regions, resulting in improved transferability compared to SOTA methods. Our theoretical analysis reveals the inherent factors of the surrogate model and AEs that contribute to the transferability of AEs, and we propose a unified metric to measure the transferability of AEs across models. The enhanced transferability of AEs produced by AWT highlights the need for robust defenses and opens up new avenues for research in adversarial robustness.

Acknowledgements

References

Caesar, H.; Bankiti, V.; Lang, A. H.; Vora, S.; Liong, V. E.; Xu, Q.; Krishnan, A.; Pan, Y.; Baldan, G.; and Beijbom, O. 2020. nuscenes: A multimodal dataset for autonomous driving. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 11621–11631.

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.; et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

d'Ascoli, S.; Touvron, H.; Leavitt, M. L.; Morcos, A. S.; Biroli, G.; and Sagun, L. 2021. Convit: Improving vision transformers with soft convolutional inductive biases. In *International conference on machine learning*, 2286–2296. PMLR.

Foret, P.; Kleiner, A.; Mobahi, H.; and Neyshabur, B. 2020. Sharpness-aware minimization for efficiently improving generalization. *arXiv preprint arXiv:2010.01412*.

Ge, Z.; Liu, H.; Xiaosen, W.; Shang, F.; and Liu, Y. 2023. Boosting adversarial transferability by achieving flat local maxima. *Advances in Neural Information Processing Systems*, 36: 70141–70161.

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

Graham, B.; El-Nouby, A.; Touvron, H.; Stock, P.; Joulin, A.; Jégou, H.; and Douze, M. 2021. Levit: a vision transformer in convnet's clothing for faster inference. In *Proceedings of the IEEE/CVF international conference on computer vision*, 12259–12269.

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

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.

Heo, B.; Yun, S.; Han, D.; Chun, S.; Choe, J.; and Oh, S. J. 2021. Rethinking spatial dimensions of vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, 11936–11945.

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.

Keskar, N. S.; Mudigere, D.; Nocedal, J.; Smelyanskiy, M.; and Tang, P. T. P. 2016. On large-batch training for deep learning: Generalization gap and sharp minima. *arXiv preprint arXiv:1609.04836*.

Kurakin, A.; Goodfellow, I.; and Bengio, S. 2016. Adversarial machine learning at scale. *arXiv preprint arXiv:1611.01236*.

Li, Y.; Bai, S.; Zhou, Y.; Xie, C.; Zhang, Z.; and Yuille, A. 2020. Learning transferable adversarial examples via ghost networks. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, 11458–11465.

Lin, J.; Song, C.; He, K.; Wang, L.; and Hopcroft, J. E. 2019. Nesterov accelerated gradient and scale invariance for adversarial attacks. *arXiv preprint arXiv:1908.06281*.

Liu, Y.; Chen, X.; Liu, C.; and Song, D. 2016. Delving into transferable adversarial examples and black-box attacks. *arXiv preprint arXiv:1611.02770*.

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*, 10012–10022.

Ma, W.; Li, Y.; Jia, X.; and Xu, W. 2023. Transferable adversarial attack for both vision transformers and convolutional networks via momentum integrated gradients. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 4630–4639.

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

Povey, D.; Ghoshal, A.; Boulianne, G.; Burget, L.; Glembek, O.; Goel, N.; Hannemann, M.; Motlicek, P.; Qian, Y.; Schwarz, P.; et al. 2011. The Kaldi speech recognition toolkit. In *IEEE 2011 workshop on automatic speech recognition and understanding*. IEEE Signal Processing Society.

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.

Qiu, C.; Duan, Y.; Zhao, L.; and Wang, Q. 2024. Enhancing Adversarial Transferability Through Neighborhood Conditional Sampling. *arXiv preprint arXiv:2405.16181*.

Radford, A.; Wu, J.; Child, R.; Luan, D.; Amodei, D.; Sutskever, I.; et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8): 9.

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

Simonyan, K.; and Zisserman, A. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Szegedy, C.; Ioffe, S.; Vanhoucke, V.; and Alemi, A. 2017. Inception-v4, inception-resnet and the impact of residual connections on learning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 31.

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.

Touvron, H.; Cord, M.; Douze, M.; Massa, F.; Sablayrolles, A.; and Jégou, H. 2021a. Training data-efficient image transformers & distillation through attention. In *International conference on machine learning*, 10347–10357. PMLR.

Touvron, H.; Cord, M.; Sablayrolles, A.; Synnaeve, G.; and Jégou, H. 2021b. Going deeper with image transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, 32–42.

Wang, H.; Dong, K.; Zhu, Z.; Qin, H.; Liu, A.; Fang, X.; Wang, J.; and Liu, X. 2024a. Transferable Multimodal Attack on Vision-Language Pre-training Models. In *IEEE Symposium on Security and Privacy, SP 2024, San Francisco, CA, USA, May 19-23, 2024*, 1722–1740. IEEE.

Wang, J.; Liu, A.; Yin, Z.; Liu, S.; Tang, S.; and Liu, X. 2021a. Dual Attention Suppression Attack: Generate Adversarial Camouflage in Physical World. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, 8565–8574. Computer Vision Foundation / IEEE.

Wang, J.; Liu, X.; Yin, Z.; Wang, Y.; Guo, J.; Qin, H.; Wu, Q.; and Liu, A. 2024b. Generate Transferable Adversarial Physical Camouflages via Triplet Attention Suppression. *Int. J. Comput. Vis.*, 132(11): 5084–5100.

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. 2021b. Admix: Enhancing the transferability of adversarial attacks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 16158–16167.

Wang, X.; Lin, J.; Hu, H.; Wang, J.; and He, K. 2021c. Boosting adversarial transferability through enhanced momentum. *arXiv preprint arXiv:2103.10609*.

Wang, Y.; Wang, J.; Yin, Z.; Gong, R.; Wang, J.; Liu, A.; and Liu, X. 2022. Generating Transferable Adversarial Examples against Vision Transformers. In Magalhães, J.; Bimbo, A. D.; Satoh, S.; Sebe, N.; Alameda-Pineda, X.; Jin, Q.; Oria, V.; and Toni, L., eds., *MM '22: The 30th ACM International Conference on Multimedia, Lisboa, Portugal, October 10 - 14, 2022*, 5181–5190. ACM.

Wu, L.; Zhu, Z.; Tai, C.; et al. 2018. Understanding and enhancing the transferability of adversarial examples. *arXiv preprint arXiv:1802.09707*.

Wu, T.; Luo, T.; and Wunsch II, D. C. 2024. LRS: Enhancing Adversarial Transferability through Lipschitz Regularized Surrogate. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 6135–6143.

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.

Yang, Z.; Li, L.; Xu, X.; Zuo, S.; Chen, Q.; Zhou, P.; Rubinstein, B.; Zhang, C.; and Li, B. 2021. Trs: Transferability reduced ensemble via promoting gradient diversity and model smoothness. *Advances in Neural Information Processing Systems*, 34: 17642–17655.

Zhang, J.; Huang, Y.; Wu, W.; and Lyu, M. R. 2023. Transferable adversarial attacks on vision transformers with token gradient regularization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 16415–16424.

Zhang, Y.; Hu, S.; Zhang, L. Y.; Shi, J.; Li, M.; Liu, X.; and Jin, H. 2024. Why does little robustness help? a further step towards understanding adversarial transferability. In *Proceedings of the 45th IEEE Symposium on Security and Privacy (S&P'24)*, volume 2.

Zhu, Y.; Chen, Y.; Li, X.; Chen, K.; He, Y.; Tian, X.; Zheng, B.; Chen, Y.; and Huang, Q. 2022. Toward understanding and boosting adversarial transferability from a distribution perspective. *IEEE Transactions on Image Processing*, 31: 6487–6501.