Boosting Adversarial Transferability by Block Shuffle and Rotation

Kunyu Wang[1], Xuanran He[2], Wenxuan Wang[1], Xiaosen Wang[3*]
[1]Chinese University of Hong Kong, [2]Nanyang Technological University,
[3]Huawei Singularity Security Lab

Abstract

Adversarial examples mislead deep neural networks with imperceptible perturbations and have brought significant threats to deep learning. An important aspect is their transferability, which refers to their ability to deceive other models, thus enabling attacks in the black-box setting. Though various methods have been proposed to boost transferability, the performance still falls short compared with white-box attacks. In this work, we observe that existing input transformation based attacks, one of the mainstream transfer-based attacks, result in different attention heatmaps on various models, which might limit the transferability. We also find that breaking the intrinsic relation of the image can disrupt the attention heatmap of the original image. Based on this finding, we propose a novel input transformation based attack called block shuffle and rotation (BSR). Specifically, BSR splits the input image into several blocks, then randomly shuffles and rotates these blocks to construct a set of new images for gradient calculation. Empirical evaluations on the ImageNet dataset demonstrate that BSR could achieve significantly better transferability than the existing input transformation based methods under single-model and ensemble-model settings. Combining BSR with the current input transformation method can further improve the transferability, which significantly outperforms the state-of-the-art methods. Code is available at https://github.com/Trustworthy-AI-Group/BSR.

1. Introduction

Deep neural networks (DNNs) have established superior performance in many tasks, such as image classification [12, 13], segmentation [21], object detection [26, 27], face recognition [39], among others. Despite their achievements, DNNs have been observed to exhibit significant vulnerability to adversarial examples [8, 34, 41], which closely resemble legitimate examples, yet can deliberately misguide deep learning models to produce unreasonable predictions. The existence of such vulnerabilities gives rise to serious concerns, particularly in security-sensitive applications, such as autonomous driving [5], face verification [30].

Raw Image Shuffled Image Reshuffled Image

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

Adversarial attacks can generally be categorized into two types: white-box attacks and black-box attacks. White-box attacks involve having complete access to the target model's architecture and parameters. In contrast, black-box attacks only have limited information about the target model, which makes them more applicable for real-world applications. In the white-box setting, some studies employ the gradient with respect to the input sample to generate adversarial examples [8]. Those crafted adversarial examples exhibit transferability across neural models [25], which is the ability of adversarial examples generated on one model to deceive not only the victim model but also other models, making them suitable for black-box attacks. However, existing attack methods [14, 22] demonstrate outstanding white-box attack performance but relatively poorer transferability, limiting their efficacy in attacking real-world applications.

Recently, several approaches have emerged to enhance adversarial transferability, including incorporating momentum into gradient-based attacks [3, 18], attacking multiple models simultaneously [19], transforming the image before gradient calculation [46, 51], leveraging victim model features [47], and modifying the forward or backward process [45, 48]. Among these, input transformation based methods, which modify the input image for gradient calculation, have demonstrated significant effectiveness in im-

*Corresponding author. Email: xiaosen@hust.edu.cn

proving transferability. However, we find that all the existing input transformation based attacks result in different attention heatmaps [29] on different models. This discrepancy in attention heatmaps could potentially limit the extent of adversarial transferability.

The attention heatmaps highlight the crucial regions for classification. Motivated by this, we aim to maintain consistency in the attention heatmaps of adversarial examples across different models. Since we only have access to a single white-box model for the attack, we initially explore methods to disrupt the attention heatmaps. As shown in Fig. 1, we can disrupt the intrinsic relation within the image by randomly shuffling the divided blocks of the image, leading to different attention heatmaps compared with the raw image. Based on this finding, we propose a novel input transformation based attack, called block shuffle and rotation (BSR), which optimizes the adversarial perturbation on several transformed images to eliminate the variance among the attention heatmaps on various models. In particular, BSR randomly divides the image into several blocks, which are subsequently shuffled and rotated to create new images for gradient calculation. To eliminate the variance of random transformation and stabilize the optimization, BSR adopts the average gradient on several transformed images.

In summary, we highlight our contributions as follows:

- We show that breaking the intrinsic relation of the image can disrupt the attention heatmaps of the deep model.
- We propose a new attack called block shuffle and rotation (BSR), which is the first input transformation based attack to disrupt attention heatmaps for better transferability.
- Empirical evaluations on the ImageNet dataset demonstrate BSR achieves much better transferability than the state-of-the-art input transformation based attacks.
- BSR is compatible with other transfer-based attacks and can be integrated with each other to boost the adversarial transferability further.

2. Related Work

Here we briefly introduce adversarial attacks and defenses and summarize the visualization of attention heatmaps.

2.1. Adversarial Attacks

Szegedy et al. [34] first identified adversarial examples, which bring a great threat to DNN applications. Recently, numerous attacks have been proposed, which mainly fall into two categories: 1) *white-box attacks* can access all the information of the target model, such as gradient, weight, architecture, *etc*. Gradient-based attacks [1, 22] that maximize the loss function using the gradient w.r.t. input are the predominant white-box attacks. 2) *black-box attacks* are only allowed limited access to the target model, which can be further categorized into *Score-based attacks* [11, 38], *Decision-based attacks* [15, 44] and *Transfer-based attacks* [3, 7, 51]. Among these, transfer-based attacks cannot access the target model, in which the attacker adopts the adversarial examples generated by the surrogate model to attack the target model directly. Hence, transfer-based attacks can be effectively deployed in the real world and have attracted wide interest.

Fast Gradient Sign Method (FGSM) [8] adds the perturbation in the gradient direction of the input. Iterative FGSM (I-FGSM) [14] extends FGSM into an iterative version, showing better white-box attack performance but poor transferability. Recently, numerous works have been proposed to improve adversarial transferability.

MI-FGSM [3] introduces momentum into I-FGSM to stabilize the optimization procedure and escape the local optima. Later, more advanced momentum based attacks are proposed to further boost transferability, such as NI-FGSM [18], VMI-FGSM [40], EMI-FGSM [43], PGN [6] and so on. Ensemble attacks [16, 19, 52] generate more transferable adversarial examples by attacking multiple models simultaneously. Several works [49, 57] disrupt the feature space to generate adversarial examples.

On the other hand, input transformation has become one of the most effective ways to improve transferability. Diverse input method (DIM) [51] first resizes the image into random size and adds the padding to fixed size before the gradient calculation. Translation invariant method (TIM) [4] translates the image into a set of images for gradient calculation, which is approximated by convolving the gradient with a Gaussian kernel. Scale invariant method (SIM) [18] calculates the gradient of several scaled images. *Admix* [42] mixes a small portion of images from other categories to the input image to craft a set of admixed images. PAM [55] augments the input images from several augmentation paths.

In this work, we propose a new input transformation, called BSR, which breaks the intrinsic semantic relation of the input image for more diverse transformed images and results in much better transferability than the baselines.

2.2. Adversarial Defenses

With the increasing interest of adversarial attacks, researchers have been struggling to mitigate such threats. Adversarial training [8, 22, 37] injects adversarial examples into the training process to boost the model robustness. Among them, Tramèr et al. [37] propose ensemble adversarial training using adversarial perturbation generated on several models, showing great effectiveness against transfer-based attacks. Denoising filter is a data preprocessing method that filters out the adversarial perturbation before feeding them into the target model. For instance, Liao et al. [17] design a high-level representation guided denoiser (HGD) based on U-Net to eliminate adversarial

perturbation. Naseer et al. [23] train a neural representation purifier (NRP) using a self-supervised adversarial training mechanism to purify the input sample. Researchers also introduce several input transformation based defenses that transform the image before prediction to eliminate the effect of adversarial perturbation, such as random resizing and padding [51], feature squeezing using bit reduction [53], feature distillation [20]. Different from the above empirical defenses, several certified defense methods provide a provable defense in a given radius, such as interval bound propagation (IBP) [9], CROWN-IBP [54], randomized smoothing (RS) [2], *etc*.

2.3. Attention Heatmaps

As the inner mechanism of deep models remains elusive to researchers, several techniques [29, 31, 33] have been developed to interpret these models. Among them, attention heatmap is a widely adopted way to interpret deep models. For instance, Zhou et al. [56] adopts global average pooling to highlight the discriminative object parts. Selvaraju et al. [29] proposed Grad-CAM using the gradient information to generate more accurate attention heatmaps. In this work, we adopt the attention heatmap to investigate how to boost adversarial transferability.

3. Methodology

In this section, we first introduce the preliminaries and motivation. Then we provide a detailed description of our BSR, and summarize the difference between RLFAT [32] and BSR.

3.1. Preliminaries

Given a victim model f with parameters θ and a clean image $\mathbf{x}$ with ground-truth label $\mathbf{y}$, the attacker aims to generate an adversarial example $\mathbf{x}^{adv}$ that is indistinguishable from original image $\mathbf{x}$ (*i.e.*, $\|\mathbf{x}^{adv}-\mathbf{x}\|_p \leq \epsilon$) but can fool the victim model $f(\mathbf{x}^{adv};\theta) \neq f(\mathbf{x};\theta)=\mathbf{y}$. Here ϵ is the perturbation budget, and $\|\cdot\|_p$ is the ℓ_p norm distance. In this paper, we adopt ℓ_∞ distance to align with existing works. To generate such an adversarial example, the attacker often maximizes the objective function, which can be formalized as:

$$\mathbf{x}^{adv} = \underset{\|\mathbf{x}^{adv}-\mathbf{x}\|_p\leq\epsilon}{\operatorname{argmax}} J(\mathbf{x}^{adv},\mathbf{y};\theta), \tag{1}$$

where $J(\cdot)$ is the corresponding loss function (*e.g.*, cross-entropy loss). For instance, FGSM [8] updates the benign sample by adding a small perturbation in the direction of the gradient sign:

$$\mathbf{x}^{adv} = \mathbf{x} + \epsilon \cdot \operatorname{sign}(\nabla_x J(\mathbf{x},\mathbf{y};\theta)). \tag{2}$$

FGSM can efficiently craft adversarial examples but showing poor attack performance. Thus, I-FGSM [14] extends FGSM into an iterative version, which iteratively updates the the adversarial example by adding small perturbation with a step size α:

$$\mathbf{x}_t^{adv} = \mathbf{x}_{t-1}^{adv} + \alpha \cdot \operatorname{sign}(\nabla_{\mathbf{x}_{t-1}^{adv}} J(\mathbf{x}_{t-1}^{adv},\mathbf{y};\theta)), \tag{3}$$

where $\mathbf{x}_0^{adv}=\mathbf{x}$. Considering the poor transferability of I-FGSM, MI-FGSM [3] integrates momentum into the gradient for more transferable adversarial examples:

$$\begin{aligned} g_t &= \mu \cdot g_{t-1} + \frac{\nabla_{\mathbf{x}_{t-1}^{adv}} J(\mathbf{x}_{t-1}^{adv},\mathbf{y};\theta)}{\|\nabla_{\mathbf{x}_{t-1}^{adv}} J(\mathbf{x}_{t-1}^{adv},\mathbf{y};\theta)\|_1}, \\ \mathbf{x}_t^{adv} &= \mathbf{x}_{t-1}^{adv} + \alpha \cdot \operatorname{sign}(g_t),\ g_0=0, \end{aligned} \tag{4}$$

where μ is the decay factor. Suppose $\mathcal{T}$ is a transformation operator, existing input transformation based attacks are often integrated into MI-FGSM to boost adversarial transferability, *i.e.*, adopting $\nabla_{\mathbf{x}_{t-1}^{adv}} J(\mathcal{T}(\mathbf{x}_{t-1}^{adv}),\mathbf{y};\theta)$ for Eq. (4).

3.2. Motivation

While different models may have distinct parameters and architectures, there are often shared characteristics in their learned features for image recognition tasks [31, 56]. In this work, we hypothesize that the adversarial perturbations which target these salient features have a greater impact on adversarial transferability. Wu et al. [49] find disrupting the attention heatmaps can enhance transferability, which also supports our hypothesis. Intuitively, when the attention heatmaps of adversarial examples exhibit consistency across various models, it is expected to yield better adversarial transferability. To explore this idea, we initially assess the consistency of attention heatmaps generated by Grad-CAM [28] for several input transformation based attacks. Unfortunately, as shown in Fig. 2, the attention heatmaps of adversarial examples on the white-box model are different from that on the target black-box model, resulting in limited adversarial transferability. This finding motivates us to investigate a new problem:

How can we generate adversarial examples with consistent attention heatmaps across different models?

To maintain the consistency of attention heatmaps across multiple models, one direct approach is to optimize the adversarial perturbation by utilizing gradients w.r.t. the input image obtained from different models, *a.k.a.* ensemble attack [19]. By incorporating gradients from various models, each associated with its own attention heatmap for the same input image, ensemble attack helps eliminate the variance among attention heatmaps, resulting in improved transferability. In practice, however, it is often challenging and costly to access multiple models, making it more feasible to work with a single surrogate model. In this work, we explore how to obtain the gradients with different attention heatmaps on a single model using input transformation.

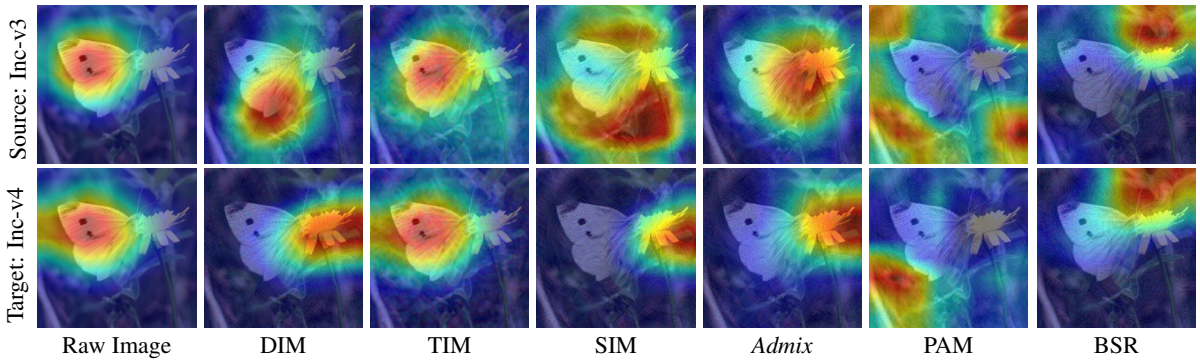

Figure 2. Attention heatmaps of adversarial examples generated by various input transformations using Grad-CAM.

With such transformation, we can optimize the perturbation to eliminate the variance among attention heatmaps of various transformed images, thus enhancing the consistency of attention heatmaps and adversarial transferability.

3.3. Block Shuffle and Rotation

With a single source model, we have to transform the input image to obtain diverse attention heatmaps when calculating the gradient. Thus, we should address the problem:

How to transform the image to disrupt the attention heatmap on a single source model?

Attention heatmap highlights the significant features that contribute to the deep model's accurate prediction. For human perception, we are capable of recognizing objects based on partial visual cues, even when they are partially obstructed by other objects. For instance, we can identify a cat by observing only a portion of its body (*e.g.*, head). This observation motivates us to employ image transformations that draw attention to specific regions of the main object, thereby varying the attention heatmap on a single model. Intuitively, randomly masking the object partially can force the deep model to focus on the remaining object, leading to various attention heatmaps. However, masking the object leads to a loss of information in the image, rendering the gradient meaningless for the masked block. Consequently, it can slow down the attack efficiency and effectiveness.

On the other hand, humans exhibit a remarkable ability to not only recognize the visible parts of the objects but also mentally reconstruct the occluded portions when the objects are partially obstructed by other elements. This cognitive process is attributed to our perception of intrinsic relationships inherent within the object, such as the understanding that a horse's legs are positioned beneath its body. Recognizing the significance of intrinsic relationships for human perception, we try to disrupt these relationships to affect attention heatmaps. In particular, we split the image into several blocks and shuffle the blocks to construct new images that appear visually distinct from the original one. As expected, the attention heatmaps are also disrupted on the transformed image, even when recovering the attention heatmaps to match the original image, as depicted in Fig. 1. Thus, we can break the intrinsic relationships for more diverse attention heatmaps to boost adversarial transferability.

Algorithm 1 Block Shuffle and Rotation

Input: A classifier f with parameters θ, loss function J; a raw example x with ground-truth label y; the magnitude of perturbation ϵ; number of iteration T; decay factor μ; the number of transformed images N; the number of blocks n; the maximum angle τ for rotation

Output: An adversarial example x^{adv}

$\alpha = \epsilon/T$, $g_0 = 0$

for $t = 1 \rightarrow T$ **do**

Generate several transformed images: $\mathcal{T}(\mathbf{x}^{adv}, n, \tau)$

Calculate the average gradient $\bar{g}_t$ by Eq. (6):

Update the momentum g_t by:

$$g_t = \mu \cdot g_{t-1} + \frac{\bar{g}_t}{\|\bar{g}_t\|_1}$$

Update the adversarial example:

$$x_t^{adv} = x_{t-1}^{adv} + \alpha \cdot \text{sign}(g_t) \tag{5}$$

end for

return x_T^{adv}

To achieve this goal, we propose a new input transformation $\mathcal{T}(\mathbf{x}, n, \tau)$, which randomly splits the image into $n \times n$ blocks followed by the random shuffling of these blocks. To further disrupt the intrinsic relationship, each block is independently rotated by an angle within the range of $-\tau \leq \beta \leq \tau$ degrees. During the rotation of each block, any portions that extend beyond the image boundaries are removed while the resulting gaps are filled with zero.

With the transformation $\mathcal{T}(\mathbf{x}, n, \tau)$, we can obtain more diverse attention heatmaps compared to those obtained from

various models. Hence, strengthening the consistency of heatmaps (*harder*) helps result in consistent attention across models (*easier*). This motivates us to adopt transformation $\mathcal{T}(\mathbf{x}, n, \tau)$ for gradient calculation to achieve more consistent heatmaps. Note that transformation $\mathcal{T}(\mathbf{x}, n, \tau)$ cannot guarantee that all transformed images are correctly classified ($\sim 86.8\%$ on Inc-v3). To eliminate the variance introduced by the intrinsic relation corruption, we calculate the average gradient on N transformed images as follows:

$$\bar{g} = \frac{1}{N}\sum_{i=0}^{N} \nabla_{\mathbf{x}^{adv}} J(\mathcal{T}(\mathbf{x}^{adv}, n, \tau), \mathbf{y}; \theta)). \tag{6}$$

The input transformation $\mathcal{T}(\mathbf{x}, n, \tau)$ is general to any existing attacks. Here we integrate it into MI-FGSM, denoted as Block Shuffle and Rotation (BSR), and summarize the algorithm in Algorithm 1.

3.4. BSR Vs. RLFAT

Song et al. [32] propose Robust Local Features for Adversarial Training (RLFAT) by minimizing the distance between the high-level feature of the original image and block shuffled image for better generalization. Since RLFAT also shuffles the image blocks, we highlight the difference between BSR and RLFAT as follows:

- **Goal.** BSR aims to generate more transferable adversarial examples, while RLFAT boosts the generalization of adversarial training.
- **Strategy.** BSR shuffles and rotates the image blocks, while RLFAT only shuffles the blocks.
- **Usage.** BSR directly adopts the transformed images for gradient calculation to craft adversarial examples, while RLFAT treats it as a regularizer for the original image.

With the different goals, strategies, and usages, BSR should definitely be a new and novel input transformation based attack to effectively boost the adversarial transferability.

4. Experiments

In this section, we conduct empirical evaluations on ImageNet dataset to evaluate the effectiveness of BSR.

4.1. Experimental Setup

Dataset. We evaluate our proposed BSR on 1000 images belonging to 1000 categories from the validation set of ImageNet dataset [24].

Models. We adopt four popular models, *i.e.*, Inception-v3 (Inc-v3) [35], Inception-v4 (Inc-v4), Inception-Resnet-v3 (IncRes-v3) [36], Resnet-v2-101 (Res-101) [12], and three ensemble adversarially trained models, *i.e.*, Inc-v3$_{ens3}$, Inc-v3$_{ens4}$, IncRes-v2$_{ens2}$ [37] as victim models to evaluate the transferability. To further verify the effectiveness of BSR, we utilize several advanced defense methods, including HGD [17], R&P [50], NIPS-r3[1], Bit-RD [53], JPEG [10], FD [20], RS [2] and NRP [23].

Baselines. To verify the effectiveness of BSR, we choose five competitive input transformation based attacks as our baselines, *i.e.* DIM [51], TIM [4], SIM [18], *Admix* [42] and PAM[55]. For fairness, all the input transformations are integrated into MI-FGSM [3].

Parameters Settings. We set the maximum perturbation $\epsilon = 16$, number of iteration $T = 10$, step size $\alpha = \epsilon/T$ and the decay factor $\mu = 1$ for MI-FGSM [3]. DIM [51] adopts the transformation probability of 0.5. TIM [4] utilizes a kernel size of 7×7. The number of copies of SIM [18] and *Admix* [42] is 5. *Admix* admixes 3 images with the strength of 0.2. The number of scale for PAM [55] is 4, and the number of augmented path is 3 . Our BSR splits the image into 2×2 blocks with the maximum rotation angle $\tau = 24^\circ$ and calculates the gradients on $N = 20$ transformed images.

4.2. Evaluation on Single Model

We first evaluate the attack performance on various input transformation based attacks, *i.e.*, DIM TIM, SIM, *Admix*, PAM, and our proposed BSR. We craft the adversaries on the four standard trained models and test them on seven models. The attack success rates, *i.e.*, the misclassification rates of the victim model on the adversarial examples, are summarized in Tab. 1. Each column denotes the model to be attacked and each row indicates that the attacker generates the adversarial examples on the corresponding models.

It can be observed that for DIM and TIM, DIM exhibits superior performance on standard trained models while TIM exhibits better transferability on adversarially trained models. SIM, as a special case of *Admix*, can achieve better performance than DIM and TIM, while *Admix* shows the best performance on standardly trained model among the four baselines and PAM exhibits better transferability on adversarially trained model. In contrast, our proposed BSR, surpasses existing input transformation based attacks while maintaining comparable performance in white-box attacks. Remarkably, on standard trained models, BSR attains an average attack success rate of 93.8%, exhibiting a substantial improvement over *Admix* by a clear margin of at least 6.5%. Similarly, on adversarially trained models, BSR achieves an average attack success rate of 56.2%, outperforming PAM by a significant margin of 11.0%. These exceptional findings substantiate the superiority of BSR in generating transferable adversarial examples, thereby highlighting the importance of maintaining attention heatmap consistency across different models as a means to enhance transferability.

[1] https://github.com/anlthms/nips-2017/tree/master/mmd

Model	Attack	Inc-v3	Inc-v4	IncRes-v2	Res-101	Inc-v3$_{ens3}$	Inc-v3$_{ens4}$	IncRes-v2$_{ens}$
Inc-v3	DIM	98.6*	64.4	60.2	53.5	18.8	18.4	9.5
	TIM	**100.0***	49.3	43.9	40.2	24.6	21.7	13.4
	SIM	**100.0***	69.5	68.5	63.5	32.3	31.0	17.4
	Admix	**100.0***	82.2	81.1	73.8	38.7	37.9	19.8
	PAM	**100.0***	76.4	75.5	69.6	39.0	38.8	20.0
	BSR	**100.0***	**96.2**	**94.7**	**90.5**	**55.0**	**51.6**	**29.3**
Inc-v4	DIM	72.0	97.6*	63.8	57.2	22.6	21.1	11.7
	TIM	59.1	99.7*	49.0	41.9	26.8	22.9	16.6
	SIM	80.4	99.7*	73.4	69.4	48.6	45.2	29.6
	Admix	89.0	**99.9***	85.3	79.0	55.5	51.7	32.3
	PAM	86.7	**99.9***	81.6	75.9	55.4	50.5	33.2
	BSR	**96.1**	**99.9***	**93.4**	**88.4**	**57.6**	**52.1**	**34.3**
IncRes-v2	DIM	70.3	64.7	93.1*	58.0	30.4	23.5	16.9
	TIM	62.2	55.6	97.4*	50.3	32.4	27.5	22.6
	SIM	85.9	80.0	98.7*	76.1	56.2	49.1	42.5
	Admix	90.8	86.3	**99.2***	82.2	63.6	56.6	49.4
	PAM	88.6	86.3	99.4*	81.6	66.0	58.3	**51.0**
	BSR	**94.6**	**93.8**	98.5*	**90.7**	**71.4**	**63.1**	**51.0**
Res-101	DIM	76.0	68.4	70.3	98.0*	34.7	31.8	19.6
	TIM	59.9	52.2	51.9	99.2*	34.4	31.2	23.7
	SIM	74.1	69.6	69.1	99.7*	42.8	39.6	25.7
	Admix	84.5	80.2	80.7	**99.9***	51.6	44.7	29.9
	PAM	77.4	73.9	75.7	**99.9***	51.2	46.3	32.2
	BSR	**97.1**	**96.6**	**96.6**	99.7*	**78.7**	**74.7**	**55.6**

Table 1. Attack success rates (%) on seven models under single model setting with various single input transformations. The adversaries are crafted on Inc-v3, Inc-v4, IncRes-v2 and Res-101 respectively. * indicates white-box attacks.

Attack	Inc-v4	IncRes-v2	Res-101	Inc-v3$_{ens3}$	Inc-v3$_{ens4}$	IncRes-v2$_{ens}$
BSR-DIM	98.3$_{\uparrow 31.9}$	94.8$_{\uparrow 34.6}$	90.1$_{\uparrow 36.6}$	57.8$_{\uparrow 39.0}$	54.5$_{\uparrow 36.1}$	32.3$_{\uparrow 22.8}$
BSR-TIM	94.7$_{\uparrow 45.4}$	92.5$_{\uparrow 48.6}$	87.0$_{\uparrow 47.0}$	71.3$_{\uparrow 46.7}$	68.4$_{\uparrow 46.7}$	47.9$_{\uparrow 34.5}$
BSR-SIM	99.4$_{\uparrow 29.9}$	98.4$_{\uparrow 29.9}$	97.8$_{\uparrow 34.3}$	84.3$_{\uparrow 52.0}$	81.4$_{\uparrow 50.4}$	59.2$_{\uparrow 41.8}$
BSR-*Admix*	98.9$_{\uparrow 16.7}$	98.8$_{\uparrow 17.7}$	98.2$_{\uparrow 24.4}$	89.1$_{\uparrow 50.4}$	86.9$_{\uparrow 49.5}$	68.0$_{\uparrow 48.2}$
BSR-PAM	98.5$_{\uparrow 22.1}$	97.3$_{\uparrow 21.8}$	96.9$_{\uparrow 27.3}$	79.4$_{\uparrow 40.4}$	75.3$_{\uparrow 36.5}$	50.5$_{\uparrow 30.5}$
Admix-TI-DIM	90.4	87.3	83.7	72.4	68.4	53.4
PAM-TI-DIM	89.3	85.5	80.7	73.6	69.1	52.1
BSR-TI-DIM	95.2	92.9	87.9	74.2	70.7	50.0
BSR-SI-TI-DIM	**98.5**	**97.1**	**95.4**	**90.6**	**90.0**	**75.1**

Table 2. Attack success rates (%) on seven models under single model setting with various input transformations combined with BSR. The adversaries are crafted on Inc-v3. ↑ indicates the increase of attack success rate when combined with BSR.

4.3. Evaluation on Combined Input Transformation

Previous works [42] have shown that a good input transformation based attack should not only exhibit better transferability, but also be compatible with other input transformations to generate more transferable adversarial examples. Following the evaluation setting of *Admix*, we combine our BSR with various input transformations, denoted BSR-DIM, BSR-TIM, BSR-SIM, BSR-*Admix* and BSR-PAM. Here we also combine BSR with multiple input transformations, denoted as BSR-TI-DIM and BSR-SI-TI-DIM to compare *Admix*-TI-DIM and PAM-TI-DIM.

Attack	Inc-v3	Inc-v4	IncRes-v2	Res-101	Inc-v3$_{ens3}$	Inc-v3$_{ens4}$	IncRes-v2$_{ens}$
DIM	99.0*	97.1*	93.4*	99.7*	57.6	51.5	35.9
TIM	99.8*	97.4*	94.7*	99.8*	61.6	55.5	45.6
SIM	99.9*	99.1*	98.5*	**100.0***	78.4	75.2	60.6
Admix	**100.0***	99.6*	99.0*	**100.0***	85.1	80.9	67.8
PAM	99.9*	99.7*	99.4*	**100.0***	86.1	81.6	69.1
BSR	**100.0***	**99.9***	99.9*	99.9*	**92.4**	**89.0**	**77.2**
Admix-TI-DIM	99.6*	98.8*	98.2*	99.8*	93.1	92.4	89.4
PAM-TI-DIM	99.8*	99.8*	99.2*	**99.8***	95.8	95.2	93.0
BSR–TI-DIM	99.8*	99.8*	99.7*	**99.8***	96.1	95.1	90.8
BSR-SI-TI-DIM	**99.9***	**99.9***	**99.9***	**99.8***	**99.1**	**99.1**	**97.0**

Table 3. Attack success rates (%) on seven models under ensemble model setting with various input transformations. The adversaries are crafted on Inc-v3, Inc-v4, IncRes-v2 and Res-101 model. * indicates white-box attacks.

Method	HGD	R&P	NIPS-r3	Bit-RD	JPEG	FD	RS	NRP	Average
Admix-TI-DIM	92.8	93.5	94.5	82.4	97.6	90.9	72.6	80.4	88.1
PAM-TI-DIM	95.4	95.3	96.4	85.9	98.4	93.4	74.0	83.8	91.6
BSR-TI-DIM	97.1	98.0	97.9	84.9	98.8	93.2	69.1	73.6	89.1
BSR-SI-TI-DIM	**98.5**	**99.1**	**99.4**	**91.4**	**99.2**	**97.1**	**83.9**	**84.2**	**94.1**

Table 4. Attack success rates (%) of eight defense methods by *Admix*, SSA and BSR input transformations. The adversaries are crafted on Inc-v3, Inc-v4, IncRes-v2 and Res-101 synchronously.

We report the attack success rates of the adversarial examples generated on Inc-v3 in Tab. 2 and the results for other models in Appendix. Our BSR significantly improves the transferability of these input transformation based attacks. In general, BSR can improve the attack success rate with a range from 16.7% to 52.0%. In particular, when combining these input transformation with BSR, the attack performance on adversarially trained models are significantly improved by a margin of $22.8\% \sim 52.0\%$. Although *Admix*-TI-DIM demonstrates the best performance among combined methods, our proposed BSR-TI-DIM surpasses *Admix*-TI-DIM with a clear margin of 2.6% on average attack success rates. Notably, when BSR is combined with SI-DI-TIM, it further enhances transferability by a margin ranging from 8.1% to 31.6%. This further supports the high effectiveness of BSR and shows its excellent compatibility with other input transformation based attacks.

4.4. Evaluation on Ensemble Model

Liu *et al*. [19] first propose ensemble attack to boost the transferability by synchronously attacking several models. To evaluate the compatibility of the proposed BSR with ensemble attack, we generate the adversarial examples on four standard trained models and test them on adversarially trained models following the setting in MI-FGSM [3]. We evaluate the attack performance of single input transformation as well as BSR combined with the existing input transformations, respectively.

As shown in Tab. 3, under ensemble model setting, PAM showcases superior white-box attack capabilities, surpassing other baselines significantly with a margin of at least 0.7% for adversarially trained models. In contrast, BSR consistently outperforms PAM on these models by a margin of 6.3% to 8.1%, and maintains comparable white-box attack performance with *Admix*. Although PAM-TI-DIM can achieve at least 93.0% attack success rate on adversarially trained models. Notably, BSR-SI-TI-DIM achieves average attack success rate of 98.4% on adversarially trained models, surpassing PAM-TI-DIM with a margin of at least 3.1%. Such remarkable attack performance validates the remarkable effectiveness of BSR for improving the transferability and poses a huge threat to security-critical applications once again.

4.5. Evaluation on Defense Method

To thoroughly evaluate the effectiveness of our proposed method, we assess the attack performance of BSR against several defense mechanisms, including HGD, R&P, NIPS-r3, Bit-RD, JPEG, FD, RS and NRP. From previous experiments, combined input transformations with ensemble attack exhibit the best attack performance. Here we adopt the adversarial examples generated by PAM-TI-DIM, BSR-

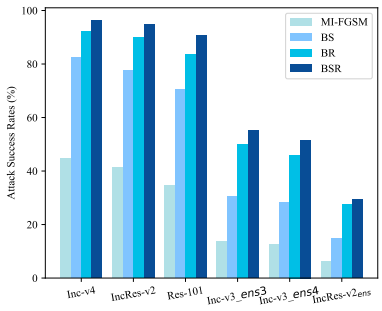

Figure 3. Attack success rates (%) of various models on the adversarial examples generated by MI-FGSM, BS, BR and BSR, respectively.

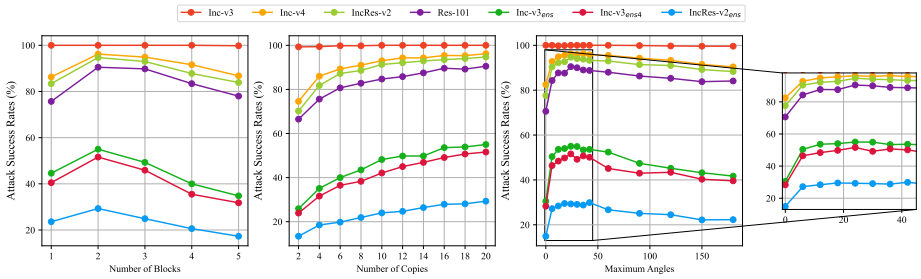

Figure 4. Attack successful rates (%) of various models on the adversarial examples generated by BSR with various numbers of blocks, number of transformed images and range of the rotation angles. The adversarial examples are crafted on Inc-v3 model and tested on the other six models under the black-box setting.

TI-DIM and BSR-SI-TI-DIM under the ensemble setting to attack these defense approaches.

As shown in Tab. 4, BSR-SI-TI-DIM achieves the average attack success rate of 94.1%, which outperforms PAM-TI-DIM by an average margin of 2.5%. Notably, even on the certified defense RS and powerful denoiser NRP, BSR-SI-TI-DIM achieve the attack success rate of 83.9% and 84.2%, which outperforms PAM-TI-DIM with a clear margin of 9.9% and 0.4%, respectively. Such excellent attack performance further shows the superiority of BSR and reveals the inefficiency of existing defenses.

4.6. Ablation Study

To further gain insight into the performance improvement of BSR, we conduct ablation and hyper-parameter studies by generating the adversarial examples on Inc-v3 and evaluating them on the other six models.

On the effectiveness of shuffle and rotation. After splitting the image into several blocks, we shuffle the image blocks and rotate each block. To explore the effectiveness of shuffle and rotation, we conduct two additional attacks, *i.e.*, block shuffle (BS) and block rotation (BR). As shown in Fig. 3, BS and BR can achieve better transferability than MI-FGSM, supporting our proposition that optimizing the adversarial perturbation on the input image with different attention heatmaps can eliminate the variance among attention heatmaps to boost the transferability. By combining BS and BR, BSR exhibits the best transferability, showing its rationality and high effectiveness in crafting transferable adversarial examples.

On the number of blocks n. As shown in Fig. 4, when $n = 1$, BSR only rotates the raw image, which cannot bring disruption on the attention heatmaps, showing the poorest transferability. When $n > 3$, increasing n results in too much variance that cannot be effectively eliminated, decreasing the transferability. Hence, a suitable magnitude of disruption on the raw image is significant to improve the transferability and we set $n = 2$ in our experiments.

On the number of transformed images N. Since BSR introduces variance when breaking the intrinsic relation, we adopt the average gradient on N transformed images to eliminate such variance. As shown in Fig. 4, when $N = 5$, BSR can already achieve better transferability than MI-FGSM, showing its high efficiency and effectiveness. When we increase N, the attack performance would be further improved and be stable when $N > 20$. Hence, we set $N = 20$ in our experiments.

On the range of rotation angles τ. We randomly rotate the image blocks with the angle $-\tau \leq \beta \leq \tau$, which also affects the magnitude of disruption on the image. As shown in Fig. 4, we conduct the experiments from $\tau = 6°$ to $\tau = 180°$. When τ is smaller than $24°$, increasing τ results in more disruption on the image so as to achieve better transferability. If we continue to increase τ, the rotation will introduce too much disruption, which decays the performance. Hence, we set $\tau = 24°$ in our experiments.

5. Conclusion

Intuitively, the consistent attention heatmaps of adversaries on different models will have better transferability. However, we find that the existing input transformation based attacks often result in inconsistent attention heatmaps on various models, limiting the transferability. To this end, we propose a novel input transformation based attack called block shuffle and rotation (BSR), which optimizes the perturbation on several transformed images with different attention heatmaps to eliminate the variance among the attention heatmaps on various models. Empirical evaluations on ImageNet dataset show that BSR achieves better transferability than the SOTA attacks under various attack settings. We hope our approach can provide new insight to improve the transferability by generating adversarial examples with more stable attention heatmaps on different models.

References

[1] Nicholas Carlini and David Wagner. Towards Evaluating the Robustness of Neural Networks. In *IEEE Symposium on Security and Privacy*, pages 39–57, 2017. 2

[2] Jeremy M Cohen, Elan Rosenfeld, and J Zico Kolter. Certified adversarial robustness via randomized smoothing. In *International Conference on Machine Learning*, 2019. 3, 5

[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. 1, 2, 3, 5, 7

[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. 2, 5

[5] Kevin Eykholt, Ivan Evtimov, Earlence Fernandes, Bo Li, Amir Rahmati, Chaowei Xiao, Atul Prakash, Tadayoshi Kohno, and Dawn Song. Robust Physical-World Attacks on Deep Learning Visual Classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1625–1634, 2018. 1

[6] Zhijin Ge, Hongying Liu, Xiaosen Wang, Fanhua Shang, and Yuanyuan Liu. Boosting Adversarial Transferability by Achieving Flat Local Maxima. In *Proceedings of the Advances in Neural Information Processing Systems*, 2023. 2

[7] Zhijin Ge, Fanhua Shang, Hongying Liu, Yuanyuan Liu, Liang Wan, Wei Feng, and Xiaosen Wang. Improving the Transferability of Adversarial Examples with Arbitrary Style Transfer. In *Proceedings of the ACM International Conference on Multimedia*, page 4440–4449, 2023. 2

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

[9] Sven Gowal, Krishnamurthy Dj Dvijotham, Robert Stanforth, Rudy Bunel, Chongli Qin, Jonathan Uesato, Relja Arandjelovic, Timothy Mann, and Pushmeet Kohli. Scalable Verified Training for Provably Robust Image Classification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4842–4851, 2019. 3

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

[11] Chuan Guo, Jacob Gardner, Yurong You, Andrew Gordon Wilson, and Kilian Weinberger. Simple Black-box Adversarial Attacks. In *International Conference on Machine Learning*, pages 2484–2493, 2019. 2

[12] 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

[13] 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. 1

[14] Alexey Kurakin, Ian Goodfellow, and Samy Bengio. Adversarial Examples in the Physical World. In *International Conference on Learning Representations (Workshop)*, 2017. 1, 2, 3

[15] Huichen Li, Xiaojun Xu, Xiaolu Zhang, Shuang Yang, and Bo Li. QEBA: Query-Efficient Boundary-Based Blackbox Attack. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1218–1227, 2020. 2

[16] Yingwei Li, Song Bai, Yuyin Zhou, Cihang Xie, Zhishuai Zhang, and Alan L. Yuille. Learning Transferable Adversarial Examples via Ghost Networks. In *AAAI Conference on Artificial Intelligence*, 2020. 2

[17] 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/CVF Conference on Computer Vision and Pattern Recognition*, pages 1778–1787, 2018. 2, 5

[18] 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. 1, 2, 5

[19] Yanpei Liu, Xinyun Chen, Chang Liu, and Dawn Song. Delving into transferable adversarial examples and black-box attacks. In *International Conference on Learning Representations*, 2017. 1, 2, 3, 7

[20] Zihao Liu, Qi Liu, Tao Liu, Nuo Xu, Xue Liu, Yanzhi Wang, and Wujie Wen. Feature distillation: Dnn-oriented jpeg compresstion against adversarial examples. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 860–868, 2019. 3, 5

[21] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully Convolutional Networks for Semantic Segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3431–3440, 2015. 1

[22] 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, 2

[23] 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. 3, 5

[24] Russakovsky Olga, Deng Jia, Su Hao, Krause Jonathan, Satheesh Sanjeev, Ma Sean, hUANG Zhiheng, Karpathy Andrej, Khosla Adiya, and Bernstein Michael et al. Imagenet large scale visual recognition challenge. In *International Journal of Computer Vision*, pages 211–252, 2015. 5

[25] Nicolas Papernot, Patrick McDaniel, Ian Goodfellow, Somesh Jha, Z Berkay Celik, and Ananthram Swami. Practical black-box attacks against machine learning. In *Proceedings of the 2017 ACM on Asia conference on computer and communications security*, pages 506–519, 2017. 1

[26] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You Only Look Once: Unified, Real-time Object

Detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 779–788, 2016. 1

[27] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in Neural Information Processing Systems*, 2015. 1

[28] 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 Conference on Computer Vision and Pattern Recognition*, pages 2921–2929, 2016. 3, 14

[29] 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 *International Conference on Computer Vision*, pages 618–626, 2017. 2, 3

[30] Mahmood Sharif, Sruti Bhagavatula, Lujo Bauer, and Michael K. Reiter. Accessorize to a Crime: Real and Stealthy Attacks on State-of-the-Art Face Recognition. In *Proceedings of the ACM SIGSAC Conference on Computer and Communications Security*, pages 1528–1540, 2016. 1

[31] Daniel Smilkov, Nikhil Thorat, Been Kim, Fernanda Viégas, and Martin Wattenberg. Smoothgrad: Removing Noise by Adding Noise. *arXiv preprint arXiv:1706.03825*, 2017. 3

[32] Chuanbiao Song, Kun He, Lin Jiadong, Liwei Wang, and E. John Hopcroft. Robust Local Features for Improving the generalization of adversarial training. In *International Conference on Learning Representaion*, 2020. 3, 5

[33] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *International conference on machine learning*, pages 3319–3328, 2017. 3

[34] 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

[35] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Wojna Zbjgniew. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2818–2826, 2016. 5, 14

[36] Christian Szegedy, Sergey Ioffe, Vincent Vanhoucke, and Alex Alemi. Inception-v4, inception-resnet and the impact of residual connections on learning. In *AAAI Conference on Artificial Intelligence*, 2017. 5, 14

[37] Florian Tramèr, Alexey Kurakin, Nicolas Papernot, Ian Goodfellow, Dan Boneh, and Patrick McDaniel. Ensemble Adversarial Training: Attacks and Defenses. *International Conference on Learning Representations*, 2018. 2, 5

[38] Jonathan Uesato, Brendan O'donoghue, Pushmeet Kohli, and Aaron Oord. Adversarial Risk and the Dangers of Evaluating against Weak Attacks. In *International Conference on Machine Learning*, pages 5025–5034, 2018. 2

[39] 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

[40] Xiaosen Wang and Kun He. Enhancing the transferability of adversarial attacks through variance tuning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2021. 2

[41] Xiaosen Wang, Kun He, Chuanbiao Song, Liwei Wang, and John E. Hopcroft. AT-GAN: A Generative Attack Model for Adversarial Transferring on Generative Adversarial Nets. *arXiv preprint arXiv:1904.07793*, 2019. 1

[42] Xiaosen Wang, Xuanran He, Jingdong Wang, and Kun He. Admix: Enhancing the transferability of adversarial attacks. In *International Conference on Computer Vision*, pages 16138–16147, 2021. 2, 5, 6

[43] Xiaosen Wang, Jiadong Lin, Han Hu, Jingdong Wang, and Kun He. Boosting Adversarial Transferability through Enhanced Momentum. In *British Machine Vision Conference*, page 272, 2021. 2

[44] Xiaosen Wang, Zeliang Zhang, Kangheng Tong, Dihong Gong, Kun He, Zhifeng Li, and Wei Liu. Triangle Attack: A Query-efficient Decision-based Adversarial Attack. In *European conference on computer vision*, 2022. 2

[45] Xiaosen Wang, Kangheng Tong, and Kun He. Rethinking the Backward Propagation for Adversarial Transferability. In *Proceedings of the Advances in Neural Information Processing Systems*, 2023. 1

[46] 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. 1

[47] Zhibo Wang, Hengchang Guo, Zhifei Zhang, Wenxin Liu, Qin Zhan, and Kui Ren. Feature importance-aware transferable adversarial attacks. In *International Conference on Computer Vision*, pages 7639–7648, 2021. 1

[48] Zhiyuan Wang, Zeliang Zhang, Siyuan Liang, and Xiaosen Wang. Diversifying the High-level Features for better Adversarial Transferability. In *Proceedings of the British Machine Vision Conference*, 2023. 1

[49] Weibin Wu, Yuxin Su, Xixian Chen, Shenglin Zhao, Irwin King, Michael R Lyu, and Yu-Wing Tai. Boosting the Transferability of Adversarial Samples via Attention. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1161–1170, 2020. 2, 3

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

[51] 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 Conference on Computer Vision and Pattern Recognition*, pages 2730–2739, 2019. 1, 2, 3, 5

[52] 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

[53] Weilin Xu, David Evans, and Yanjun Qi. Feature squeezing: Detecting adversarial examples in deep neural networks. *Network and Distributed System Security Symposium*, 2018. 3, 5

[54] Huan Zhang, Hongge Chen, Chaowei Xiao, Sven Gowal, Robert Stanforth, Bo Li, Duane Boning, and Cho-Jui Hsieh. Towards Stable and Efficient Training of Verifiably Robust Neural Networks. In *International Conference on Learning Representations*, 2019. 3

[55] 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, 5, 12

[56] 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

[57] Wen Zhou, Xin Hou, Yongjun Chen, Mengyun Tang, Xiangqi Huang, Xiang Gan, and Yong Yang. Transferable adversarial perturbations. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 452–467, 2018. 2