Enhancing Transferability of Targeted Adversarial Examples via Inverse Target Gradient Competition and Spatial Distance Stretching

Zhankai Li[1], Weiping Wang[1], Jie Li[1], Shigeng Zhang[1], Yunan Hu[1], Song Guo[2]
[1]Central South University
[2]The Hong Kong University of Science and Technology
janlzk7@gmail.com, {wpwang, lijie55, sgzhang, hyn318}@csu.edu.cn, songguo@ust.hk

Abstract

In the field of AI security, deep neural networks (DNNs) are highly sensitive to adversarial examples (AEs), which can cause incorrect predictions with minimal input perturbations. Although AEs exhibit transferability across models, targeted attack success rates (TASRs) are low due to differences in feature dimensions and decision boundaries. To enhance targeted AE transferability, we propose a novel approach using Inverse Target Gradient Competition (ITC) and Spatial Distance Stretching (SDS) in the optimization process. Specifically, we employ a siamese-network-like framework to generate both non-targeted and targeted AEs. The ITC mechanism applies non-targeted adversarial gradients each epoch to impede the optimization of targeted perturbations, thereby improving robustness. Additionally, a top-k SDS strategy guides AEs to penetrate target class regions in the latent space while distancing from non-targeted regions, achieving optimal transferability. Compared to state-of-the-art competition-based attacks, our method significantly improves transferable TASRs by 16.1% and 21.4% on mainstream CNNs and ViTs, respectively, and demonstrates superior defense-breaking capabilities. Our code is available here.

1. Introduction

Adversarial attacks involve adding small, often imperceptible perturbations to input data, which can lead otherwise high-performing models to produce incorrect outputs. The concept of adversarial attacks was first introduced by [23] and further developed by [7], who demonstrated the vulnerability of deep neural networks (DNNs) to these perturbations. With the rapid advancement of DNNs in recent years, the security risks revealed by adversarial attacks have become a critical issue. Adversarial attacks are now widely used to assess model security, improve model robustness, and drive research in defense technologies.

However, the significance of adversarial attacks extends beyond these applications, as they have inadvertently become a widely used defense mechanism in real-world scenarios. A prime example is Google's login verification system, reCAPTCHA, which aims to distinguish real users from bots to prevent abuse. Adversarial examples (AEs) are one of the primary image verification methods employed, as illustrated in Figure 1. This raises an important question: is it secure enough? Since we do not know the bot's model architecture or training domain, the transferability of AEs becomes especially crucial. Most current research on adversarial transferability focuses on non-targeted attacks [14, 40], even extending to cross-architecture and cross-domain attacks [15, 37], achieving notable performance. In contrast, targeted attacks face a significant increase in difficulty due to the substantial changes in decision boundaries, and as a result, they still maintain a relatively low success rate. However, for login verification, given that the output category of non-targeted attacks is uncontrollable, and even the original verification images without perturbation may have "low-quality" issues (leading to misidentification by bots), in such cases, targeted AEs are often safer and more reliable than non-targeted AEs that may cause the deployed bots to misjudge into the ground truth category.

Currently, effective targeted attack methods are primarily based on transformation [1, 18, 29, 34], achieving transferability through input augmentation. However, the relentless pursuit and stacking of a greater variety of transformations, although enhancing transferability to some extent, does not guarantee robust performance in white-box attacks. Moreover, this stacking approach based on affine and perspective transformations offers limited insight into the complexity and vulnerability of DNNs. Recently, competition-based attacks have been proposed [2, 30], offering a unique perspective. Their advantage lies in compelling AEs to consider the more complex and diverse feature representations within DNNs during their generation, simulating various noise interferences and attack strategies,

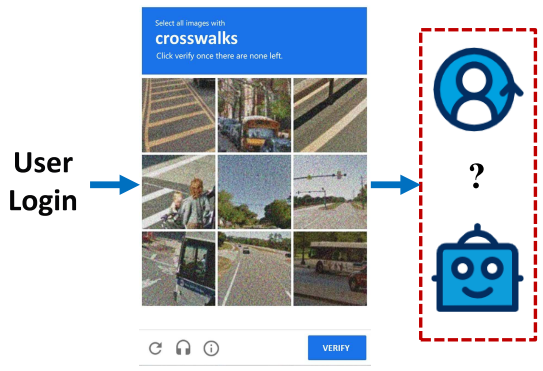

Figure 1. Example of Google's web login verification system, re-CAPTCHA. It determines whether a user is human or a bot based on whether they can correctly match text to corresponding images.

and promoting the learning of more robust feature combinations. The competitive mechanism, by delving into the model's decision boundaries, helps to uncover the latent vulnerabilities of DNNs and provides a new perspective for understanding and improving the model's generalization capabilities. However, no matter what the attack method is, it cannot conceal the widespread insufficiency in the transferability performance of current targeted adversarial attacks.

To address this challenge, we propose a new targeted adversarial attack method called ITDS, which consists of two components: Inverse Target Gradient Competition (ITC) and Spatial Distance Stretching (SDS), aimed at enhancing the transferability of targeted AEs. In the first part, we adopt a siamese-network-like framework, innovatively leveraging the ground truth gradients of multiple target class non-targeted AEs at intermediate stages to exert inverse "competitiveness" on non-target class targeted AEs during the optimization process. In the second part, to tackle the multi-dimensional complexity of decision boundaries in targeted attacks, we introduce a top-k SDS strategy to seek stable perturbation update directions. This guides AEs to penetrate target class regions within the latent multi-dimensional space while globally distancing from several closest non-target region boundaries. The experimental results demonstrate that the proposed ITDS method has achieved significant advantages in enhancing the transferability and defense-breaking capabilities of targeted AEs, surpassing the existing state-of-the-art (SOTA) methods.

2. Related Work

In this section, we will briefly overview the background of adversarial attacks. Then focus on various targeted attack techniques that demonstrate effective transferability.

2.1. Adversarial Attacks

Goodfellow et.al first designed a simple yet effective method called Fast Gradient Sign Method (FGSM) [7]. This method generates AEs by making small adjustments to the input x (original sample) along the direction of the gradient, i.e., $x^{adv} = x + \delta$. In the case of targeted attacks, FGSM updates the input x in the direction that minimizes the classification loss for a given target class y, with the objective function as follows:

$$\arg\min_{x^{adv}} \mathcal{J}(x^{adv}, y; \theta) \;\; \text{s.t.} \;\; \|x - x^{adv}\|_\infty \leq \epsilon, \qquad (1)$$

where θ is the model parameters, $\mathcal{J}$ denotes the Cross-Entropy (CE) loss function, and the ℓ_∞ norm is used to constrain the perturbation within the range ϵ.

I-FGSM [13] is an iterative extension of FGSM, which gradually refines adversarial perturbation with small step sizes. To address gradient vanishing issues in targeted attacks, [39] proposes using Logits loss instead of CE loss, significantly enhancing targeted attack effectiveness. Based on this, [33] further proposes downscaling logit calibration with a temperature factor and an adaptive margin. While these methods provide a foundation for enhancing targeted attack effectiveness, their transferability remains limited.

2.2. Transferable Adversarial Attacks

In this section, we introduce advanced transferable attacks that can successfully deceive other models using AEs generated on a single surrogate model without additional training. We categorize these methods into transformation-based and competitive-based attacks, focusing primarily on transferable targeted attacks, along with certain non-targeted attacks that can also be adapted for targeted attacks.

Transformation-Based Attacks. DIM [34] is a representative transformation-based attack, which resizes and pads images with a certain probability at each step of the multi-step attack process. SSM [18] applies frequency domain augmentations through spectrum transformations on the input. ODIM [1] method significantly improves the targeted attack transferability by mapping input images to random 3D objects and applying varied rendering techniques. SIA [31] generates structurally diverse AEs by applying various random transformations to the images. Both ODIM and SIA integrate DIM into their implementation principles and share a similar lineage in transformation strategies. BSR [29] randomly shuffles and rotates the blocks within images, calculates the ensemble gradients of this new set of images, reduces attention heatmap differences across various models, and significantly enhances transferability.

Competition-based Attacks. Admix [30] combines features from different categories in the input domain, creating multiple images for gradient computation, and enhancing transferability without degrading white-box performance. FIA [32] employs aggregated gradients to disrupt essential object-aware features while suppressing model-specific features, boosting AE transferability. RPA [38] adopts a similar approach to FIA, introducing random patch transforma-

tions to benign images. To alleviate the overfitting dilemma common in an AE crafted by simple iterative attacks, FFT [35] encourages features conducive to the target class while discouraging those associated with the original class in an intermediate layer of the source model. CFM [2] introduces a competitive strategy, crafting targeted adversarial perturbations with diverse features by introducing two types of competitor noises: adverse perturbations toward different target classes and friendly perturbations towards the correct class, achieving SOTA results.

3. The Design of ITDS

This section consists of two parts: the Inverse Target Gradient Competition mechanism and the Spatial Distance Stretching strategy. In each part, we first explain the motivation and then propose the corresponding method.

3.1. Inverse Target Gradient Competition (ITC)

3.1.1. Motivation of ITC

The motivation for the first part of our research arises from a question: *What is a competition mechanism*? From existing work related to competition mechanisms, we summarize it as a way to enhance the transferability of AEs by promoting diversified attack strategies through simulating various noise interferences. Since noise interference is involved, the essence of competitiveness can be interpreted as an inverse force. For targeted attacks, could we identify the most competitive noise? We believe this is the key issue. To address this critical question, we turn to target class samples and consider the gradients of the target label for these samples. These gradients essentially represent the optimal force linking target class samples with the target label, which can be used in both forward and inverse directions. Therefore, we have a bold hypothesis: when using this gradient in inverse, it not only effectively generates non-targeted AEs but also serves as the optimal competitiveness for targeted attacks.

3.1.2. Method of ITC

Figure 2 shows the overall framework of our Inverse Target Gradient Competition mechanism, which is jointly driven by both non-targeted and targeted attacks within a siamese-network-like architecture. Our strategy is to utilize the inverse target gradient to introduce resistance during the optimization process of targeted AEs. This inverse target gradient can be generated for any samples; but which inverse target gradient should we choose? As mentioned earlier, we seek the strongest resistance, which is the inverse gradient of the target label with respect to the target class sample. The approach of generating the inverse target gradient is illustrated in the lower part of Figure 2, where we use 'cock' as the target label as an example. In each sub-epochs (epochs$_{sub}$), multiple target class images that are correctly classified by the siamese classifier are selected to generate non-targeted AEs. The optimization goal is to minimize the logit value of the target label. We set epochs$_{sub}$ to be one-tenth of the total epochs, as we believe that target class samples need a certain number of iterations to transform into robust non-targeted AEs. Accordingly, the non-targeted AEs generated in each epochs$_{sub}$ contain different levels and increasingly deeper latent features of target class. Therefore, we consider the inverse target gradient perturbations generated in each epochs$_{sub}$ to also have diverse potential competitive characteristics, and during the generation of non-targeted AEs in each epoch, multiple inverse target perturbations are generated to enrich competitive characteristics. The formulas are as follows:

$$\tilde{g}_t^z = \nabla_{z_t^{adv}} \mathcal{L}(z_t^{adv}, y; \theta), \tag{2}$$

$$g_{t+1}^z = \mu \cdot g_t^z + \tilde{g}_t^z, \tag{3}$$

$$z_{t+1}^{adv} = \mathrm{Clip}_{z,\epsilon}\{z_t^{adv} - \alpha \cdot \mathrm{sign}(g_{t+1}^z)\}, \tag{4}$$

where z is the target class sample, $\tilde{g}_t^z$ represents the inverse target gradient, and we follow [1, 2, 39] to use the Logit loss $\mathcal{L}$ (opposite to CE loss optimization). g_{t+1}^z represents the momentum gradient without normalization for z, and z_{t+1}^{adv} is the non-targeted AE generated in each round.

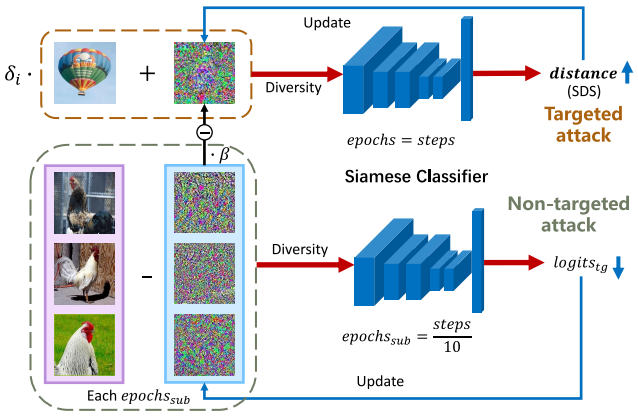

Figure 2. The overall framework of the ITC, e.g., combined with DIM, utilizes a siamese-network-like architecture to generate targeted and non-targeted AEs simultaneously. The upper part focuses on targeted attacks against non-target class samples, while the lower part addresses non-targeted attacks on target class samples. Note that for non-targeted attacks, the optimization condition utilizes the logits value of the target label, whereas the goal of the targeted attack is to stretch the spatial distance proposed in SDS.

The process of generating targeted AEs is shown in the upper part of Figure 2, where the input is a non-target image, and the optimization goal is to maximize the spatial distance, which we introduce in the SDS of the next section. Each round of generated targeted AEs competes with

a certain proportion of inverse target gradients from different stages. The formulas are as follows:

$$\bar{g}_t^{cpt} = \frac{1}{n \cdot m} \sum_{z \in Z} \sum_{i=0}^{m-1} \nabla_{x_t^{adv}} \mathcal{L}\big(\delta_i \cdot (x_t^{adv} - \beta \cdot \tilde{g}_t^z), y; \theta\big), \tag{5}$$

$$g_{t+1} = \mu \cdot g_t + \bar{g}_t^{cpt}, \tag{6}$$

$$x_{t+1}^{adv} = \mathrm{Clip}_{x,\epsilon}\{x_t^{adv} + \alpha \cdot \mathrm{sign}(g_{t+1})\}, \tag{7}$$

where n represents the number of target class samples in per epochs$_{sub}$, δ_i and m are the scaling factor and level, β is the competition ratio, and each round of targeted AEs competes with the corresponding inverse target gradient.

There are two notable points. The first point (P1) is whether, for targeted or non-targeted attacks, the calculation of momentum gradients differs from the traditional MI-FGSM [3], and we have not employed norm normalization (N-norm). One of the reasons is the absence of gradient explosion in our experimental evaluations, and more importantly, our experiments have shown that using N-norm would limit our attack performance. We believe that N-norm might restrict the magnitude of gradient updates, resulting in less aggressive generation of AEs. Without N-norm, gradient updates could become more flexible, allowing for more effective exploration of the model's decision boundary, thereby generating more aggressive AEs.

The second point (P2) is that we did not use the momentum gradient as the competitive gradient. The reason is that the momentum gradient is essentially a cumulative gradient. It weakens the strength of the inverse target features in the corresponding round. Since we want to maintain the integrity of the inverse target features, using it without any normalization operation in conjunction with P1 would cause excessive damage to the input features. This would be catastrophic for targeted attacks. These two points will also be discussed in subsequent ablation experiments.

(a) (b) (c) (d)

Figure 3. Comparison of targeted AEs with strong and weak transferability. Panels (a) and (c) show AEs with weak transferability, while (b) and (d) show AEs with strong transferability.

3.2. Spatial Distance Stretching (SDS)

3.2.1. Motivation of SDS

The motivation for the second part of our study arises from an interesting phenomenon commonly observed in highly transferable targeted AEs. As shown in Figure 3, we present two 'cock'-targeted examples as representatives: targeted AEs generated using ResNet50 with strong and weak transferability, from left to right. In (a) and (c), we have AEs that can only fool fewer than three black-box models into predicting the target class, while (b) and (d) show AEs capable of misleading over twelve black-box models to the target class. Table 1 presents the top-5 average confidence scores for each AE on the successfully misled black-box models. It can be seen that the average confidence for the target la-

Label	Figure	1	2	3	4	5
motor_scooter	(a)	42.0%	5.5%	4.3%	3.6%	3.2%
	(b)	83.3%	4.9%	2.1%	1.1%	0.3%
siamese_cat	(c)	49.6%	5.6%	5.3%	5.1%	5.0%
	(d)	90.5%	2.1%	1.3%	0.7%	0.2%

Table 1. The average top-5 confidence on black-box models successfully attacked by targeted AEs as shown in Figure 3.

Algorithm 1 The ITDS Attack Algorithm

Input: Siamese classifier f with parameters θ, Logit loss $\mathcal{L}$, our Distance loss $\mathcal{D}$, and an original sample x with target label y
Input: The maximum perturbation ϵ, the step size α, the number of epochs T and epochs$_{sub}$ T_{sub}
Input: The decay factor μ, scaling level m, and n random target class samples Z ($z \in Z$)
Input: The competition ratio β, and $topk$ classifications k
Output: An adversarial example x^{adv}.

$g_0^z = 0$; $g_0 = 0$; $z_0^{adv} = z$; $x_0^{adv} = x$;
for t **in** $range(T)$ **do**
  if t % T_{sub} == 0 **then**
    Reselect n random target class samples Z
  end if
  Get the inverse target gradient for each $\tilde{g}_t^z$ of n random target class samples by Eq.(2)
  Get the momentum inverse target gradient for each g_{t+1}^z by Eq.(3)
  Update each z_{t+1}^{adv} by Eq.(4)
  Get the set Q of top-k classifications by $f(x_t^{adv})$
  Calculate the integrate gradient of competition examples by:

$$\bar{g}_t^{cpt} = \frac{1}{n \cdot m} \sum_{z \in Z} \sum_{i=0}^{m-1} \nabla_{x_t^{adv}} \mathcal{D}\big(\delta_i \cdot (x_t^{adv} - \beta \cdot \tilde{g}_t^z), y\big)$$

  Get the momentum gradient by Eq.(6)
  Update the adversarial example x_{t+1}^{adv} by Eq.(7)
end for
return $x^{adv} = x_T^{adv}$.

bel (top-1) in highly transferable AEs is significantly higher than that in low-transferability AEs. Notably, highly transferable AEs not only exhibit higher top-1 confidence but also have lower confidence scores for subsequent rankings (rank-2 and beyond) compared to low-transferability AEs, with each subsequent rank showing an even larger gap.

3.2.2. Method of SDS

Based on this observation, we naturally propose a hypothesis by inverse inference: increasing the gap between the target class and other classes under a top-k setting can enhance adversarial transferability. Therefore, the key question becomes how to widen this gap. To address this, we devised two strategies from different perspectives, as shown in the following formulas:

$$f_{untg}(x) = \frac{1}{k-1}\sum_{q\in Q} f_q(x) \quad \text{s.t.} \quad y \notin Q, \tag{8}$$

$$\mathcal{D}_{S1}(x_t^{adv}, y) = f_y(x_t^{adv}) - f_{untg}(x_t^{adv}), \tag{9}$$

$$\mathcal{D}_{S2}(x_t^{adv}, y) = \frac{f_y(x_t^{adv}) - f_{untg}(x_t^{adv})}{\left\| \nabla_{x_t^{adv}} f_y(x_t^{adv}) - \nabla_{x_t^{adv}} f_{untg}(x_t^{adv}) \right\|_2}. \tag{10}$$

The first strategy (S1) maximizes the target class logit value relative to the aggregated logits of other classes within the top-k setting, enhancing target class features in the latent feature space and diminishing those of similar other classes, thus widening the latent feature gap, as shown in Equation (9). The second strategy (S2), inspired by DeepFool's approach to calculating point-to-hyperplane distances [19], guides AEs to deeply infiltrate the target class region in multi-dimensional geometric space while maintaining a global distance from the aggregated boundaries of the closest non-target regions, as detailed in Equation (10). In these equations, $f(\cdot)$ represents the classifier, Q denotes the set of top-k classes (excluding y), and $f_q(x)$ and $f_y(x)$ represent the output values of $f(x)$ corresponding to q and y, respectively. The algorithm of ITDS is shown in Algorithm 1.

4. Experiments

We first outline the experimental setup, then evaluate the targeted-attack performance of both competition-based and transformation-based methods separately. Afterward, we evaluate the ability of our ITDS to break through defenses and conduct ablation studies.

4.1. Experimental Setup

Dataset. To ensure rigorous evaluation, we randomly selected 2,000 high-quality images from 1,000 classes in the ILSVRC2012 validation set [21]. These images were correctly identified by all models tested, including CNNs, ViTs and adversarial training models, using intersection extraction, with more than 90% top-1 average confidence. The target class images of our method were randomly drawn from the corresponding category in the ILSVRC2012 dataset.

Models. We choose four normally pre-trained CNNs, i.e., ResNet50 (RN50) [9], VGG16 [22], MobileNet-v3-large (MN-v3) [11], RegNet-y-32gf (RegN) [20] as surrogate (white-box) models. Concurrently, we choose four white-box independent normally pre-trained CNNs, i.e., Inception-v3 (Inc-v3) [24], ResNet101 (RN101), DenseNet161 (DN161) [12], EfficientNet-b7 (ENet) [25], and four ViTs, i.e., VisionTransformer-b-16 (ViT) [5], DeiT-base (DeiT) [26], ConViT-base (ConViT) [6] and PiT-base (PiT) [10] to serve as target black-box models.

Baselines. We adopt six challenging adversarial attacks as our baselines, which have released the code and provide the necessary parameters for reproduction, i.e., *Competition-based*: Admix [30], CFM [2], *Transformation-Based*: DIM [34], ODIM [1], SIA [31], and BSR [29]. All the baselines are combined with MI-FGSM.

Defenses. To further demonstrate the effectiveness of ITDS, we consider four additional advanced defense methods, i.e., JPEG [8], FD [17], GDMP [28], Score-Opt (SO) [36], and three adversarial training defense models [27], i.e., Def_{linf4} (Def_4), Def_{linf8} (Def_8) and Def_{l2-3} (Def_{l2}) from ResNet50, which are proven to be robust to adversarial attacks on ImageNet datasets.

Attack setting. For all methods, we set the maximum perturbation of $\epsilon = 16$, the number of epochs $T = 100$, the step size $\alpha = 1.6$. We adopt the decay factor $\mu = 1.0$ for all the methods that employ the MI-FGSM, and a transformation probability of $p = 0.7$ for DIM and ODIM. Furthermore, all parameters for Admix, SIA, CFM, and BSR are set as described in their respective published papers. For our ITDS, we set the number of epochs_{sub} to $T/10$, scaling level $m = 5$ with $\delta_i = 1/2^i$, the number of random target class samples $n = 3$, the competition ratio $\beta = 1.5$, and $k = 10$ of the top-k. For each AE of the trial, the target label is randomly chosen from the other 999 classes excluding its true label, and the final results are averaged over three trials.

4.2. Evaluation on Competition-based Attacks

We first evaluate the attack performance of various competition-based attacks. Since the FIA and RPA introduced in related work are only applicable to non-targeted attacks, and FFT must be combined with existing methods to be usable, we ultimately chose Admix and CFM as competitors for ITDS. To demonstrate the versatility of our approach, the surrogate models we use to craft AEs are structurally independent of each other. The targeted attack success rates (TASRs) are shown in Table 2, where the upper row indicates the targeted model under attack, and the left column lists the tested surrogate models. To test the feasi-

Model	Attack	RN50	VGG16	MN-v3	RegN	Inc-v3	RN101	DN161	EffN	ViT	DeiT	ConViT	PiT	BAvg.
RN50	Admix	100*	7.9	2.1	4.9	1.6	38.6	19.8	0.6	0.3	0.1	0.2	0.1	6.9
	CFM	100*	45.0	17.7	35.5	9.5	87.7	60.7	5.6	2.0	1.5	1.5	3.9	24.6
	ITDS$_{fs}$	**100***	**47.6**	**34.8**	**50.8**	**26.9**	**90.1**	**81.5**	**15.4**	**9.3**	**3.7**	**4.7**	**6.4**	**33.7**
	ITDS$_{gs}$	**100***	**46.2**	**35.0**	**49.3**	**28.4**	**87.6**	**80.4**	**17.9**	**10.0**	**4.1**	**4.4**	**6.6**	**33.6**
VGG16	Admix	4.3	88.4*	1.4	7.7	0.7	0.9	4.1	0.5	0.1	0.1	0.1	0.5	1.8
	CFM	3.9	86.1*	1.2	8.4	0.8	1.0	2.8	0.6	0.1	0.2	0.1	0.3	1.7
	ITDS$_{fs}$	**18.8**	**100***	**12.0**	**20.6**	**8.0**	**7.6**	**18.4**	**5.6**	**1.0**	**0.6**	**0.5**	**1.5**	**8.6**
	ITDS$_{gs}$	**18.7**	**100***	**12.0**	**19.9**	**9.0**	**8.2**	**18.5**	**5.7**	**1.4**	**0.5**	**0.8**	**1.9**	**8.7**
MN-v3	Admix	0.6	0.4	99.3*	0.3	0.6	0.3	0.4	0.4	0.2	0.1	0.2	0.1	0.3
	CFM	10.4	8.9	100*	10.4	7.4	9.6	7.7	14.3	7.2	2.7	4.0	3.1	7.8
	ITDS$_{fs}$	**16.4**	**9.6**	**100***	**15.2**	**11.8**	**16.0**	**15.2**	**19.8**	**14.5**	**3.8**	**4.2**	**4.4**	**11.9**
	ITDS$_{gs}$	**16.7**	**9.9**	**100***	**14.9**	**11.7**	**15.9**	**15.6**	**16.8**	**14.4**	**3.2**	**4.8**	**3.8**	**11.6**
RegN	Admix	1.4	1.1	1.6	95.8*	0.4	0.5	1.6	0.5	0.3	0.2	0.2	0.4	0.7
	CFM	21.4	23.3	14.1	92.3*	4.4	12.1	19.6	9.0	1.4	1.2	1.1	5.9	10.3
	ITDS$_{fs}$	**38.2**	**16.0**	**25.0**	**100***	**17.8**	**24.8**	**40.6**	**23.9**	**11.1**	**7.6**	**7.2**	**14.0**	**20.5**
	ITDS$_{gs}$	**41.2**	**17.7**	**26.8**	**99.9***	**19.7**	**27.7**	**43.1**	**25.0**	**13.3**	**9.6**	**8.4**	**17.4**	**22.7**

Table 2. TASRs (%) on twelve pre-trained models with various competition-based attacks. The AEs are crafted on the RN50, VGG16, MN-v3, and RegN models, respectively. An asterisk (*) indicates white-box attacks, and BAvg means average black-box TASR. ITDS$_{fs}$ and ITDS$_{gs}$ represent the integration of S1 and S2 in SDS, respectively, and the boldface represents their results.

Model	Attack	RN50	VGG16	MN-v3	RegN	Inc-v3	RN101	DN161	EffN	ViT	DeiT	ConViT	PiT	BAvg.
RN50	DIM	100*	33.6	14.1	31.9	15.3	65.0	59.1	13.5	3.1	1.9	2.6	3.8	22.2
	ODIM	99.9*	52.8	33.2	49.8	37.5	70.2	65.8	30.1	14.5	9.2	10.4	15.5	35.3
	SIA	100*	86.8	56.3	86.5	40.4	99.0	93.5	34.3	30.6	21.7	24.0	40.3	55.7
	BSR	100*	87.6	48.7	87.2	31.0	97.6	92.8	27.8	18.1	16.9	16.2	41.5	51.4
	Admix-DIM	100*	51.9	25.5	53.6	29.1	83.8	77.2	28.8	8.0	4.3	5.1	10.1	34.3
	CFM-DIM	99.7*	68.4	47.9	66.0	49.2	88.4	79.5	41.9	18.0	12.5	14.9	20.0	46.0
	ITDS-DIM	**100***	**79.3**	**70.3**	**82.8**	**69.6**	**95.6**	**94.6**	**62.0**	**44.4**	**23.6**	**24.7**	**31.6**	**61.7**
VGG16	DIM	12.9	85.8*	3.4	18.2	2.9	3.5	10.6	3.5	0.2	0.3	0.2	1.4	5.2
	ODIM	33.5	78.8*	10.0	28.2	11.2	13.1	25.1	13.6	2.6	1.4	1.3	4.1	13.1
	SIA	50.4	81.7*	13.5	55.6	7.3	20.5	43.7	7.7	2.1	1.2	1.6	9.2	19.3
	BSR	24.3	84.7*	3.8	34.6	1.7	7.5	25.1	1.4	0.4	0.3	0.3	2.1	9.2
	Admix-DIM	20.2	88.0*	6.8	23.9	5.7	5.4	16.0	8.0	0.8	0.5	0.5	2.1	8.2
	CFM-DIM	19.2	86.3*	4.1	23.7	3.9	5.9	14.0	6.1	0.4	0.6	0.4	1.6	7.2
	ITDS-DIM	**42.0**	**100***	**24.8**	**34.8**	**25.7**	**19.8**	**35.3**	**24.9**	**8.5**	**2.8**	**3.2**	**7.8**	**20.9**
MN-v3	DIM	6.3	2.5	99.2*	4.9	6.9	4.5	5.5	16.2	6.0	1.8	2.5	2.3	5.4
	ODIM	24.9	17.3	99.4*	21.2	22.1	22.8	23.2	36.1	20.0	9.6	11.6	11.2	20.0
	SIA	32.7	21.4	98.7*	29.2	14.1	26.1	25.4	24.2	19.1	7.9	10.6	10.3	20.1
	BSR	22.0	12.5	98.9*	18.3	8.7	14.4	13.0	15.9	11.0	4.4	6.5	7.8	12.2
	Admix-DIM	18.0	6.3	99.4*	11.9	14.8	12.1	13.0	32.0	12.5	3.7	5.5	4.4	7.5
	CFM-DIM	34.7	24.9	99.8*	31.6	29.4	33.4	27.8	44.2	28.1	11.8	14.2	11.6	26.5
	ITDS-DIM	**49.5**	**27.3**	**100***	**39.9**	**39.2**	**44.9**	**41.6**	**58.0**	**46.7**	**18.1**	**17.0**	**13.7**	**36.0**
RegN	DIM	24.8	15.2	12.4	92.9*	9.3	12.3	31.4	21.4	3.6	3.6	3.0	13.0	13.6
	ODIM	51.3	43.3	31.2	90.3*	29.2	35.6	56.9	49.3	15.7	12.9	11.6	26.1	33.0
	SIA	71.2	72.1	38.6	95.7*	15.6	48.0	72.0	33.1	19.7	17.9	18.6	47.7	41.3
	BSR	49.0	63.0	22.1	94.1*	6.1	24.9	47.7	15.6	7.0	8.8	8.9	38.0	26.4
	Admix-DIM	39.4	25.2	20.6	92.6*	14.8	23.1	47.7	33.1	7.6	7.1	7.3	22.9	22.6
	CFM-DIM	71.5	61.9	50.8	93.0*	38.3	61.0	72.3	61.6	22.5	21.7	22.1	43.3	47.9
	ITDS-DIM	**82.2**	**50.5**	**68.9**	**99.9***	**64.6**	**73.1**	**84.9**	**75.6**	**56.5**	**43.0**	**40.8**	**54.7**	**63.2**

Table 3. TASRs (%) on twelve pre-trained models with various transformation-based attacks. The AEs are crafted on the RN50, VGG16, MN-v3, and RegN models, respectively. An asterisk (*) indicates white-box attacks. ITDS-DIM represents the combination of ITDS, which integrates S1, with DIM, and the boldface represents the results of this combination.

bility of the two strategies in SDS, ITDS is implemented in two forms: ITDS$_{fs}$ for S1 (feature space) and ITDS$_{gs}$ for S2 (geometric space).

It can be seen that, compared to other competition-based attacks, ITDS demonstrates an overwhelming advantage. Taking ITDS$_{gs}$ as an example, its average black-box TASR (BAvg.) is 9.0%, 7.0%, 3.8%, and 12.4% higher than the existing SOTA method CFM on RN50, VGG16, MN-v3, and RegN, respectively. From the perspective of cross-architecture models, the largest gaps in average black-box

Attack	RN50	VGG16	MN-v3	RegN	Inc-v3	RN101	DN161	EffN	ViT	DeiT	ConViT	PiT	BAvg.
ITDS$_I$	100*	41.5	24.3	40.4	18.9	84.6	70.2	11.5	5.5	2.9	2.7	4.2	27.9
ITDS$_{I+S}$	100*	47.6	34.8	50.8	26.9	90.1	81.5	15.4	9.3	3.7	4.7	6.4	33.7
ITDS$_I$-DIM	99.9*	71.1	58.6	73.3	56.8	92.5	87.1	48.4	35.1	18.8	19.1	23.6	53.1
ITDS$_{I+S}$-DIM	**100***	**79.3**	**70.3**	**82.8**	**69.6**	**95.6**	**94.6**	**62.0**	**44.4**	**23.6**	**24.7**	**31.6**	**61.7**

Table 4. Evaluation of the effectiveness of two modules in ITDS, where 'I' represents ITC and 'S' represents SDS. The AEs are crafted in RN50. An asterisk (*) indicates white-box attacks. Boldface represents the best results. Note that single SDS requires combination with existing attack methods, which we discuss in the appendix.

TASR between it and CFM on CNNs and ViTs are 13.9% and 9.7% on RegN, respectively. It should be noted that the white-box attack performance of our ITDS is also significantly superior, being the best in terms of white-box TASR in any surrogate model, especially on VGG16 and RegN, where it is 13.9% and 7.6% higher than CFM, respectively. In the comparison between ITDS$_{fs}$ and ITDS$_{gs}$, the results seem to be evenly matched, with each showing their own merits in the three trials. Therefore, we believe that both strategies are viable, which precisely proves the effectiveness and diversity of the proposed SDS principle.

4.3. Evaluation on Transformation-based Attacks

Currently, transformation-based methods are the main approach to adversarial attacks and are the most effective way to improve the transferability of attacks. Numerous studies [2, 4, 16] have shown that non-transformation-based methods combined with DIM can further enhance the transferability of crafted AEs. This combination approach has become a common standard in industry, frequently compared to transformation-based methods for fairness. In this section, we evaluate the TASR of transformation-based and DIM-combined competition-based adversarial attacks. The former includes DIM, ODIM, SIA, and BSR, while the latter includes Admix-DIM, CFM-DIM, and ITDS-DIM. It should be noted that, due to the similar performance of S1 and S2 in SDS, in all subsequent experiments involving ITDS-DIM, we selected ITDS$_{fs}$ as the representative to combine with DIM. The results are shown in Table 3.

It can be observed that all competition-based methods show a significant improvement in attack performance after being combined with DIM. Compared to the strongest competition-based attack CFM-DIM, ITDS-DIM achieved average black-box TASR improvements of 15.7%, 13.7%, 9.5%, and 15.3% on RN50, VGG16, MN-v3, and RegN, respectively. The maximum average black-box TASR gap between the two on CNNs and ViTs is 16.1% on RN50 and 21.4% on RegN. Even when compared with the SOTA transformation-based attacks, our method is not inferior, and even surpasses them in all aspects. Compared to the strongest transformation-based attack SIA, ITDS-DIM achieved average black-box TASR improvements of 15.9% and 21.9% on MN-v3 and RegN, respectively. The maximum average black-box TASR gap between the two on CNNs and ViTs is 21.3% and 22.8% on RegN, respectively.

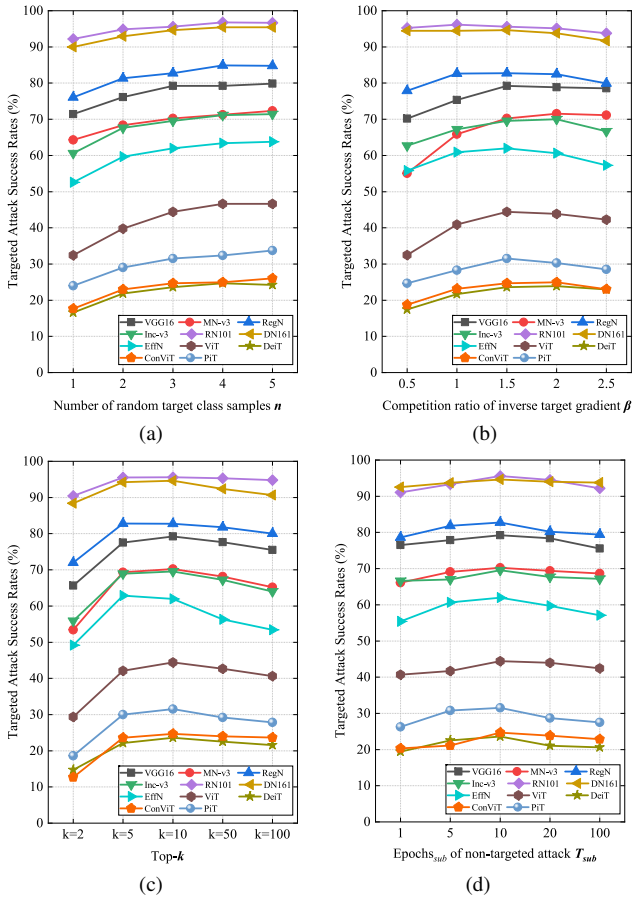

Figure 4. Ablation studies on the RN50 model. (a) - (d): TASRs (%) on the other eleven models with the adversaries crafted by ITDS-DIM, where the default value for n, β, k and T_{sub} are set to 3, 1.5, 10 and 10 respectively, when test parameters for each other.

4.4. Ablation Studies

We first present a series of ablation studies conducted on RN50 to assess the impact of each parameter in ITDS on attack performance, including the number of random target class samples n, the competitive ratio of the inverse target gradient β, the value of k in top-k, and the number of epochs$_{sub}$ for non-targeted attack T_{sub}. In all ablation experiments, unless otherwise specified, all parameters other

Ablation		Target Model												
Norm (P1)	MI (P2)	RN50	VGG16	MN-v3	RegN	Inc-v3	RN101	DN161	EffN	ViT	DeiT	ConViT	PiT	BAvg.
	✓	7.1*	0.5	1.0	0.4	0.4	0.8	0.3	0.2	0.2	0.2	0.1	0.1	0.4
✓		100*	72.8	66.1	75.8	65.1	90.1	88.2	54.7	37.2	16.9	17.4	23.7	55.3
✓	✓	100*	59.7	41.3	66.1	52.8	90.9	91.3	45.1	27.8	14.5	16.2	21.1	47.9
		100*	**79.3**	**70.3**	**82.8**	**69.6**	**95.6**	**94.6**	**62.0**	**44.4**	**23.6**	**24.7**	**31.6**	**61.7**

Table 5. TASRs (%) of ITDS-DIM by ablating inner modules of the ITC. The AEs are crafted in RN50. An asterisk (*) indicates white-box attacks. In which Norm and MI correspond to P1 and P2 in Section 3.1.2, respectively. Boldface represents the best results.

Attack	JPEG	FD	SO	GDMP	Def_4	Def_8	Def_{l2}
SIA	83.2	89.8	44.3	1.5	39.0	14.0	54.7
Admix-DIM	80.0	92.9	47.7	1.9	34.8	14.5	53.1
CFM-DIM	82.5	92.6	51.5	4.1	25.2	8.7	45.5
ITDS-DIM	**99.4**	**99.9**	**91.3**	**31.2**	**50.2**	**25.6**	**66.8**

Table 6. TASRs (%) of breaking-through four defense methods and three adversarial training defense models. The results for defense methods are based on the average performance of AEs crafted on RN50, VGG16, MN-v3, and RegN respectively, while the results for defense models are based on the performance of AEs crafted by the corresponding models. Bolder for the best.

than the test parameters are fixed according to the experimental setup.

As shown in Figure 4a, TASR increases with n, as the diversity of target class images directly affects attack performance. However, the curve levels off after $n = 3$, and considering computational costs, we selected $n = 3$. In Figure 4b, most models achieve optimal performance at $\beta = 1.5$, with a few exceptions (e.g., MN-v3). We believe a moderate competitive gradient provides an "exercise" effect, while an overly strong gradient disrupts input features, reducing attack performance. In Figure 4c, optimal performance generally occurs at $k = 5$ or $k = 10$. We posit that an overly small k under-covers classes and fails to enlarge global margins, whereas an overly large one blurs the optimization. At $k = 10$, cross-architecture transferability is even better. In Figure 4d, evaluating epochs$_{sub}$ (T_{sub}) for non-targeted attacks, we found that larger values do not always lead to better performance. As discussed in the ITC section, inverse target gradient perturbations generated at each T_{sub} have diverse potential competitive features, boosting attack performance. At $T_{sub} = 1$, though it involves more target class samples, it lacks deeper latent features of the target class, limiting attack performance. Beyond a certain T_{sub} value, missing target class features reduce the diversity of inverse gradients, causing performance to regress.

Furthermore, to verify the effectiveness of the ITC and SDS components in ITDS, we conducted experimental evaluations as shown in Table 4. The results indicate that ITDS using only ITC has already surpassed the attack performance of CFM in Table 2. Regardless of whether DIM is combined or not, SDS can bring additional performance improvements to ITDS. These findings collectively support the key roles of ITC and SDS in ITDS.

To verify the impact of internal modules within the ITC on attack performance, we conducted ablation experiments as shown in Table 5. It can be observed that the use of N-norm may limit the scale of gradient updates, which could result in generated AEs lacking sufficient aggressiveness. On the contrary, if we do not use N-norm (P1), gradient updates will gain greater freedom, which helps to explore the model's decision boundary more deeply, and may thus produce more aggressive AEs. Simultaneously, momentum gradients, as a technique for accumulating previous gradients, will weaken the intensity of inverse target features in the current iteration. Since we wish to maintain the integrity of inverse target features and avoid any normalization, the inverse target momentum gradients in each round will compete with the AE when combined with P1, leading to excessive deformation of AE features, which is disastrous for implementing precisely targeted attacks.

4.5. Evaluation on Advanced Defenses

In this section, we evaluate the capabilities of various attacks to break through different defense mechanisms. As shown in Table 6, our method exhibits exceptional robustness against a wide range of defense strategies and models. For example, when the Score-OPT method is employed, the average TASR of our approach is nearly 40% higher than that of CFM-DIM. Furthermore, our method achieves the best performance in all three defense models evaluated.

5. Conclusion

This paper introduces a novel method named ITDS, which significantly improves the transferability of targeted AEs by incorporating the ITC mechanism and the SDS strategy. Experimental results demonstrate that ITDS achieves a substantial increase in average transferable TASRs across various mainstream CNN and ViT models, outperforming SOTA competition-based and transformation-based methods. Our ablation studies further confirmed the key roles of the ITC and SDS components in enhancing attack performance. Moreover, ITDS exhibits superior effectiveness in countering multiple defense methods and models. These findings confirm the effectiveness and practicality of ITDS in enhancing the transferability of targeted AEs and provide a new perspective for the field of adversarial research.

Acknowledgments

This work was supported by funding from the National Natural Science Foundation of China (NSFC) under Grants 62272486, 62172154, and 62372473, the Natural Science Foundation of Hunan Province under Grant 2023JJ70016, the Hong Kong Research Grants Council (RGC) General Research Fund under Grants 152244/21E, 152169/22E, 152228/23E, and 162161/24E, the Research Impact Fund under Grant R5011-23F, the Collaborative Research Fund under Grant C1042-23GF, the NSFC/RGC Collaborative Research Scheme under Grant CRS_HKUST602/24, the Areas of Excellence Scheme under Grant AoE/E-601/22-R, and the InnoHK initiative (HKGAI). Professor Shigeng Zhang is the corresponding author of this paper.

References

[1] Junyoung Byun, Seungju Cho, Myung-Joon Kwon, Hee-Seon Kim, and Changick Kim. Improving the transferability of targeted adversarial examples through object-based diverse input. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15244–15253, 2022. 1, 2, 3, 5

[2] Junyoung Byun, Myung-Joon Kwon, Seungju Cho, Yoonji Kim, and Changick Kim. Introducing competition to boost the transferability of targeted adversarial examples through clean feature mixup. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24648–24657, 2023. 1, 3, 5, 7

[3] Yinpeng Dong, Fangzhou Liao, Tianyu Pang, Hang Su, Jun Zhu, Xiaolin Hu, and Jianguo Li. Boosting adversarial attacks with momentum. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 9185–9193, 2018. 4

[4] Yinpeng Dong, Tianyu Pang, Hang Su, and Jun Zhu. Evading defenses to transferable adversarial examples by translation-invariant attacks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4312–4321, 2019. 7

[5] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 5

[6] Stéphane d'Ascoli, Hugo Touvron, Matthew L Leavitt, Ari S Morcos, Giulio Biroli, and Levent Sagun. Convit: Improving vision transformers with soft convolutional inductive biases. In *International Conference on Machine Learning*, pages 2286–2296. PMLR, 2021. 5

[7] Goodfellow et al. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014. 1, 2

[8] Chuan Guo, Mayank Rana, Moustapha Cisse, and Laurens Van Der Maaten. Countering adversarial images using input transformations. *arXiv preprint arXiv:1711.00117*, 2018. 5

[9] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 5

[10] Byeongho Heo, Sangdoo Yun, Dongyoon Han, Sanghyuk Chun, Junsuk Choe, and Seong Joon Oh. Rethinking spatial dimensions of vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11936–11945, 2021. 5

[11] Andrew Howard, Mark Sandler, Grace Chu, Liang-Chieh Chen, Bo Chen, Mingxing Tan, Weijun Wang, Yukun Zhu, Ruoming Pang, Vijay Vasudevan, et al. Searching for mobilenetv3. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1314–1324, 2019. 5

[12] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708, 2017. 5

[13] Kurakin et al. Adversarial examples in the physical world. In *Artificial intelligence safety and security*, pages 99–112. Chapman and Hall/CRC, 2018. 2

[14] Zhankai Li, Weiping Wang, Jie Li, Kai Chen, and Shigeng Zhang. Foolmix: Strengthen the transferability of adversarial examples by dual-blending and direction update strategy. *IEEE Transactions on Information Forensics and Security*, 2024. 1

[15] Zhankai Li, Weiping Wang, Jie Li, Kai Chen, and Shigeng Zhang. Ucg: A universal cross-domain generator for transferable adversarial examples. *IEEE Transactions on Information Forensics and Security*, 2024. 1

[16] Jiadong Lin, Chuanbiao Song, Kun He, Liwei Wang, and John E Hopcroft. Nesterov accelerated gradient and scale invariance for adversarial attacks. In *International Conference on Learning Representations*, 2020. 7

[17] Zihao Liu, Qi Liu, Tao Liu, Nuo Xu, Xue Lin, Yanzhi Wang, and Wujie Wen. Feature distillation: Dnn-oriented jpeg compression against adversarial examples. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 860–868. IEEE, 2019. 5

[18] Yuyang Long, Qilong Zhang, Boheng Zeng, Lianli Gao, Xianglong Liu, Jian Zhang, and Jingkuan Song. Frequency domain model augmentation for adversarial attack. In *European Conference on Computer Vision*, pages 549–566. Springer, 2022. 1, 2

[19] Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, and Pascal Frossard. Deepfool: a simple and accurate method to fool deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2574–2582, 2016. 5

[20] Ilija Radosavovic, Raj Prateek Kosaraju, Ross Girshick, Kaiming He, and Piotr Dollár. Designing network design spaces. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10428–10436, 2020. 5

[21] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large

scale visual recognition challenge. *International journal of computer vision*, 115:211–252, 2015. 5

[22] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 5

[23] C Szegedy. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013. 1

[24] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826, 2016. 5

[25] Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning*, pages 6105–6114. PMLR, 2019. 5

[26] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. In *International conference on machine learning*, pages 10347–10357. PMLR, 2021. 5

[27] Francisco Utrera, Evan Kravitz, N Benjamin Erichson, Rajiv Khanna, and Michael W Mahoney. Adversarially-trained deep nets transfer better: Illustration on image classification. *arXiv preprint arXiv:2007.05869*, 2021. 5

[28] Jinyi Wang, Zhaoyang Lyu, Dahua Lin, Bo Dai, and Hongfei Fu. Guided diffusion model for adversarial purification. *arXiv preprint arXiv:2205.14969*, 2022. 5

[29] Kunyu Wang, Xuanran He, Wenxuan Wang, and Xiaosen Wang. Boosting adversarial transferability by block shuffle and rotation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24336–24346, 2024. 1, 2, 5

[30] Xiaosen Wang, Xuanran He, Jingdong Wang, and Kun He. Admix: Enhancing the transferability of adversarial attacks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 16158–16167, 2021. 1, 2, 5

[31] Xiaosen Wang, Zeliang Zhang, and Jianping Zhang. Structure invariant transformation for better adversarial transferability. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4607–4619, 2023. 2, 5

[32] Zhibo Wang, Hengchang Guo, Zhifei Zhang, Wenxin Liu, Zhan Qin, and Kui Ren. Feature importance-aware transferable adversarial attacks. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 7639–7648, 2021. 2

[33] Juanjuan Weng, Zhiming Luo, Shaozi Li, Nicu Sebe, and Zhun Zhong. Logit margin matters: Improving transferable targeted adversarial attack by logit calibration. *IEEE Transactions on Information Forensics and Security*, 18:3561–3574, 2023. 2

[34] Cihang Xie, Zhishuai Zhang, Yuyin Zhou, Song Bai, Jianyu Wang, Zhou Ren, and Alan L Yuille. Improving transferability of adversarial examples with input diversity. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2730–2739, 2019. 1, 2, 5

[35] Zeng et al. Enhancing targeted transferability via feature space fine-tuning. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4475–4479. IEEE, 2024. 3

[36] Boya Zhang, Weijian Luo, and Zhihua Zhang. Enhancing adversarial robustness via score-based optimization. *Advances in Neural Information Processing Systems*, 36:51810–51829, 2023. 5

[37] Qilong Zhang, Xiaodan Li, YueFeng Chen, Jingkuan Song, Lianli Gao, Yuan He, et al. Beyond imagenet attack: Towards crafting adversarial examples for black-box domains. In *International Conference on Learning Representations*, 2022. 1

[38] Yaoyuan Zhang, Yu-an Tan, Tian Chen, Xinrui Liu, Quanxin Zhang, and Yuanzhang Li. Enhancing the transferability of adversarial examples with random patch. In *IJCAI*, pages 1672–1678, 2022. 2

[39] Zhao et al. On success and simplicity: A second look at transferable targeted attacks. *Advances in Neural Information Processing Systems*, 34:6115–6128, 2021. 2, 3

[40] Rongyi Zhu, Zeliang Zhang, Susan Liang, Zhuo Liu, and Chenliang Xu. Learning to transform dynamically for better adversarial transferability. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24273–24283, 2024. 1