Leveraging Spatial Invariance to Boost Adversarial Transferability

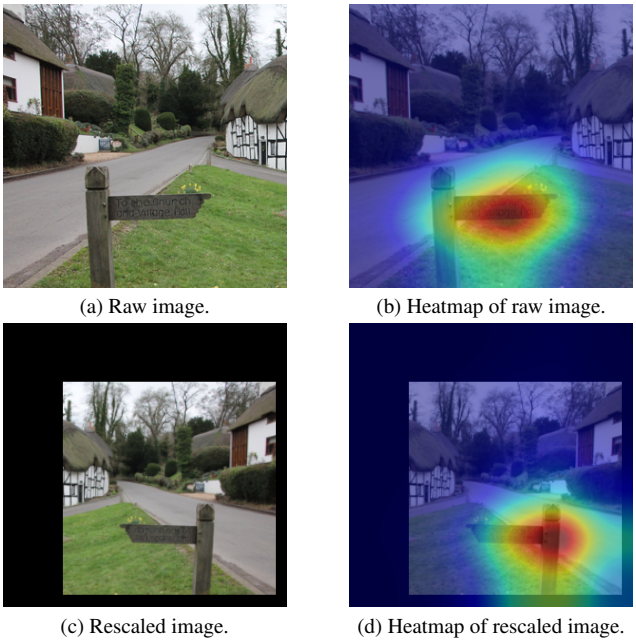
Zihan Zhou[1] Li Li[1] Yanli Ren[1] Chuan Qin[2] Guorui Feng[1,*]
Shanghai University[1] University of Shanghai for Science and Technology[2]
{zhouzihan, llichn, renyanli, grfeng}@shu.edu.cn, qin@usst.edu.cn

Abstract

Adversarial examples, crafted with imperceptible perturbations, reveal a significant vulnerability of Deep Neural Networks (DNNs). More critically, the transferability of adversarial examples allows attackers to induce unreasonable predictions without requiring knowledge about the target model. DNNs exhibit spatial invariance, meaning that the position of an object does not affect the classification result. However, existing input transformation-based adversarial attacks solely focus on behavioral patterns at a singular position, failing to fully exploit the spatial invariance exhibited by DNNs across multiple positions, thus constraining the transferability of adversarial examples. To address this, we propose a multi-scale, multi-position input transformation-based attack called Spatial Invariance Diversity (SID). Specifically, SID uses hybrid spatial-spectral fusion mechanisms within localized receptive fields, followed by multi-scale spatial downsampling and positional perturbations via random transformations, thereby crafting an ensemble of inputs to activate diverse behavioral patterns of DNNs for effective adversarial perturbations. Extensive experiments on the ImageNet dataset demonstrate that SID could achieve better transferability than the current state-of-the-art input transformation-based attacks. Additionally, SID can be flexibly integrated with other input transformation-based or gradient-based attacks, further enhancing the transferability of adversarial examples. The code is available at https://github.com/TheMoss7/SID.

1. Introduction

With the continuous advancement of Deep Neural Networks (DNNs) [8, 11, 12, 30], they have been successfully applied to various fields *e.g.*, image classification [11], semantic segmentation [4, 37], and face recognition [25, 31], demonstrating remarkable performance. However, the vast number of parameters within DNNs makes them difficult to interpret, leading to a black-box nature. Recent studies [9, 26] have shown that adding imperceptible perturbations to the legitimate input can effectively mislead DNNs. The emergence of adversarial examples has further heightened concerns regarding the reliability of DNNs.



(a) Raw image. (b) Heatmap of raw image. (c) Rescaled image. (d) Heatmap of rescaled image.

Figure 1. The attention heatmaps of raw image, and rescaled image generated on Inception-v3 model using Grad-CAM.

Adversarial attacks are typically divided into white-box [9, 13, 20] and black-box attacks [1, 3]. In white-box attacks, the architecture and parameters of the target model are accessible. Therefore, attackers can easily launch attacks on the white-box target model. However, the white-box attack scenario is overly idealized, as the target model is typically inaccessible. In contrast, black-box attacks are more realistic, where the attacker has no knowledge of the target model. Recently, many studies have found that adversarial examples generated through white-box attacks exhibit transferability [2, 32], allowing them to successfully attack other models with different architectures and parameters. It is entirely feasible to use a surrogate model to conduct

* Corresponding author.

black-box attacks. Therefore, improving the transferability of adversarial examples has become an important issue.

Recently, many studies have explored different approaches to improve the adversarial transferability, including input transformation [33, 43, 45], model ensemble [18, 41], gradient manipulation [6, 17, 32], and improving adversarial loss function [42]. Among these, input transformation methods have shown great potential. By applying random transformation to input images before gradient calculation, this approach captures more diverse gradients, significantly enhancing the adversarial transferability. DNNs exhibit spatial invariance [14], consistently achieving correct classification regardless of changes in object positions within the image. As illustrated in Figure 1, while the pre-trained model's attention heatmap [23] maintains focus on the correct area, the target model demonstrates robust recognition capability by accurately identifying objects even when images are reduced in size and randomly repositioned. This indicates that there are weights associated with the relevant category in the different regions. However, we observe that, when computing adversarial perturbations, existing methods usually keep the position of correct objects unchanged and fail to activate the DNNs' behavioral patterns at multiple positions, leading to a lack of diversity in the perturbations. We believe this oversight limits the diversity of perturbations, which causes lower adversarial transferability.

Motivated by the spatial invariance of DNNs, we aim to activate different behavioral patterns at different spatial positions and scales. Therefore, in this work, we propose a novel input transformation-based attack, called Spatial Invariance Diversity (SID). Specifically, we use hybrid spatial-spectral fusion mechanisms within localized receptive fields. The resulting image is then downsampled to multiple scales, with the resampled images placed in random positions via padding to craft transformed images for gradient calculation. This approach preserves the overall content of the image while fusing the local image blocks with the global input image to activate more behavioral patterns. In summary, our contributions are as follows:

- We design a new image transformation method that no longer focuses on a single position or quantity of image content. While ensuring the global semantics, spatial-frequency domain self-enhancement is applied at multiple scales and positions to generate more diverse images.
- We propose SID, which leverages the spatial invariance of DNNs to achieve improved transferability of adversarial examples by activating the different behavioral patterns of DNNs at different spatial scales and positions.
- Experiments conducted on the ImageNet dataset demonstrate that our SID exhibits superior transferability compared with the state-of-the-art input transformation-based attacks.

2. Related Work

2.1. Adversarial Attacks

Since the discovery of adversarial examples by Szegedy *et al*. [26], many studies [2, 6, 33] have been proposed, continuously highlighting the vulnerability of DNNs. Among them, adversarial attacks based on Fast Gradient Sign Method (FGSM) [9] have proven to be one of the most effective approaches. FGSM is a simple and fast white-box attack method. Subsequently, Kurakin *et al*. introduce an iterative version, I-FGSM [13], which improves the effectiveness of adversarial examples in white-box attacks.

Gradient-based attack methods improve the transferability of adversarial examples by operating on gradients. Dong *et al*. [6] incorporate the concept of momentum from optimization algorithms into I-FGSM, proposing MI-FGSM, successfully improve the transferability. Lin *et al*. [17] propose the use of Nesterov accelerated gradient to maintain the gradient direction. Wang *et al*. [34] introduce variance tuning, which adjusts the current gradient by considering the gradient variance from previous iterations.

In addition to gradient-based methods, input transformation-based attacks [19, 33, 35, 36, 43] can also significantly enhance the transferability of adversarial examples. Diverse Input Method (DIM) [40] randomly pads the image before calculating gradients through the model. Translation Invariant Method (TIM) [7] leverages translation invariance by approximating the gradient of the shifted image using pre-defined convolutional kernels. Spectrum Simulation Attack (SSA) [19] applies Gaussian noise in the spatial domain and random masks in the frequency domain. Structure Invariant Attack (SIA) [36] applies random augmentations to each image block, creating more diverse images without compromising their content. Block Shuffle and Rotation (BSR) [33] disrupts the attention heatmap of the model through image block shuffle and rotation, reducing variance in attention across different models.

2.2. Adversarial Defenses

To address the threat posed by adversarial attacks, many defense methods have been introduced to reduce model vulnerabilities to adversarial examples. Liao *et al*. [16] introduce a High-level representation Guided Denoiser (HGD), an adversarial purifier based on the U-Net architecture. Xie *et al*. [39] employ random resizing and padding to counteract the effects of adversarial examples. Guo *et al*. [10] utilize multiple input image transformations, such as JPEG compression, to counter adversarial examples. Nie *et al*. [22] utilize a diffusion architecture model to remove adversarial perturbations. Naseer *et al*. [21] use a self-supervised mechanism to train a Neural Representation Purifier (NRP) to reduce the impact of adversarial perturbations. Cohen *et*

al. [5] propose that Randomized Smoothing (RS) enables the training of a robust ImageNet classifier with a strong robustness guarantee. In addition to pre-processing inputs, enhancing model robustness is another effective strategy to combat adversarial examples. One of the most effective methods is adversarial training, which involves using adversarial examples during the training process. Goodfellow *et al*. [9] incorporate adversarial examples into the training of classification models on MNIST. Tramèr *et al*. [29] introduce ensemble adversarial training, using adversarial examples generated on several models, showing great robustness against adversarial attack. Wong *et al*. [38] explore an efficient approach to adversarial training by using a weaker and cheaper adversary at a significantly reduced cost.

3. Methodology

3.1. Preliminaries

For a classification model f with parameters θ serving as the victim model, and a clean input image x with the ground-truth label y, the attacker's goal is to craft an adversarial example $x_{adv} = x + \delta$ that can mislead the victim model, *i.e*., $f(x_{adv}; \theta) \neq y$, where δ represents the adversarial perturbation. To ensure the stealth of the attack, the adversarial perturbation should satisfy the L_p-norm constraint $\|\delta\|_p \leq \epsilon$. We follow previous studies [33, 36, 40], employing the L_∞-norm for constraints. Therefore, adversarial attack can be expressed as the following optimization problem:

$$\arg\max_{x_{adv}} J(x_{adv}, y; \theta), \quad s.t. \|\delta\|_\infty \leq \epsilon, \tag{1}$$

where $J(\cdot)$ represents the loss function. We use cross-entropy loss function consistent with other input transformation-based methods [19, 33, 36]. For the one-step adversarial attack FGSM, it can be expressed using the following formula:

$$x_{adv} = x + \epsilon \cdot \text{sign}(\nabla_x J(x_{adv}, y; \theta)), \tag{2}$$

where $sign(\cdot)$ is the sign function, which represents the direction of the gradient used as the adversarial perturbation. Furthermore, for the iterative version of FGSM, known as I-FGSM, with a step size denoted as α, the middle iteration can be expressed as follows:

$$x_{adv}^t = \text{clip}_{x,\epsilon}(x_{adv}^{t-1} + \alpha \cdot \text{sign}(\nabla_{x_{adv}^{t-1}} J(x_{adv}^{t-1}, y; \theta))), \tag{3}$$

where $clip(\cdot)$ represents the pixel clip function, which constrains the perturbation. MI-FGSM, combined with input transformation-based attack method, can be expressed using the following formula:

$$\begin{aligned} g_t &= \mu \cdot g_{t-1} + \frac{\nabla_{x_{adv}^{t-1}} J(T(x_{adv}^{t-1}), y; \theta)}{\|\nabla_{x_{adv}^{t-1}} J(T(x_{adv}^{t-1}), y; \theta)\|_1}, \\ x_{adv}^t &= \text{clip}_{x,\epsilon}(x_{adv}^{t-1} + \alpha \cdot \text{sign}(g_t)), \end{aligned} \tag{4}$$

where $g_0 = 0$, μ is the decay factor, and $T(\cdot)$ is the transformation operator from input transformation-based attack. In MI-FGSM, $T(\cdot)$ is the identity mapping. Given that MI-FGSM effectively escapes local optima [6], the proposed method will be integrated into this framework.

3.2. Motivation

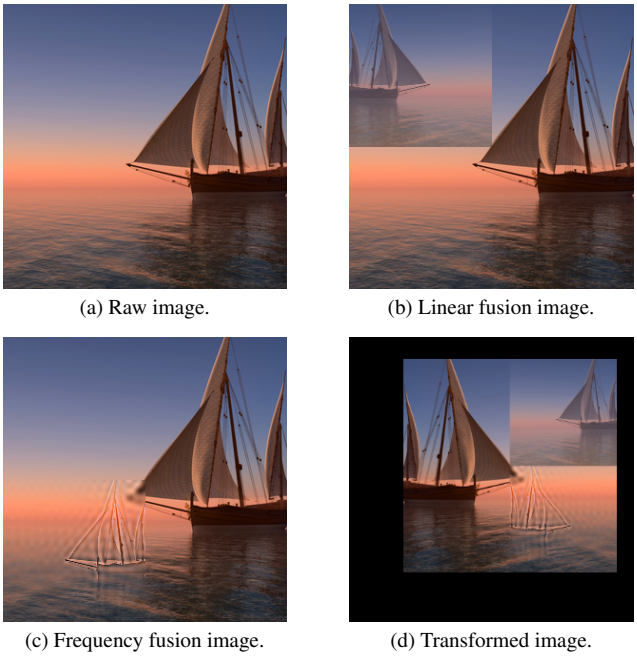
For the image classification task, the models exhibit spatial invariance [14], allowing a well-trained model to classify objects correctly regardless of their position in the image. This is the same as human perception. Considering that the end of DNNs typically consists of fully connected layers, features extracted from inputs in different positions are likely to be similar. Some work [24, 44] have proven that even though the structures and parameters of different models are not the same, the features extracted by the models often possess the same characteristics.

From this premise, we assert that, in a well-trained model, the weights associated with the input image class also exhibit spatial invariance. This means that there are diverse behavioral patterns of DNNs at different positions. However, existing work overlook this point. Most of the work do not change the position of the correct object in the image. Therefore, in the input space, the behavioral patterns at many positions have not been activated. This means that the adversarial perturbations obtained only contain the behavioral pattern of a single position, while ignoring the behavior patterns at other positions.

In this work, we aim to activate diverse behavioral patterns of a single model across spatial positions. Specifically, we achieve this through image transformation. With a specific image transformation method, we hope that the adversarial perturbations can contain various behavioral patterns at different spatial positions to enhance the effectiveness. Meanwhile, some work [19, 33] on adversarial attacks have also demonstrated that the remarkable common characteristics extracted from different models have a greater impact on the transferability of adversarial examples. Based on the analysis, we activate diverse behavioral patterns at different positions to find as many common characteristics as possible, so as to enhance the adversarial transferability.

3.3. Spatial Invariance Diversity

When designing the method of image transformation, the semantic information of the image should be ensured to remain unchanged. For instance, the four limbs of a cat should be positioned below its body. This is also consistent with human perception. Therefore, when utilizing the spatial invariance of the models, the basic structure of the image should not be altered. On this basis, we choose two methods to change the position of the content of the image category. One is random image padding, and the other is image flipping. In this paper, we do not consider



(a) Raw image. (b) Linear fusion image. (c) Frequency fusion image. (d) Transformed image.

Figure 2. The raw and transformed images using proposed input transformation method SID.

changing the size of the image because the size of the input image for some models is fixed. Therefore, we perform downsampling on the image and then conduct random image padding. To make use of the multi-scale information, we correlate the downsampling ratio of the image with the number of intermediate iterations.

Assuming the image is randomly transformed N times, the downsampling ratio for the transformation can be expressed as $(1-\beta\cdot\frac{n}{N})$, where n represents the current iteration number, and the range of β is restricted to between 0 and 1. As the number of iterations increases, the ratio gradually decreases, meaning that the image size also decreases. During the middle iterations, the transformed image exhibits multi-scale features. This approach helps to activate the different types of behavioral patterns of the surrogate model in different scales and positions.

For the image flipping, there are two situations depending on the content of the image. One situation is that the content of the image is not in the central position. In this case, the image flipping operation will change the position of the image, which is consistent with our starting point. The other situation is that the content is right in the center of the image. In this case, although the image flipping does not change the position of the content, it still activates different behavioral patterns of the model. According to the experiment, we find that the improvement of the transferability of adversarial examples by vertical flipping is not as good as that by horizontal flipping. The experiments in work [36] also prove this point. Therefore, in the implementation, we choose horizontal flipping. As shown in Figure 2, we randomly flip the image and image blocks for more behavioral patterns.

3.4. Local Image Fusion

We find that, in previous work, the adversarial perturbations obtained usually contain only one behavioral pattern of the model, because the image contains only an object of the correct class. We consider this approach to be inefficient. To further utilize the property of spatial invariance to obtain more effective adversarial perturbations, we propose local image fusion to make the perturbations contain more behavioral patterns. As shown in Figure 2b and Figure 2c, we fuse the image itself, through the designed method, at random positions of the original image. This approach has two advantages. Firstly, a single image contains the content of the correct class in multiple positions, so that the adversarial perturbation obtained by enhancing the image contains multiple behavioral patterns. Secondly, the local image fusion does not destroy the original structure of the image, ensuring that the semantic information remains unchanged and guaranteeing the usability of the perturbation. Specifically, we design two types of local image fusion methods: linear fusion and frequency fusion.

To ensure the use of two different fusion methods in a single image, we adopt the image block strategy. Specifically, the image is divided into $k\times k$ blocks. For each block, there exists a probability $1-p$ of retaining its original state, while with a probability of p, one of the two fusion methods is randomly selected and applied. For linear fusion, we downsample the input image to match the dimensions of the local image block and then perform a linear combination of the local image block and the rescaled image:

$$\begin{aligned} x_{linear} =& \omega\cdot x_{rescale}+(1-\omega)\cdot x_{block}, \\ & s.t. 0\le\omega\le 1, \end{aligned} \tag{5}$$

where ω is the weight for the linear fusion, $x_{rescale}$ is the rescaled input image, x_{block} is the local image block, x_{linear} is the linear fusion image block. The linear fusion image is shown in Figure 2b. For frequency fusion, we also rescale the input image to match the dimensions of the image blocks. Then we combine the low-frequency content of the image block with the high-frequency content of the rescaled image:

$$\begin{aligned} x_{frequency} = IDCT(HP(DCT(x_{rescale})) \\ +LP(DCT(x_{block}))), \end{aligned} \tag{6}$$

where $DCT(\cdot)$ is discrete cosine transform, $IDCT(\cdot)$ is inverse discrete cosine transform, $HP(\cdot)$ is the high-pass filter, $LP(\cdot)$ is the low-pass filter, $x_{frequency}$ is the frequency

Algorithm 1 Spatial Invariance Diversity

Input: A classifier f with parameters θ; a clean input image x with ground-truth label y; the number of iteration T; the maximum range of the perturbation ϵ; the number of image transformation N; the decay factor μ; the weight of downsampling factor β; the number of image blocks k; the probabilities of image block fusion p; the weight of linear fusion ω.

Output: The adversarial example x_{adv};

$\alpha = \epsilon/T, g_0 = 0, x^0_{adv} = x$;
for $t = 1 \to T$ **do**
  for $n = 1 \to N$ **do**
    Get transformation output: $x'_i = T(x^t_{adv}, \beta, n, k, p, \omega)$
    Calculate gradient: $g'_i = \nabla_{x^t_{adv}} J(x^{'}_i, y; \theta)$
  end for
  Calculate average gradient: $g' = \frac{1}{N}\sum_{i=1}^{N} g'_i$
  Update the momentum: $g_t = \mu \cdot g_{t-1} + \frac{g'}{\|g'\|_1}$
  Update the adversarial example: $x^t_{adv} = \text{clip}_{x,\epsilon}(x^{t-1}_{adv} + \alpha \cdot \text{sign}(g_t))$
end for
return x^T_{adv}.

fusion image block. The frequency fusion image is shown in Figure 2c.

By fusing global information with the local image blocks, the enhanced images activate as many behavioral patterns of the surrogate model related to the input images as possible. Furthermore, regardless of the fusion method used, we strive to retain the consistency of the global image while embedding information, ensuring that its semantic content remains unchanged. For simplicity, we denote the transformation methods introduced in Sec. 3.3 and Sec. 3.4 with the following notation:

$$x_{aug} = T(x, \beta, n, k, p, \omega), \tag{7}$$

where x_{aug} is the transformed image, and $T(\cdot)$ is the final transformation method. The transformed image is shown in Figure 2d. We integrate Equation (7) into MI-FGSM and summarize the algorithm in Algorithm 1.

4. Experiment

4.1. Setup

Dataset. Following previous work [19, 36], we perform evaluation experiments on the ImageNet-compatible dataset containing 1,000 images sampled from ImageNet.

Models. To evaluate the adversarial performance, we utilize nine victim models, which include normally trained models *i.e.*, Inception-v3 [27] (IncV3), Inception-v4 (IncV4), Inception-ResNet-v2 [28] (IncResV2), ResNet-v2-50 (Res50), ResNet-v2-152 (Res152) [11], Vision Transformer (ViT-B) [8], and adversarially trained models (IncV3_{ens3}, IncV3_{ens3}, and IncResV2_{ens}) [29]. To further evaluate the adversarial performance, we select several defense methods, including HGD [16], NRP [21], R&P [39], RS [5], AT [38], DiffPure [22], JPEG [10], and Res-De [15].

Competitors. We select five input transformation-based attack methods as baselines for comparison to demonstrate the effectiveness of the proposed method. We compare SID with two similar methods, DIM [40] and TIM [7]. Additionally, we select the state-of-the-art attack methods in the past three years, including SSA [19], SIA [36], and BSR [33], for comparison. Furthermore, we combine different methods for comparison, *e.g.*, SI-NI-TIM (the combined version of SI-NI-FGSM [17] and TI-FGSM). In the experiments, we combine all methods with MI-FGSM to ensure that all approaches are evaluated on the same baseline.

Parameters Settings. We follow the parameters settings in MI-FGSM, the number of iteration $T = 10$, the maximum perturbation boundary $\epsilon = 16$, the step size $\alpha = \epsilon/T = 1.6$, and the decay factor $\mu = 1$. For TIM, we choose the Gaussian kernel and set the kernel size to 7×7. The transformation probability of DIM [40] is 0.5. For SI-NI-FGSM, the number of copies is set to $m_1 = 5$. For SSA, the tuning factor of the spectrum mask $\rho = 0.5$, and the standard deviation of the spatial noise $\sigma = 1.6$. For SIA, the number of blocks $s = 3$. For BSR, the number of blocks $n = 2$, and the maximum rotation angle $\tau = 24°$. For our SID, the downsampling factor $\beta = 0.1$, the number of blocks $k = 2$, the probabilities of image block fusion $p = 0.5$, the weight of linear fusion $\omega = 0.5$. For SSA, SIA, BSR, and SID, the input image is transformed for $N = 20$ times. The parameters of all compared methods are the default parameters from papers.

4.2. Evaluation on Trained Models

In this section, we evaluate the adversarial performance of various attack methods on six popular models and three adversarially trained models. The adversarial examples are generated using four models, normally trained on ImageNet and provided by PyTorch. Here, we use the attack success rate to evaluate the effectiveness of adversarial examples.

The results are shown in Table 1. For DIM and TIM, DIM demonstrates better transferability on normally trained models, while TIM performs better on adversarially trained models. Compared with them, SSA shows improvements in performance for both white-box and black-box attacks. As the state-of-the-art methods, SIA and BSR demonstrate significant transferability. It can be observed that our SID achieves performance comparable to SIA and BSR on normally trained models, with an average improvement of 3.5% and 4.0%. Notably, our method shows significant improvements on adversarially trained models, achieving an average improvement of 16.8% over SIA and 13.7% over

Model	Attack	IncV3	IncV4	Res50	Res152	IncResV2	ViT-B	IncV3$_{ens3}$	IncV3$_{ens4}$	IncResV2$_{ens}$
IncV3	DIM	99.8*	71.5	63.3	59.7	65.5	23.8	31.6	31.0	17.5
	TIM	**100.0***	52.5	46.6	40.7	45.7	25.6	30.0	29.8	19.7
	SSA	99.5*	88.1	82.3	81.0	85.7	39.2	56.8	56.2	35.7
	SIA	**100.0***	95.2	92.4	88.9	95.2	46.4	61.6	59.9	36.9
	BSR	**100.0***	96.2	91.5	87.1	94.7	47.0	55.0	51.6	29.3
	SID	**100.0***	**97.3**	**93.9**	**92.0**	**95.7**	**58.1**	**77.3**	**73.6**	**52.8**
IncV4	DIM	76.8	99.0*	61.2	56.1	66.0	23.7	28.5	26.0	16.1
	TIM	60.8	99.8*	45.6	41.5	47.9	25.5	28.4	27.3	20.5
	SSA	90.7	99.5*	83.4	82.3	86.5	43.0	58.6	53.4	36.9
	SIA	**97.0**	**99.9***	90.1	87.1	94.0	44.9	56.9	53.8	35.3
	BSR	96.1	**99.9***	85.6	82.3	93.4	34.7	57.6	52.1	34.3
	SID	96.8	**99.9***	**92.1**	**90.5**	**94.5**	**58.4**	**71.7**	**68.1**	**52.4**
IncResV2	DIM	73.9	72.3	64.0	59.5	97.0*	24.2	32.7	30.5	22.1
	TIM	62.5	58.1	54.8	49.3	98.6*	26.9	34.9	31.2	26.4
	SSA	89.9	89.3	86.0	85.0	98.1*	48.7	69.0	63.3	55.4
	SIA	**96.5**	**95.8**	92.9	90.4	**99.7***	48.6	69.7	64.8	51.2
	BSR	94.6	93.8	92.4	90.7	98.5*	51.9	71.4	63.1	51.0
	SID	96.1	95.7	**94.0**	**92.9**	**99.7***	**65.4**	**84.0**	**79.4**	**68.9**
Res50	DIM	79.0	76.4	79.1*	73.6	68.2	23.8	28.3	28.2	16.1
	TIM	60.9	53.2	61.5*	53.4	45.6	27.1	28.7	30.2	20.4
	SSA	88.2	86.3	92.2*	87.1	81.1	38.5	48.3	46.5	31.8
	SIA	94.8	95.6	95.7*	93.8	90.4	38.5	50.3	45.1	27.2
	BSR	96.8	**97.6**	96.6*	95.3	94.2	50.0	69.6	63.9	46.3
	SID	**97.2**	**97.6**	**97.3***	**97.2**	**94.9**	**62.9**	**78.3**	**72.4**	**54.0**
ViT-B	DIM	49.3	43.4	47.5	44.0	36.1	99.8*	47.0	51.5	23.8
	TIM	42.1	34.9	37.9	34.7	27.2	**100.0***	39.1	41.0	27.1
	SSA	66.4	62.2	67.9	63.9	55.6	99.8*	62.5	63.1	38.5
	SIA	80.1	75.9	79.5	76.9	69.3	**100.0***	79.5	81.1	38.5
	BSR	78.1	73.9	75.6	73.6	68.3	99.0*	80.2	82.6	50.0
	SID	**81.0**	**76.5**	**82.4**	**78.2**	**73.0**	99.8*	**81.9**	**86.5**	**62.9**

Table 1. The attack success rates (%) on nine pre-trained models. The adversarial examples are crafted on IncV3, IncV4, IncResV2, and Res50, respectively. * indicates white-box attacks.

Attack	IncV4	Res50	Res152	IncResV2	ViT-B	IncV3$_{ens3}$	IncV3$_{ens4}$	IncResV2$_{ens}$	AVG.
SSA-NI	84.3	80.5	76.0	82.8	36.0	36.5	35.3	19.0	56.3
SIA-NI	**94.9**	89.3	**87.4**	**93.5**	38.9	51.3	49.0	28.2	66.5
BSR-NI	91.9	86.1	81.9	89.1	41.3	55.3	54.3	33.8	66.7
SID-NI	94.8	**90.3**	**87.4**	92.8	**53.5**	**67.0**	**63.2**	**41.6**	**73.8**
TI-DIM	71.3	60.4	55.5	64.8	29.9	42.7	42.0	29.0	49.4
SSA-TI-DIM	92.1	87.6	87.0	90.6	59.8	81.8	81.6	69.8	81.2
SIA-TI-DIM	**97.1**	92.4	90.3	94.9	61.3	82.5	78.8	65.9	82.9
BSR-TI-DIM	95.2	86.6	82.3	92.9	52.7	74.2	70.7	50.0	75.6
SID-TIM	95.8	89.9	87.6	92.4	64.3	83.9	81.7	70.1	83.2
SID-SI-TIM	96.3	**93.6**	**92.1**	**95.8**	**75.5**	**90.9**	**89.6**	**80.3**	**89.2**

Table 2. The attack success rates (%) of black-box attacks on eight pre-trained models. The adversarial examples are crafted on IncV3 with the gradient-based method and various input transformations.

BSR. This demonstrates the effectiveness of SID.

4.3. Evaluation on Ensemble Attacks

FGSM-based adversarial attacks can be flexibly combined, and we show the results of different combinations in Table 2. The momentum method is replaced with Nesterov accelerated gradient, denoted as SSA-NI, SIA-NI, BSR-NI, and SID-NI. Additionally, we combine SID with TIM and SIM, denoted as SID-TIM and SID-SI-TIM, to compare TI-DIM, SSA-TI-DIM, SIA-TI-DIM, and BSR-TI-DIM.

When combined with Nesterov accelerated gradient, SIA-NI demonstrates competitive performance on the CNN architecture models, closely matching that of our proposed SID-NI. BSR-NI demonstrates superior performance relative to SIA-NI on ViT-B and adversarially trained models, though it still falls short of SID-NI by 7.8% to 12.2%. On average, our SID-NI achieves a performance improvement of 7.1% and 7.3%, compared with BSR-NI and SIA-NI. Given that DIM also manipulates image size similarly to SID, we combine it with the comparative methods for fairness. It is noteworthy that SIA-TI-DIM demonstrates strong performance. However, in the case of adversarially trained models, SID-TIM shows a consistent improvement. On average, SID-TIM without DIM outperforms SIA-TI-DIM by 0.3%. Furthermore, SID-SI-TIM, which incorporates SIM, achieves a significant improvement of 6.3% in average.

Attack	IncV3	IncV4	Res50	Res152	IncResV2	ViT-B	IncV3$_{ens3}$	IncV3$_{ens4}$	IncResV2$_{ens}$
DIM	97.3*	93.5*	74.3*	71.2	88.9*	41.8	55.0	52.6	41.7
TIM	97.4*	95.7*	89.0*	86.3	94.1*	41.0	54.8	52.4	36.0
SSA	97.6*	97.5*	96.0*	94.8	96.0*	65.8	83.5	81.3	70.7
SIA	**100.0***	99.9*	98.8*	98.6	99.7*	72.6	89.4	84.3	69.6
BSR	99.9*	**100.0***	98.7*	98.1	**99.8***	72.2	92.1	88.9	73.7
SID	99.9*	99.9*	**98.9***	**99.0**	99.6*	**86.5**	**95.8**	**93.8**	**86.1**
SSA-TI-DIM	98.6*	98.6*	96.6*	96.2	97.6*	85.7	94.1	93.5	90.3
SIA-TI-DIM	99.9*	99.7*	**98.5***	97.3	99.6*	85.5	96.1	95.2	91.6
BSR-TI-DIM	99.9*	99.9*	97.3*	97.3	99.8*	77.2	95.5	93.2	86.6
SID-TIM	99.9*	99.8*	98.0*	98.0	99.4*	89.6	96.8	96.3	93.9
SID-SI-TIM	**100.0***	**100.0***	**98.5***	**98.9**	**99.9***	**93.1**	**98.6**	**97.6**	**95.9**

Table 3. The attack success rates (%) on nine pre-trained models under ensemble model setting with various input transformations. The adversarial examples are crafted on IncV3, IncV4, IncResV2, and Res50. * indicates white-box attacks.

Method	R&P	NRP	HGD	AT	Res-De	DiffPure	JPEG	RS	AVG.
SSA-TI-DIM	92.0	73.0	90.3	53.8	96.5	59.2	91.5	79.3	79.5
SIA-TI-DIM	93.9	53.4	96.3	52.3	**99.3**	48.4	92.4	77.4	76.7
BSR-TI-DIM	91.5	47.3	95.5	51.9	98.5	43.6	90.8	76.7	74.5
SID-TIM	94.5	63.9	96.5	54.4	98.4	59.4	92.7	79.9	80.0
SID-SI-TIM	**96.7**	**78.9**	**97.3**	**57.2**	98.4	**70.6**	**93.8**	**82.2**	**84.3**

Table 4. The attack success rates (%) on eight defense methods. The adversarial examples are crafted on IncV3, IncV4, IncResV2, and Res50 simultaneously.

4.4. Evaluation on Ensemble Models

Ensemble model attack [18, 41] is an effective method to enhance the adversarial transferability. We utilize IncV3, IncV4, Res50, and IncResV2 to generate adversarial examples. Additionally, we discuss two situations: single input transformation and ensemble input transformation.

As shown in Table 3, for the single input transformation, SIA and BSR exhibit comparable performance, with our SID slightly outperforming both attack methods on normally trained CNN architecture models. However, on ViT-B and adversarially trained models, our SID shows an improvement of 4.9% to 14.3% over the best-performing BSR. On average, SID outperforms SIA and BSR by 5.2% and 4.0%, respectively. For the ensemble input transformation, SIA-TI-DIM performs excellently across all models, with its average performance surpassing that of other attack methods. In contrast, SID-TIM, which integrates only TIM, still outperforms SIA-TI-DIM. Notably, SID-TIM shows a 1.0% improvement on average. Besides, SID-SI-TIM achieves a 98.1% attack success rate on average, showing an improvement of over 2.2% compared with other attack methods. This further demonstrates the significant effectiveness of SID in improving the adversarial transferability.

4.5. Evaluation on Defense Methods

We have demonstrated the significant effectiveness of our SID against adversarially trained models in Secs. 4.2 to 4.4. To further evaluate its effectiveness, we assess the attack performance of SID against eight different defense methods. The results are shown in Table 4. In light of the previous results, we combine input transformations with ensemble attacks to achieve better performance.

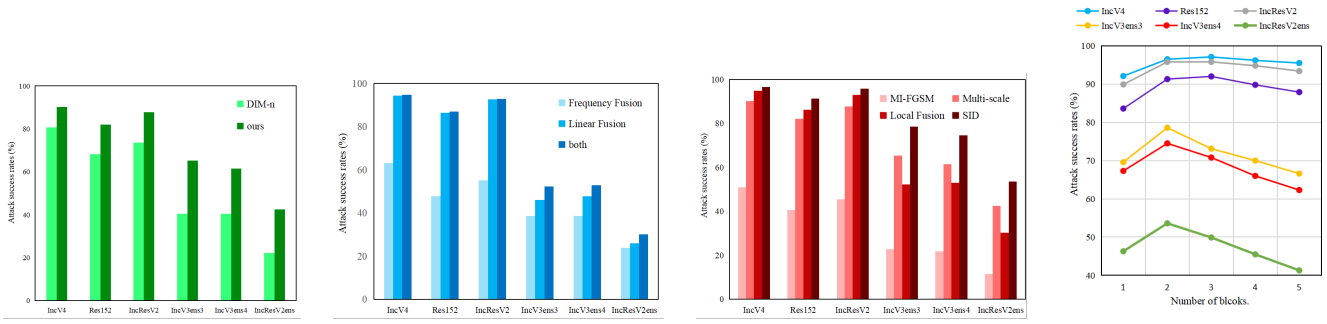
Our SID-SI-TIM achieves an average attack success rate of 84.3% against various defense methods, surpassing SSA-TI-DIM by 4.8%. Besides, our SID-TIM outperforms SSA-TI-DIM by 0.5%. Notably, when facing the most effective defense AT, SID-SI-TIM and SID-TIM exceed SSA-TI-DIM by 3.4% and 0.6%, respectively. Additionally, in the case of the diffusion-based defense method, DiffPure, SID-SI-TIM outperforms SSA-TI-DIM by 11.4%. These results further demonstrate the effectiveness of SID.

4.6. Ablation Study

Here we use IncV3 to craft the adversarial examples.

Multi-scale Vs. Fixed-scale. DIM transforms images by random padding. However, unlike DIM, we downsample the image content before padding, ensuring that the dimensions of the transformed images remain unchanged. Additionally, our method adjusts the size according to the number of transformations, generating multiple scaled transformed images in a single iteration. For fairness, we apply N times transformations to DIM, denoted as DIM-n, and SID does not employ the local image fusion. As shown in Figure 3a, our multi-scale method outperforms DIM-n across multiple models. This indicates that multi-scale effectively enhances the adversarial transferability.

The impact of Linear Fusion and Frequency Fusion. In local image fusion, we propose two fusion methods that integrate the global image from both spatial and frequency domains. As shown in Figure 3b, we explore the impact of two fusion methods. When using the two fusion methods individually, we set the probabilities to $p = 0.5$ and do



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

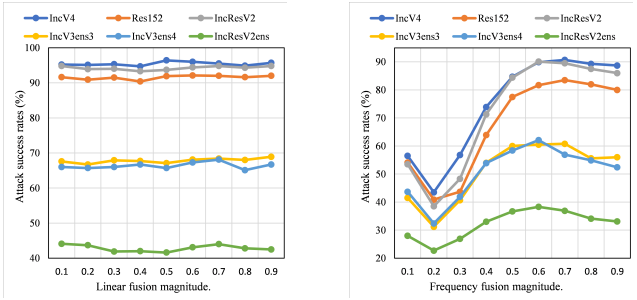
Figure 3. The black-box attack success rates (%) on six trained models with (a) DIM-n and our multi-scale method, (b) different local fusion methods, (c) MI-FGSM and different parts of our SID, (d) various numbers of blocks.

not employ the multi-scale approach. It can be observed that when using linear fusion, the transferability significantly exceeds that of frequency fusion. Notably, when both methods are employed simultaneously, the transferability exceeds that of either fusion method used individually.

The impact of Multi-scale and Local Image Fusion. From the perspective of spatial invariance, we design image transformation at both the global and local levels. As shown in Figure 3c, we further explore the effectiveness of each component and compare them with MI-FGSM. It is evident that in normally trained models, the local image fusion contributes more significantly to enhancing transferability, while in adversarially trained models, the multi-scale method has a greater impact. Consequently, our SID, which combines the advantages of both methods, achieves the best performance across different models.

The size of Local Image Fusion. As shown in Figure 3d, when $k = 2$, the transferability of adversarial examples reaches its maximum. Further increases in k result in a decline in performance. The number of blocks determines the size of the local fusion. When the number of blocks $k = 1$, it corresponds to the global image fusion. As the number of blocks increases, the fusion size decreases. When the fusion size becomes sufficiently small, the model is unable to effectively recognize the fused local information, leading to a decrease in transferability.

The magnitude of Local Image Fusion. To better demonstrate the effects brought about by the image fusion technology, we use two fusion methods separately and control the fusion magnitude. In linear fusion, the fusion magnitude represents the value of ω. As shown in Figure 4a, the change in the fusion magnitude does not bring about obvious changes. In frequency fusion, the fusion magnitude is the proportion of the low-pass part of the original image block after two-dimensional DCT transformation. As shown in Figure 4b, it can be significantly observed that the value reaches the maximum when the low-pass proportion is 0.6, which is also the parameter value we selected in the experiment. As the magnitude increases, the original image block is modified, resulting in the loss of global semantics and a decrease in the attack success rate.



(a) (b)

Figure 4. The black-box attack success rates (%) on six trained models with (a) various magnitude of the linear fusion, (b) various magnitude of the frequency fusion.

5. Conclusion

We discover that adjusting the position of image content can activate diverse behavioral patterns of DNNs regarding the input image, and neglecting these patterns limits the transferability of adversarial examples. To address this, we propose a novel input transformation-based attack called Spatial Invariance Diversity (SID). This approach activates diverse behavioral patterns of DNNs at different scales and positions by randomly adjusting the global image content position and employing local image fusion, resulting in various adversarial perturbations. Extensive experiments demonstrate that SID effectively improves the transferability of adversarial examples. From the perspective of spatial invariance, SID provides new insights into adversarial attacks by activating diverse behavioral patterns to improve the transferability of adversarial examples.

6. Acknowledgment

This work was supported by Eastern Talent Plan Leading Project under Grant BJKJ2024011.

References

[1] Wieland Brendel, Jonas Rauber, and Matthias Bethge. Decision-based adversarial attacks: Reliable attacks against black-box machine learning models. In *International Conference on Learning Representations*, 2018. 1

[2] Zhaoyu Chen, Bo Li, Shuang Wu, Kaixun Jiang, Shouhong Ding, and Wenqiang Zhang. Content-based unrestricted adversarial attack. *Advances in Neural Information Processing Systems*, 2024. 1, 2

[3] Shuyu Cheng, Yinpeng Dong, Tianyu Pang, Hang Su, and Jun Zhu. Improving black-box adversarial attacks with a transfer-based prior. *Advances in neural information processing systems*, 2019. 1

[4] Seun-An Choe, Ah-Hyung Shin, Keon-Hee Park, Jinwoo Choi, and Gyeong-Moon Park. Open-set domain adaptation for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 23943–23953, 2024. 1

[5] Jeremy Cohen, Elan Rosenfeld, and Zico Kolter. Certified adversarial robustness via randomized smoothing. In *international conference on machine learning*, pages 1310–1320, 2019. 3, 5

[6] 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. 2, 3

[7] 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. 2, 5

[8] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021. 1, 5

[9] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In *International Conference on Learning Representations*, 2015. 1, 2, 3

[10] Chuan Guo, Mayank Rana, Moustapha Cisse, and Laurens van der Maaten. Countering adversarial images using input transformations. In *International Conference on Learning Representations*, 2018. 2, 5

[11] 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. 1, 5

[12] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 2012. 1

[13] Alexey Kurakin, Ian J Goodfellow, and Samy Bengio. Adversarial examples in the physical world. In *International Conference on Learning Representations*, 2017. 1, 2

[14] Yann LeCun, Yoshua Bengio, et al. Convolutional networks for images, speech, and time series. *The handbook of brain theory and neural networks*, page 1995, 1995. 2, 3

[15] Yingwei Li, Qihang Yu, Mingxing Tan, Jieru Mei, Peng Tang, Wei Shen, Alan Yuille, et al. Shape-texture debiased neural network training. In *International Conference on Learning Representations*, 2021. 5

[16] Fangzhou Liao, Ming Liang, Yinpeng Dong, Tianyu Pang, Xiaolin Hu, and Jun Zhu. Defense against adversarial attacks using high-level representation guided denoiser. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1778–1787, 2018. 2, 5

[17] 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. 2, 5

[18] Yanpei Liu, Xinyun Chen, Chang Liu, and Dawn Song. Delving into transferable adversarial examples and black-box attacks. *arXiv preprint arXiv:1611.02770*, 2016. 2, 7

[19] 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, 2022. 2, 3, 5

[20] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations*, 2018. 1

[21] Muzammal Naseer, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, and Fatih Porikli. A self-supervised approach for adversarial robustness. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 262–271, 2020. 2, 5

[22] Weili Nie, Brandon Guo, Yujia Huang, Chaowei Xiao, Arash Vahdat, and Animashree Anandkumar. Diffusion models for adversarial purification. In *International Conference on Machine Learning*, pages 16805–16827, 2022. 2, 5

[23] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017. 2

[24] Daniel Smilkov, Nikhil Thorat, Been Kim, Fernanda Viégas, and Martin Wattenberg. Smoothgrad: removing noise by adding noise. *arXiv preprint arXiv:1706.03825*, 2017. 3

[25] Luchuan Song, Xiaodan Li, Zheng Fang, Zhenchao Jin, YueFeng Chen, and Chenliang Xu. Face forgery detection via symmetric transformer. In *Proceedings of the 30th ACM international conference on multimedia*, pages 4102–4111, 2022. 1

[26] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. In *International Conference on Learning Representations*, 2014. 1, 2

[27] 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

[28] Christian Szegedy, Sergey Ioffe, Vincent Vanhoucke, and Alexander Alemi. Inception-v4, inception-resnet and the impact of residual connections on learning. In *Proceedings of the AAAI conference on artificial intelligence*, 2017. 5

[29] Florian Tramèr, Alexey Kurakin, Nicolas Papernot, Ian Goodfellow, Dan Boneh, and Patrick McDaniel. Ensemble adversarial training: Attacks and defenses. In *International Conference on Learning Representations*, 2018. 3, 5

[30] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, 2017. 1

[31] Hao Wang, Yitong Wang, Zheng Zhou, Xing Ji, Dihong Gong, Jingchao Zhou, Zhifeng Li, and Wei Liu. Cosface: Large margin cosine loss for deep face recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5265–5274, 2018. 1

[32] Jiafeng Wang, Zhaoyu Chen, Kaixun Jiang, Dingkang Yang, Lingyi Hong, Pinxue Guo, Haijing Guo, and Wenqiang Zhang. Boosting the transferability of adversarial attacks with global momentum initialization. *Expert Systems with Applications*, 255:124757, 2024. 1, 2

[33] 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. 2, 3, 5

[34] Xiaosen Wang and Kun He. Enhancing the transferability of adversarial attacks through variance tuning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1924–1933, 2021. 2

[35] 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. 2

[36] 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, 3, 4, 5

[37] Zhaoyang Wei, Pengfei Chen, Xuehui Yu, Guorong Li, Jianbin Jiao, and Zhenjun Han. Semantic-aware sam for point-prompted instance segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3585–3594, 2024. 1

[38] Eric Wong, Leslie Rice, and J Zico Kolter. Fast is better than free: Revisiting adversarial training. In *International Conference on Learning Representations*, 2020. 3, 5

[39] Cihang Xie, Jianyu Wang, Zhishuai Zhang, Zhou Ren, and Alan Yuille. Mitigating adversarial effects through randomization. In *International Conference on Learning Representations*, 2018. 2, 5

[40] 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. 2, 3, 5

[41] Yifeng Xiong, Jiadong Lin, Min Zhang, John E Hopcroft, and Kun He. Stochastic variance reduced ensemble adversarial attack for boosting the adversarial transferability. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14983–14992, 2022. 2, 7

[42] Chaoning Zhang, Philipp Benz, Adil Karjauv, Jae Won Cho, Kang Zhang, and In So Kweon. Investigating top-k white-box and transferable black-box attack. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 15085–15094, 2022. 2

[43] Jianping Zhang, Jen-tse Huang, Wenxuan Wang, Yichen Li, Weibin Wu, Xiaosen Wang, Yuxin Su, and Michael R Lyu. Improving the transferability of adversarial samples by path-augmented method. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8173–8182, 2023. 2

[44] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2921–2929, 2016. 3

[45] 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. 2