Improving the Transferability of Adversarial Examples with Resized-Diverse-Inputs, Diversity-Ensemble and Region Fitting

Junhua Zou[1][0000−0003−4655−7173], Zhisong Pan[1][0000−0001−8615−7313], Junyang Qiu[2][0000−−0001−8288−6180], Xin Liu[1][0000−0003−3051−4793], Ting Rui[1][0000−−0002−2949−5874], and Wei Li[1][0000−0003−1733−7956]

[1] Army Engineering University of PLA, China
[2] Jiangnan Institute of Computing Technology, China
278287847@qq.com

Abstract. We introduce a three stage pipeline: resized-diverse-inputs (RDIM), diversity-ensemble (DEM) and region fitting, that work together to generate transferable adversarial examples. We first explore the internal relationship between existing attacks, and propose RDIM that is capable of exploiting this relationship. Then we propose DEM, the multi-scale version of RDIM, to generate multi-scale gradients. After the first two steps we transform value fitting into region fitting across iterations. RDIM and region fitting do not require extra running time and these three steps can be well integrated into other attacks. Our best attack fools six black-box defenses with a 93% success rate on average, which is higher than the state-of-the-art gradient-based attacks. Besides, we rethink existing attacks rather than simply stacking new methods on the old ones to get better performance. It is expected that our findings will serve as the beginning of exploring the internal relationship between attack methods.

Keywords: Adversarial examples, the internal relationship, region fitting, resized-diverse-inputs, diversity-ensemble

1 Introduction

Recent work has demonstrated that deep neural networks (DNNs) are challenged by their vulnerability to adversarial examples [11,28], i.e., inputs with carefully-crafted perturbations that are almost indistinguishable from the original images can be misclassified by DNNs. Moreover, a more severe and complicated security issue is the transferability of adversarial examples [18,20], i.e., adversarial examples generated by a given DNN can also mislead other unknown DNNs. Fig. 1 shows the transferability of an adversarial example. The threat of adversarial examples can even extend to the physical world [2,10,14], and has motivated extensive research on security-sensitive applications. These defenses include adversarial training [11,19,29], input denoising [16], input transformation [12,31], theoretically-certified defenses [23,30] and others [22,25]. Although adversarial

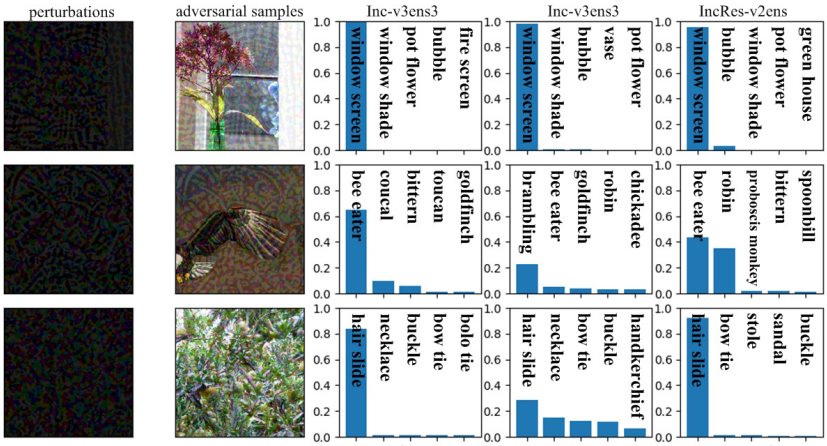


Fig. 1. Perturbations and adversarial samples along their predicted confidence scores of 3 black-box models. Inc-v3ens3, Inc-v3ens4 and IncRes-v2ens [29] are defense models

examples are security threats to the practical deployment, they can help DNNs to identify the vulnerability before they are applied in reality [19].

We focus on the gradient-based attacks of the classification task in this paper. With the knowledge of a given DNN, the gradient-based attacks are the most commonly used methods, and can attack black-box models based on the transferability of adversarial examples. Usually, existing attack methods are combined together to achieve higher attack success rates.

Motivation. Among the gradient-based attacks, the diverse-inputs method (DIM) [32] applies random and differentiable transformations to the inputs with probability p, then feeds these transformed inputs into a white-box model for gradient calculation. Usually, DIM is combined with the translation-invariant method (TIM) [9] to achieve state-of-the-art results. Based on these two methods, our simple observations are shown as follow:

1. TIM can be considered as a Gaussian blur process for gradients. As shown in Fig. 2, TIM can blur a normal image (the first row), but cannot blur an image with vertical and horizontal stripes (the second row).
2. As shown in Fig. 3, the gradients of a diverse input have many vertical and horizontal stripes (here we visualize the gradients as images by setting non-zero values to 255 to highlight zero values). The number of stripes depends on the diversity scale.

Intuitively, DIM can alleviate the loss of gradient information caused by Gaussian blur, and thus generate more transferable adversarial examples. However, DIM sets up the transformation probability p and limits the maximum diversity size to a really small size to avoid success rates dropping. The hyper-parameters of DIM restrict the number of stripes of the gradients, and cannot benefit TIM as much as possible. The intuition reveals the other two clues. One

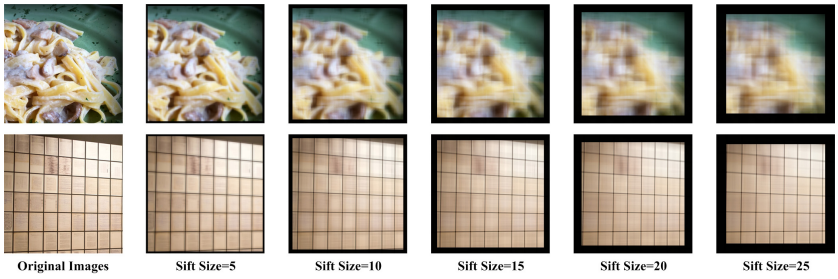


Fig. 2. Two rows of images generated by the translation-invariant method [9] with different sift size ranging from 5 to 25. Images of the first row gradually become blurred as the sift size increases while images of the second row remain stable

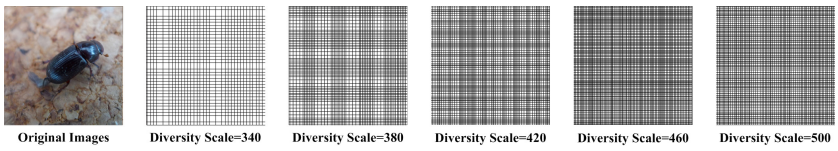


Fig. 3. A set of visualized gradients of a diverse example with different diversity scale [32]. The number of black stripes of these images increases as the diversity scale increases. In addition, gradients are visualized by setting non-zero values to 255 to highlight zero values. Hence, the black stripes indicate zero values of gradients while the white regions indicate non-zero values

is that multi-scale gradient information benefits the transferability of adversarial examples. The other is that DIM divides the gradient information into many regions, and Gaussian Filter with large kernel may blur image edges. The two characteristics of these two methods indicate that region fitting plays a more important role than value fitting in adversarial example generation.

Methods. In this paper, we introduce a three stage pipeline: resized-diverse-inputs (RDIM), diversity-ensemble (DEM) and region fitting, that work together to generate transferable adversarial examples. We first explore the internal relationship between DIM [32] and TIM [9] based on the observations above, and propose a **resized-diverse-inputs method** (RDIM) that is more suitable to characterize this relationship. Compared with DIM, RDIM removes the transformation probability p, sets a much larger diversity size and finally resizes the diverse inputs to the original size at each iteration. We combine TIM and RDIM, and then conduct extensive experiments on the ImageNet dataset. The results show that this combination can achieve higher attack success rates on defense models comparing with the state-of-the-art results. We then propose a **diversity-ensemble method** (DEM), the multi-scale version of RDIM, to further boost the success rates. We show that DEM can further promote TIM because DEM generates multi-scale gradients with different numbers of vertical and horizontal stripes for TIM. After the first two steps we transform value

fitting into **region fitting** across iterations. RDIM and region fitting do not require extra running time and these three steps can be well integrated into other attacks, such as model-ensemble methods [8]. Our best attack fools six black-box defenses with a 93% attack success rate on average, which is higher than the state-of-the-art multi-model gradient-based attacks.

Rather than simply stacking the new methods on the old ones to get better performance, we rethink the proposed methods. It is expected that our findings will serve as the beginning of exploring the internal relationship between attack methods. In summary, our contributions are as follows:

1. We are the first to explore the internal relationship between attack methods. We find that the gradients of diverse inputs have many vertical and horizontal stripes, and these gradients can be used to alleviate the loss of gradient information caused by TIM.
2. Based on the internal relationship between DIM and TIM, we propose RDIM to exploit this relationship. We show that RDIM further boosts the attack success rates against black-box defenses.
3. We propose DEM which can generate multi-scale gradients for TIM. DEM can further promote TIM because DEM generates multi-scale gradients with different numbers of vertical and horizontal stripes for TIM. We also transform value fitting into region fitting across iterations to further boost the success rates against black-box defenses.
4. Our best attack fools six black-box defenses with a 93% attack success rate on average, which is higher than the state-of-the-art gradient-based attacks.

2 Related Work

Recent work has demonstrated that DNNs are challenged by their vulnerability to adversarial examples [3,28]. The primary purposes of adversarial example generating methods are high attack success rates with minimal size of perturbations [6]. Attack methods in the classification task can be categorized into three types—the gradient-based attacks [8,11,15], the score-based attacks [21] and the decision-based attacks [5,7]. In addition, adversarial examples exist in face recognition [4], object detection [10], semantic segmentation [1], etc.. In this paper, we focus on gradient-based attacks of the classification task.

Gradient-based attacks can be categorized into three parts—the gradient processing part, the ensemble part and the input preprocessing part. In the gradient processing part, Goodfellow et al. [11] proposed the fast gradient sign method (FGSM) to craft adversarial examples by performing one-step update efficiently. Kurakin et al. [14] extended FGSM to the basic iterative method (BIM) and showed the powerful ability of BIM in white-box attacks but lousy performance in black-box attacks. Dong et al. [8] proposed the momentum iterative fast gradient sign method (MI-FGSM) to boost success rates in black-box attacks by integrating a momentum term into BIM. Lin et al. [17] proposed the Nesterov iterative fast gradient sign method (NI-FGSM) to further improve the transferability of adversarial examples by adapting Nesterov accelerated gradient into

MI-FGSM. In the ensemble part, Dong et al. [8] proposed a model ensemble method to fool robust black-box models obtained by ensemble adversarial training. Lin et al. [17] used a set of scaled images to achieve model augmentation and named it scale-invariant attack method (SIM). In the input preprocessing part, Dong et al. [9] proposed the translation-invariant attack method (TIM) to generate adversarial examples that are less sensitive to the discriminative regions. Xie et al. [32] proposed the input diversity (DIM) to generate adversarial examples by iteratively applying the random transformation to input examples.

Most of defenses can be categorized into two types—adversarial training and input modification. Adversarial training [11] mainly augmented the training dataset by its adversarial examples during the training process to broaden the discriminative regions [15]. Additionally, Tramèr et al. [29] further improved the robustness of defense models and proposed the ensemble adversarial training by augmenting clean examples with adversarial examples crafted for various models. Input modification aimed to reduce the influence of adversarial examples on models by mitigating adversarial perturbations through different modification methods. Xie et al. [31] employed random resizing and padding to defense against the adversarial attacks. Liao et al. [16] reduced the effects of adversarial perturbations using high-level representation guided denoiser.

3 Methodology

Given an input example X which we call a clean example, and it can be correctly classified to the ground-truth label y_{true} by deep model $f(\cdot)$ to $f(X) = y_{true}$. It is possible to construct two types of adversarial examples to attack model $f(\cdot)$ by adding different adversarial perturbations to the clean example X. In non-targeted attack, an adversarial example X^{adv} is generated with the ground-truth label y_{true} to mistaken the model as $f(X^{adv}) \neq y_{true}$. In targeted attack, a targeted adversarial example X^{adv} is classified to the specified target class y_{target} as $f(X^{adv}) = y_{target}$, where $y_{target} \neq y_{true}$. In the standard case, in order to generate indistinguishable adversarial example X^{adv}, the distortion between adversarial example X^{adv} and clean example X is measured as L_p norm of the adversarial noise as $\left\|X^{adv} - X\right\|_p \leq \varepsilon$, where p could be 0, 1, 2, ∞, and ε is the size of the adversarial perturbation.

3.1 Gradient-Based Attack Methods

In this subsection, we present a brief introduction of the family of the gradient-based attack methods.

Fast Gradient Sign Method (FGSM) [11] generates an adversarial example X^{adv} by maximizing the loss function $J\left(X^{adv}, y_{true}\right)$ of a pre-trained DNN. FGSM can efficiently craft an adversarial example as

$$X^{adv} = X + \varepsilon \cdot sign\left(\nabla_X J\left(X, y_{true}\right)\right), \tag{1}$$

where $\nabla_X J(\cdot,\cdot)$ computes the gradient of the loss function w.r.t. X, $sign(\cdot)$ is the sign function, and ε is the required scalar value that basically restricts the L_∞ norm of the perturbation.

Iterative Fast Gradient Sign Method (I-FGSM) [14] applies FGSM multiple times with a small steps size α, while FGSM generates an adversarial example by taking a single large step in the direction. The basic iterative method (BIM) [14] starts with $X_0^{adv} = X$, and iteratively computes as

$$X_{t+1}^{adv} = X_t^{adv} + \alpha \cdot sign\left(\nabla_{X_t^{adv}} J\left(X_t^{adv}, y_{true}\right)\right), \tag{2}$$

where X_t^{adv} denotes the adversarial example generated at the t-th iteration, and $X_0^{adv} = X$.

Momentum Iterative Fast Gradient Sign Method (MI-FGSM) [8] enhances the transferability of adversarial examples in black-box attacks and maintains the success rates in white-box attacks. The updating procedures are

$$g_{t+1} = \mu \cdot g_t + \frac{\nabla_{X_t^{adv}} J\left(X_t^{adv}, y_{true}\right)}{\left\|\nabla_{X_t^{adv}} J\left(X_t^{adv}, y_{true}\right)\right\|_1}, \tag{3}$$

$$X_{t+1}^{adv} = X_t^{adv} + \alpha \cdot sign\left(g_{t+1}\right), \tag{4}$$

where g_t denotes the accumulated gradient at the t-th iteration, and μ is the decay factor of g_t.

Nesterov Iterative Fast Gradient Sign Method (NI-FGSM) [17] integrates Nesterov accelerated gradient into gradient-based attack methods to avoid the "missing" of the global maximum as

$$X_t^{nes} = X_t^{adv} + \alpha \cdot \mu \cdot g_t, \tag{5}$$

$$g_{t+1} = \mu \cdot g_t + \frac{\nabla_{X_t^{adv}} J\left(X_t^{nes}, y^{true}\right)}{\left\|\nabla_{X_t^{adv}} J\left(X_t^{nes}, y^{true}\right)\right\|_1}, \tag{6}$$

$$X_{t+1}^{adv} = X_t^{adv} + \alpha \cdot \mathrm{sign}\left(g_{t+1}\right). \tag{7}$$

Diverse-Inputs Method (DIM) [32] generates adversarial examples by applying the random transformation to input examples at each iteration where the transformation function $T(X_t^{adv}, p)$ is

$$T(X_t^{adv}, p) = \begin{cases} T(X_t^{adv}) & \text{with probability } p \\ X_t^{adv} & \text{with probability 1-}p \end{cases} \tag{8}$$

Translation-Invariant Method (TIM) [9] uses a set of translated images to form an adversarial example as

$$\begin{aligned} &X_{t+1}^{adv} = \textstyle\sum_{i,j} T_{ij}(X_t^{adv}), \\ &s.t.\ \|X_t^{adv} - X^{real}\|_\infty \le \epsilon, \end{aligned} \tag{9}$$

where $T_{ij}(X_t^{adv})$ denotes the translation function that respectively shifts input X_t^{adv} by i and j pixels along the two-dimensions.

Algorithm 1 RDIM

Input: An example X; the original size S; the diversity scale S_1.
Output: A diverse example X^d.

1: $a \sim \text{Unif}(S, S_1)$; // get the random size a
2: $X^r = resize(X, (a, a))$; // resize the input image to the random size a
3: $H = S_1 - a$; // get the padding size H
4: $top, left \sim \text{Unif}(0, H)$; // get the random top and left padding size
5: $bottom = H - top, right = H - left$; // get the bottom and right padding size
6: $X^p = padding(X^r, (top, bottom, left, right))$; // get the padding image X^p
7: **Return** $X^d = resize(X^p, (S, S))$. // resize the padding image to the original size

3.2 Observation Analyses

Our simple observations are shown in Fig. 2 and Fig. 3. TIM fails to blur an image with vertical and horizontal stripes, while the gradients of a diverse input have many vertical and horizontal stripes. Intuitively, DIM can alleviate the loss of gradient information caused by Gaussian blur, and thus generate more transferable adversarial examples. We present the analyses as follow:

1. Compared with the normal size ($299 \times 299 \times 3$), the input of DIM is a larger example ($S_1 \times S_1 \times 3$, where $S_1 > 299$), which leads to the deviation of the model output. DIM does not resize the diverse inputs to the original size after the process. Hence, diversity scale of DIM is limited to 330 to avoid the vast size difference between the original inputs and the diverse inputs. Additionally, the probability p of DIM also limits the diversity.
2. TIM can be considered as a Gaussian blur process and cause the loss of gradient information. Lin et al. [17] show that TIM with a smaller kernel is better in multi-model attack. In this paper, we find another way to alleviate the loss of gradient information. Gaussian blur cannot blur an image with vertical and horizontal stripes while RDIM fills this gap.
3. Multi-group of gradients with different diversity scales can satisfy the need of TIM for blurring images with different types of stripes.
4. DIM divides the gradient information into many regions, and Gaussian Filter with large kernel may blur image edges. The two characteristics of these two methods indicate that region fitting plays a more important role than value fitting in adversarial example generation.

3.3 Resized-Diverse-Inputs Method

Based on the analyses above, we propose a **resized-diverse-inputs method** (RDIM) that is more suitable for the internal relationship with TIM [9]. Compared with DIM, RDIM removes the transformation probability p, sets a much larger diversity size and finally resizes the diverse inputs to the original size at each iteration. These three improvements of RDIM correspond to the first two analyses in Sec. 3.2. The algorithm of the RDIM is presented in Algorithm 1.

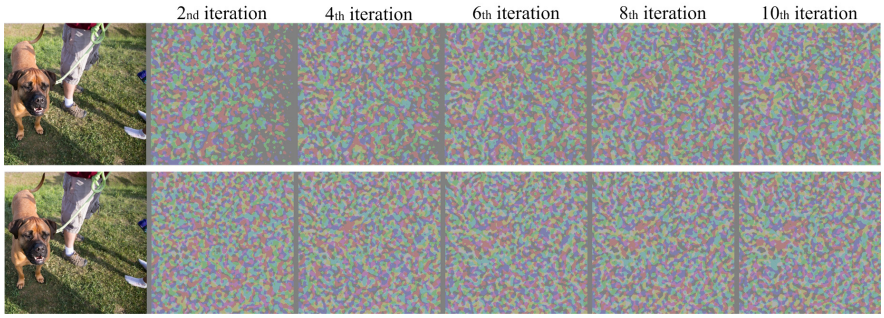


Fig. 4. Visualization of perturbations respectively generated with value fitting (the first row) and region fitting (the second row) in ten iterations. The value fitting cannot craft perturbations with detailed texture in the first four iterations

3.4 Diversity-Ensemble Method

For multi-scale setting, we also propose a **diversity-ensemble method** (DEM), the multi-scale version of RDIM, to improve the transferability of adversarial examples. Inspired by the third analysis in Sec. 3.2, we propose DEM, which generates multi-scale gradients with different numbers of vertical and horizontal stripes for TIM. DEM can satisfy the need of TIM for blurring images with different types of stripes. Similar to the ensemble-in-logits method [8], we fuse the logits of K diversity scales as

$$l\left(X\right) = \sum\nolimits_{k=1}^{K} \omega_k l\left(T\left(X, S_k\right)\right), \tag{10}$$

where $l\left(T\left(X, S_k\right)\right)$ denotes the logits of resized diverse inputs with k_{th} scale, ω_k denotes the ensemble weight with $\omega_k \geq 0$ and $\sum_{k=1}^{K} \omega_k = 1$.

3.5 Region Fitting

TIM can be considered as a Gaussian blur process with a large kernel (15×15) for gradients while DIM divides the gradients into many regions. These two methods for gradients mainly process the pixel region while normal iterative methods fit pixel value iteratively. Hence, we transform value fitting into region fitting across iterations. Compared with the updating procedure Eq. (7), region fitting can be expressed as

$$X_{t+1}^{adv} = Clip_{\varepsilon}\left\{X_t^{adv} + \varepsilon \cdot \operatorname{sign}\left(g_{t+1}\right)\right\}. \tag{11}$$

The difference between Eq. (7) and Eq. (11) across iterations is that we change α into ε. Eq. (7) iteratively increases the perturbation size with step size α, and finally makes the perturbation size reach ε. Eq. (11) makes the perturbation size reach ε at each iteration, and generates adversarial examples to meet the L_∞ norm bound by clipping function. Dong et al. [9] show that the

Algorithm 2 RF-DE-TIM

Input: A clean example X and ground-truth label y_{true}; the logits of K diversity scales $l\left(T\left(X,S_1\right)\right), l\left(T\left(X,S_2\right)\right), \ldots, l\left(T\left(X,S_K\right)\right)$; ensemble weights $\omega_1, \omega_2, \ldots, \omega_k$;
Input: The perturbation size ε; iterations T and decay factor μ.
Output: An adversarial example X^{adv}.

$\alpha = \varepsilon / T$;
$g_0 = 0$; $X_0^{adv} = X$;
for $t = 0$ to $T - 1$ **do**
    Input X_t^{adv};
    Get logits $l\left(X_t^{adv}\right)$ by Eq. (10); // fuse the logits of K diversity scales
    Get the gradient $\nabla_X J\left(X_t^{adv}, y_{true}\right)$;
    Process the gradient by $W * \nabla_X J\left(X_t^{adv}, y_{true}\right)$; // Gaussian blur for gradient
    Update g_{t+1} by Eq. (6); // accumulate the gradient
    Update X_{t+1}^{adv} by Eq. (11); // apply the region fitting
Return $X^{adv} = X_t^{adv}$.

classifiers rely on different discriminative regions for predictions. Region Fitting can accelerate the process of searching the discriminative regions as shown in Fig. 4.

We summarize RF-DE-TIM (the combination of TIM, RDIM, DEM, region fitting and MI-FGSM) in Algorithm 2.

4 Experiments

To validate the effectiveness of our methods, we present extensive experiments on ImageNet dataset. Table 1 introduces the abbreviations used in the paper.

We first provide experimental settings in Sec. 4.1. Then we report the internal relationship between RDIM and TIM with the opposite results of different combinations of attack methods in Sec. 4.2. Finally, we compare the results of our methods with the baseline methods in Sec. 4.3 and Sec. 4.4.

4.1 Experimental Settings

Dataset. We utilize an ImageNet-compatible dataset[3] [24] used in the NIPS 2017 adversarial competition to comprehensively compare the results of our methods with the baseline methods. The image size is $299 \times 299 \times 3$.

Models. We consider six defense models—Inc-v3ens3, Inc-v3ens4, IncRes-v2ens [29], high-level representation guided denoiser (HGD) [16], input transformation through random resizing and padding (R&P) [31], and rank-3 submission[4] in the NIPS 2017 adversarial competition, as the robust black-box defense models. To attack these models mentioned above, we also consider four

[3] https://github.com/tensorflow/cleverhans/tree/master/examples/nips17_adversarial_competition/dataset
[4] https://github.com/anlthms/nips-2017/tree/master/mmd

Table 1. Abbreviations used in the paper

Abbreviation	Explanation
RDI-FGSM	The combination of RDIM and FGSM
RDI-MI-FGSM	The combination of RDIM and MI-FGSM
TI-RDIM	The combination of RDIM, TIM and MI-FGSM
TI-DIM	The combination of DIM, TIM and MI-FGSM
NI-TI-RDIM	The combination of RDIM, TIM and NI-FGSM
NI-TI-DIM	The combination of DIM, TIM and NI-FGSM
DE-TIM	The combination of RDIM, DEM, TIM and MI-FGSM
SI-TIM	The combination of SIM, DIM, TIM and MI-FGSM
DE-NI-TIM	The combination of RDIM, DEM, TIM and NI-FGSM
SI-NI-TIM	The combination of SIM, DIM, TIM and NI-FGSM
RF-TI-RDIM	The combination of region fitting, RDIM, TIM and MI-FGSM
RF-DE-TIM	The combination of region fitting, RDIM, DEM, TIM and MI-FGSM

normally trained models—Inception v3 (Inc-v3) [27], Inception v4 (Inc-v4), Inception ResNet v2 (IncRes-v2) [26] and ResNet v2-152 (Res-v2-152) [13], as the white-box models to craft adversarial examples. It should be noted that adversarial examples crafted for four normally trained models are unaware of any defense strategies and will be used to attack six defense models, including top-3 defense solutions of NIPS 2017 adversarial competition.

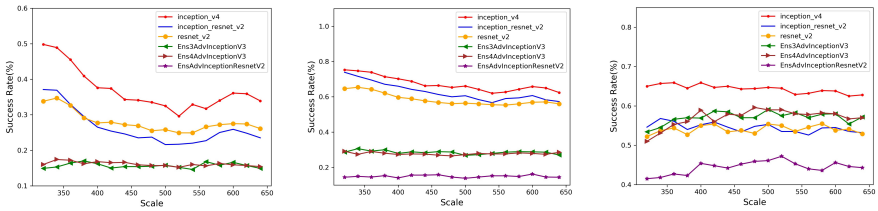
Baselines. In our experiments, we first explore the internal relationship between attack methods by RDI-FGSM, RDI-MI-FGSM, and TI-RDIM . Then in single-scale attack manner, we respectively compare TI-RDIM and NI-TI-RDIM with two baseline methods, TI-DIM and NI-TI-DIM. In the multi-scale attack manner, we respectively compare DE-TIM and DE-NI-TIM with two baseline methods, SI-TIM and SI-NI-TIM. We also include RF-DE-TIM, SI-NI-TIM and SI-NI-TI-DIM [17] in ensemble-based attacks for comparison.

Hyper-parameters. We follow the settings in TIM [9] with the number of iteration as $T = 10$, the maximum perturbation as $\varepsilon = 16$, the decay factor as $\mu = 1.0$. For TIM, We set the kernel size to 15×15. For SI-NI-FGSM, we follow the settings in NIM [17] with the number of the scale copies as $m = 5$. For DEM, we set the diversity list to $[340, 380, 420, 460, 500]$. Please note that the hyper-parameters settings for all attacks are the same.

4.2 The Internal Relationship

In this subsection, we attack the Inc-v3 model by RDI-FGSM, RDI-MI-FGSM, and TI-RDIM with different diversity scales and show the success rates against six black-box models in Fig. 5. It can be seen that the success rates of RDI-FGSM and RDI-MI-FGSM decrease as diversity scale increasing, while success rates of TI-RDIM continue increasing at first and slightly dropping after the diversity scale exceeds 520.

Based on Fig. 5, we further explore the internal relationship between RDIM and TIM. We find that images with vertical and horizontal stripes are more



(a) RDI-FGSM (b) RDI-MI-FGSM (c) TI-RDIM

Fig. 5. The success rates (%) of black-box attacks against six black-box models—Inc-v4, IncRes-v2, Res-v2-152,Inc-v3ens3, Inc-v3ens4 and IncRes-v2ens. The adversarial examples are crafted for Inc-v3 respectively using RDI-FGSM, RDI-MI-FGSM and TI-RDIM with the diversity scale ranging from 320 to 500

likely to fail when attacking DNNs even if they are perturbed by the translation-invariant method. We present two sets of perturbed images in Fig. 2. Additionally, we show gradients of diverse inputs (here we visualize the gradients as images) which have many vertical and horizontal stripes in Fig. 3. These three figures indicate that DIM can reduce the effect of stripes on TIM, and thus make adversarial examples generated by the combination of these two methods more transferable. However, without noticing that a certain number of stripes benefit TIM, DIM sets up the transformation probability p and limits the maximum diversity scale to 330 to avoid success rates dropping. Hence, we propose a **resized-diverse-inputs method** (RDIM) by removing the transformation probability p, setting a much larger diversity size and resizing the diverse inputs to their original size at each iteration. These three groups of interesting results of Fig. 2, Fig. 3 and Fig. 5 show that RDIM can reduce the effect of stripes on TIM, and thus make adversarial examples generated by the combination of RDIM and TIM more transferable. The experimental results validate the first three analyses of Sec. 3.2.

4.3 Single-Model Attacks

In this subsection, we categorize the experiments of single-model attacks into two types—single-scale attacks and multi-scale attacks based on time efficiency, e.g., all methods of single-scale attacks have similar runtime in generating adversarial examples. We compare the black-box success rates of the resized-diverse-inputs based methods with single-scale attacks and multi-scale attacks, respectively. In single-scale attacks, we generate adversarial examples for four normally trained models respectively using TI-DIM, TI-RDIM, NI-TI-DIM and NI-TI-RDIM. We then use six defense models to defense the crafted adversarial examples. We present the success rates in Table 2 for the comparison of TI-DIM and TI-RDIM, and Table 3 for the comparison of NI-TI-DIM and NI-TI-RDIM.

It can be observed from the tables that our method RDIM can further boost the success rates against these six defense models by a large margin when in-

Table 2. The success rates (%) of black-box attacks against six defense models under single-model setting. The adversarial examples are generated for Inc-v3, Inc-v4, IncRes-v2, Res-v2-152 respectively using TI-DIM and TI-RDIM

	Attack	Inc-v3ens3	Inc-v3ens4	IncRes-v2ens	HGD	R&P	NIPS-r3
Inc-v3	TI-DIM	46.6	47.6	38.1	38.1	37.4	42.8
	TI-RDIM	**59.1**	**59.0**	**46.1**	**48.3**	**47.5**	**52.1**
Inc-v4	TI-DIM	48.2	48.3	39.3	41.2	40.7	42.5
	TI-RDIM	**61.7**	**62.0**	**50.8**	**53.2**	**51.5**	**55.7**
IncRes-v2	TI-DIM	61.3	60.8	59.3	59.7	60.9	62.1
	TI-RDIM	**69.5**	**69.0**	**67.1**	**66.8**	**67.7**	**69.7**
Res-v2-152	TI-DIM	56.2	54.9	50.1	52.6	51.1	53.1
	TI-RDIM	**61.5**	**64.1**	**53.8**	**53.4**	**52.7**	**59.0**

Table 3. The success rates (%) of black-box attacks against six defense models under single-model setting. The adversarial examples are generated for Inc-v3, Inc-v4, IncRes-v2, Res-v2-152 respectively using NI-TI-DIM and NI-TI-RDIM

	Attack	Inc-v3ens3	Inc-v3ens4	IncRes-v2ens	HGD	R&P	NIPS-r3
Inc-v3	NI-TI-DIM	50.0	48.7	36.7	37.5	36.5	42.6
	NI-TI-RDIM	**53.4**	**52.6**	**39.8**	**42.3**	**41.0**	**46.3**
Inc-v4	NI-TI-DIM	52.5	52.7	40.1	43.2	40.7	42.5
	NI-TI-RDIM	**57.9**	**56.5**	**45.3**	**48.9**	**47.6**	**50.7**
IncRes-v2	NI-TI-DIM	61.1	60.2	60.3	60.7	61.2	62.7
	NI-TI-RDIM	**66.1**	**65.5**	**62.8**	**65.8**	**64.3**	**66.0**
Res-v2-152	NI-TI-DIM	56.1	55.9	51.2	50.1	49.1	53.7
	NI-TI-RDIM	**60.1**	**60.4**	**59.9**	**52.4**	**51.4**	**57.6**

tegrated into the state-of-the-art attacks. In general, the resized-diverse-inputs based methods outperform the baseline methods by 2% $\sim$ 14%. It demonstrates that our method RDIM is better than DIM, and can serve as a powerful method to improve the transferability of adversarial examples.

In multi-scale attacks, we also generate adversarial examples for four normally trained models respectively using SI-TIM, DE-TIM, SI-NI-TIM and DE-NI-TIM. We then evaluate the crafted adversarial examples by attacking six defense models. We present the success rates in Table 4 for the comparison of SI-TIM and DE-TIM. Table 5 presents the comparison of SI-NI-TIM and DE-NI-TIM.

We can observe from the tables that our method DEM can further improve the success rates against these six defense models by a large margin when integrated into the state-of-the-art attacks. In general, methods combined with DEM outperform the baseline methods by 11% $\sim$ 24%. In particular, when using DE-TIM, the combination of our method and TIM, to attack IncRes-v2 model, the adversarial examples achieve no less than 78% success rates against all six defense models. In Table 2, Table 3, Table 4 and Table 5, it should be noted that the adversarial examples crafted for a non-defense model can fool six defense models with no less than 78% success rates. The results not only validate

Table 4. The success rates (%) of black-box attacks against six defense models under single-model setting. The adversarial examples are generated for Inc-v3, Inc-v4, IncRes-v2, Res-v2-152 respectively using SI-TIM and DE-TIM

	Attack	Inc-v3ens3	Inc-v3ens4	IncRes-v2ens	HGD	R&P	NIPS-r3
Inc-v3	SI-TIM	48.4	51.2	37.5	36.3	34.6	40.0
	DE-TIM	**70.1**	**70.3**	**58.0**	**61.2**	**59.3**	**64.2**
Inc-v4	SI-TIM	51.2	50.9	42.9	41.9	39.5	42.5
	DE-TIM	**71.1**	**69.2**	**59.6**	**64.2**	**63.4**	**65.1**
IncRes-v2	SI-TIM	68.8	66.1	65.4	60.6	59.4	62.7
	DE-TIM	**79.8**	**79.5**	**78.2**	**80.0**	**79.3**	**80.1**
Res-v2-152	SI-TIM	54.7	55.3	48.0	45.2	43.4	48.4
	DE-TIM	**77.5**	**75.8**	**69.4**	**73.9**	**71.8**	**75.0**

Table 5. The success rates (%) of black-box attacks against six defense models under single-model setting. The adversarial examples are generated for Inc-v3, Inc-v4, IncRes-v2, Res-v2-152 respectively using SI-NI-TIM and DE-NI-TIM

	Attack	Inc-v3ens3	Inc-v3ens4	IncRes-v2ens	HGD	R&P	NIPS-r3
Inc-v3	SI-NI-TIM	52.1	52.8	40.7	39.5	37.3	44.4
	DE-NI-TIM	**66.4**	**66.8**	**52.7**	**56.2**	**55.4**	**59.2**
Inc-v4	SI-NI-TIM	55.6	54.1	44.7	43.1	41.4	46.3
	DE-NI-TIM	**67.3**	**65.2**	**56.4**	**60.9**	**59.2**	**62.7**
IncRes-v2	SI-NI-TIM	68.6	66.5	64.1	57.9	58.4	61.9
	DE-NI-TIM	**77.5**	**75.5**	**74.2**	**75.1**	**76.2**	**77.9**
Res-v2-152	SI-NI-TIM	57.6	55.8	48.7	47.9	46.2	53.3
	DE-NI-TIM	**74.5**	**74.8**	**67.5**	**69.3**	**68.6**	**73.0**

the effectiveness of RDIM and DEM, but also indicate that the current defenses fail to meet the demand of practical security.

4.4 Ensemble-based Attacks

In the subsection, we further show the performance of adversarial examples crafted for an ensemble of models. Similar to Sec. 4.3, we categorize the experiments of Ensemble-based attacks into single-scale attacks and multi-scale attacks. We generate adversarial examples for the ensemble of Inc-v3, Inc-v4, IncRes-v2 and Res-v2-152 with equal ensemble weights.

In single-scale attacks, we generate adversarial examples respectively using TI-DIM, TI-RDIM, NI-TI-DIM and RF-TI-RDIM, and evaluate the effectiveness of crafted adversarial examples by attacking six defenses. Table 6 shows the results of black-box attacks against six defenses. The results indicate that the proposed method RDIM can also boost the success rates over the baselines attacks in ensemble-based attacks.

In multi-scale attacks, we further present adversarial examples respectively using SI-NI-TIM, SI-NI-TI-DIM, DE-NI-TIM, DE-TIM and RF-DE-TIM, and then employ six defense models to defense the generated adversarial examples.

Table 6. The success rates (%) of black-box attacks against six defense models under multi-model setting. The adversarial examples are generated for the ensemble of Inc-v3, Inc-v4, IncRes-v2, Res-v2-152 using TI-DIM, TI-RDIM, NI-TI-DIM and RF-TI-RDIM

Attack	Inc-v3ens3	Inc-v3ens4	IncRes-v2ens	HGD	R&P	NIPS-r3
TI-DIM	83.8	83.1	78.5	83.0	81.7	83.7
TI-RDIM	85.0	84.9	79.1	82.1	81.2	83.9
NI-TI-DIM	86.4	84.9	79.4	82.3	81.0	84.2
RF-TI-RDIM	**91.3**	**90.1**	**82.0**	**87.9**	**86.1**	**90.7**

Table 7. The success rates (%) of black-box attacks against six defense models under multi-model setting. The adversarial examples are generated for the ensemble of Inc-v3, Inc-v4, IncRes-v2, Res-v2-152 using SI-NI-TIM, SI-NI-TI-DIM, DE-NI-TIM, DE-TIM and RF-DE-TIM

Attack	Inc-v3ens3	Inc-v3ens4	IncRes-v2ens	HGD	R&P	NIPS-r3
SI-NI-TIM	79.5	79.1	70.3	73.4	71.5	77.2
SI-NI-TI-DIM	87.2	85.6	77.7	82.3	81.6	84.5
DE-NI-TIM	81.5	79.6	69.8	76.1	74.8	78.6
DE-TIM	91.2	90.7	88.2	90.5	90.1	91.1
RF-DE-TIM	**94.7**	**94.5**	**89.1**	**93.2**	**92.7**	**93.9**

Table 7 shows that our method DEM and region fitting can be easily integrated into state-of-the-art attack methods and improve the transferability of adversarial examples. The experimental results prove the fourth analysis in Sec. 3.2. In particular, our best attack RF-DE-TIM fools six defense models with a 93% success rate on average. Such high success rates mean that there is an urgent need to develop more defensive methods to resist adversarial examples.

5 Conclusion

In this paper, we introduce a three stage pipeline: resized-diverse-inputs (RDIM), diversity-ensemble (DEM) and region fitting, that work together to generate transferable adversarial examples. We first explore the internal relationship between DIM and TIM, and propose RDIM that is more suitable to characterize this relationship. Combined with TIM, RDIM can balance the contradiction between loss of gradient information and stripes demand. Then we propose DEM, the multi-scale version of RDIM, to generate multi-scale gradients with different numbers of vertical and horizontal stripes for TIM. After the first two steps we transform value fitting into region fitting across iterations. RDIM and region fitting do not require extra running time and these three steps can be well integrated into other attacks. Our best attack RF-DE-TIM fools six black-box defenses with a 93% attack success rate on average, which is higher than the state-of-the-art multi-model attacks. We hope that our findings about attack methods will shed light into potential future directions for adversarial attacks.

References

1. Arnab, A., Miksik, O., Torr, P.H.S.: On the robustness of semantic segmentation models to adversarial attacks. In: 2018 IEEE Conference on Computer Vision and Pattern Recognition. pp. 888–897 (2018) 4
2. Athalye, A., Engstrom, L., Ilyas, A., Kwok, K.: Synthesizing robust adversarial examples. In: Proceedings of the 35th International Conference on Machine Learning. pp. 284–293 (2018) 1
3. Biggio, B., Corona, I., Maiorca, D., Nelson, B., Srndic, N., Laskov, P., Giacinto, G., Roli, F.: Evasion attacks against machine learning at test time. In: Machine Learning and Knowledge Discovery in Databases. pp. 387–402 (2013) 4
4. Bose, A.J., Aarabi, P.: Adversarial attacks on face detectors using neural net based constrained optimization. In: 20th IEEE International Workshop on Multimedia Signal Processing. pp. 1–6 (2018) 4
5. Brendel, W., Rauber, J., Bethge, M.: Decision-based adversarial attacks: Reliable attacks against black-box machine learning models. In: 6th International Conference on Learning Representations (2018) 4
6. Carlini, N., Wagner, D.A.: Towards evaluating the robustness of neural networks. In: 2017 IEEE Symposium on Security and Privacy. pp. 39–57 (2017) 4
7. Chen, J., Jordan, M.I.: Boundary attack++: Query-efficient decision-based adversarial attack. CoRR **abs/1904.02144** (2019) 4
8. Dong, Y., Liao, F., Pang, T., Su, H., Zhu, J., Hu, X., Li, J.: Boosting adversarial attacks with momentum. In: 2018 IEEE Conference on Computer Vision and Pattern Recognition. pp. 9185–9193 (2018) 4, 5, 6, 8
9. Dong, Y., Pang, T., Su, H., Zhu, J.: Evading defenses to transferable adversarial examples by translation-invariant attacks. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 4312–4321 (2019) 2, 3, 5, 6, 7, 8, 10
10. Eykholt, K., Evtimov, I., Fernandes, E., Li, B., Rahmati, A., Xiao, C., Prakash, A., Kohno, T., Song, D.: Robust physical-world attacks on deep learning visual classification. In: 2018 IEEE Conference on Computer Vision and Pattern Recognition. pp. 1625–1634 (2018) 1, 4
11. Goodfellow, I.J., Shlens, J., Szegedy, C.: Explaining and harnessing adversarial examples. In: 3rd International Conference on Learning Representations (2015) 1, 4, 5
12. Guo, C., Rana, M., Cissé, M., van der Maaten, L.: Countering adversarial images using input transformations. In: 6th International Conference on Learning Representations (2018) 1
13. He, K., Zhang, X., Ren, S., Sun, J.: Identity mappings in deep residual networks. In: Computer Vision - ECCV 2016 - 14th European Conference. pp. 630–645 (2016) 10
14. Kurakin, A., Goodfellow, I.J., Bengio, S.: Adversarial examples in the physical world. In: 5th International Conference on Learning Representations (2017) 1, 4, 6
15. Kurakin, A., Goodfellow, I.J., Bengio, S.: Adversarial machine learning at scale. In: 5th International Conference on Learning Representations (2017) 4, 5
16. Liao, F., Liang, M., Dong, Y., Pang, T., Hu, X., Zhu, J.: Defense against adversarial attacks using high-level representation guided denoiser. In: 2018 IEEE Conference on Computer Vision and Pattern Recognition. pp. 1778–1787 (2018) 1, 5, 9
17. Lin, J., Song, C., He, K., Wang, L., Hopcroft, J.E.: Nesterov accelerated gradient and scale invariance for improving transferability of adversarial examples. CoRR **abs/1908.06281** (2019) 4, 5, 6, 7, 10

18. Liu, Y., Chen, X., Liu, C., Song, D.: Delving into transferable adversarial examples and black-box attacks. In: 5th International Conference on Learning Representations (2017) 1
19. Madry, A., Makelov, A., Schmidt, L., Tsipras, D., Vladu, A.: Towards deep learning models resistant to adversarial attacks. In: 6th International Conference on Learning Representations (2018) 1, 2
20. Moosavi-Dezfooli, S., Fawzi, A., Fawzi, O., Frossard, P.: Universal adversarial perturbations. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition. pp. 86–94 (2017) 1
21. Narodytska, N., Kasiviswanathan, S.P.: Simple black-box adversarial perturbations for deep networks. CoRR **abs/1612.06299** (2016) 4
22. Pang, T., Du, C., Zhu, J.: Max-mahalanobis linear discriminant analysis networks. In: Proceedings of the 35th International Conference on Machine Learning. pp. 4013–4022 (2018) 1
23. Raghunathan, A., Steinhardt, J., Liang, P.: Certified defenses against adversarial examples. In: 6th International Conference on Learning Representations (2018) 1
24. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M.S., Berg, A.C., Li, F.: Imagenet large scale visual recognition challenge. International Journal of Computer Vision **115**(3), 211–252 (2015) 9
25. Samangouei, P., Kabkab, M., Chellappa, R.: Defense-gan: Protecting classifiers against adversarial attacks using generative models. In: 6th International Conference on Learning Representations (2018) 1
26. Szegedy, C., Ioffe, S., Vanhoucke, V., Alemi, A.A.: Inception-v4, inception-resnet and the impact of residual connections on learning. In: Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence. pp. 4278–4284 (2017) 10
27. Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z.: Rethinking the inception architecture for computer vision. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition. pp. 2818–2826 (2016) 10
28. Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I.J., Fergus, R.: Intriguing properties of neural networks. In: 2nd International Conference on Learning Representations (2014) 1, 4
29. Tramèr, F., Kurakin, A., Papernot, N., Goodfellow, I.J., Boneh, D., McDaniel, P.D.: Ensemble adversarial training: Attacks and defenses. In: 6th International Conference on Learning Representations (2018) 1, 2, 5, 9
30. Wong, E., Kolter, J.Z.: Provable defenses against adversarial examples via the convex outer adversarial polytope. In: Proceedings of the 35th International Conference on Machine Learning. pp. 5283–5292 (2018) 1
31. Xie, C., Wang, J., Zhang, Z., Ren, Z., Yuille, A.L.: Mitigating adversarial effects through randomization. In: 6th International Conference on Learning Representations (2018) 1, 5, 9
32. Xie, C., Zhang, Z., Zhou, Y., Bai, S., Wang, J., Ren, Z., Yuille, A.L.: Improving transferability of adversarial examples with input diversity. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR2019. pp. 2730–2739 (2019) 2, 3, 5, 6