World Scientific Annual Review of Artificial Intelligence
To appear in 2023

Black-Box Attack using Adversarial Examples: A New Method of Improving Transferability

Tao Wu

Department of Computer Science, Missouri University of Science and Technology
Rolla, MO 65409, United States
wuta@mst.edu

Tie Luo

Department of Computer Science, Missouri University of Science and Technology
Rolla, MO 65409, United States
tluo@mst.edu

Donald C. Wunsch

Department of Electrical and Computer Engineering, Missouri University of Science and Technology
Rolla, MO 65409, United States
dwunsch@mst.edu

Received 18 August 2022
Revised 24 November 2022
Accepted 8 December 2022

Adversarial examples (AE) are malicious test-data samples (typically images) generated by applying carefully calculated perturbations to clean samples. The added perturbations are usually human-imperceptible but the AE can fool a machine learning (ML) model to make misclassifications. Although multiple methods were proposed to generate AE, the ability to *generalize* is very limited; that is, they easily overfit to their source, single, white-box ML models and the generated AE rarely work for other models. In this paper, we propose a *black-box attack approach* that crafts *transferable* AE that can attack a wide range of ML models without knowing those model details. Our novel method consists of an *elastic momentum* (EM) that expedites gradient descent to avoid early overfitting, and a *random erasure* (RE) technique that increases the diversity of perturbations and reduces gradient fluctuations. Our method can be applied to any gradient-based attacks to make those attacks become more transferable. We evaluate our proposed method by attacking seven state-of-the-art (SOTA) deep learning models and compare against five SOTA attacks; we also attack nine advanced defense mechanisms that are integrated into the above models. Our results demonstrate significant improvement on the attack success rate (ASR) and transferability when using our method alone, and that it can also be easily applied to other baseline methods (which are gradient-based) to substantially improve *their* performance as well.

Keywords: Adversarial example, deep learning, neural networks, computer vision, fast gradient sign method.

1. Introduction

Deep Neural Networks (DNNs) have made resounding success in computer vision tasks. However, they are vulnerable to *adversarial examples* (AE), which are data samples (typically images) that are perturbed by human-imperceptible noises yet result in misclassifications. This can cause serious safety and security consequences in applications such as autonomous driving and medical diagnosis. The *transferability* of AE is an active research area [7, 8, 12, 18, 20, 22, 35, 37, 38, 40, 45] that studies how well an AE created to attack (fool) a "source" model can successfully fool other "target" models as well. The rationale of studying this is that (1) from an attacker's perspective, good transferability implies that one can launch *black-box attacks* on target models (without knowing their internal structure, algorithmic details, or parameters); (2) from a defender's perspective, studying it provides insight into understanding the failure and vulnerability of DNNs and how to design DNNs that are robust to AE.

The techniques proposed in the literature to improve the transferability of AE include gradient or momentum based methods [7, 20, 35, 37], ensemble methods [18, 22], image transformations based methods [8,20,40,45], and network architecture alterations [12,38]. A major issue of these techniques attempt to address is that AE created on a source model (in order to attack it) can be easily trapped into the exclusive blind spots of the source model and can hardly generalize to other (target) models; in other words, this can be viewed as an problem of AE *overfitting*.

In this paper, we propose a new method of crafting AE and thereby improving their transferability. This method consists of two techniques: *elastic momentum* (EM) and *random erasure* (RE). We first introduce EM into the AE generation process to compute gradients in a much expedited manner insofar as the training will converge earlier than reaching the overfitting region. We also propose to incorporate RE, which is a data augmentation technique, into the AE crafting procedure for the first time. The contributions of this paper are summarized as follows:

- We introduce a new black-box approach of crafting transferable AE by proposing EM and a new usage of RE. EM generalizes the conventional momentum and the Nesterov's momentum methods by computing gradients over a flexible look-ahead horizon, and RE increases the diversity of adversarial perturbations and helps stabilize gradient fluctuations.
- Besides transferability, our proposed method is very flexible in that it can be applied to any existing gradient-based attacks to enhance *their* effectiveness.
- Through extensive evaluation with 5 recent baseline methods, 7 target deep learning models, and 9 advanced defense mechanisms, we demonstrate the superior transferability of our proposed black-box attack approach.

The reminder of this paper is organized as follows.Section 2 reviews the related work on AE attacks and defenses. Preliminary formulations on AE are discussed in section 3. Section 4 presents our proposed approach to boosting AE transferability. The performance of our method is then evaluated in comparison to the state-of-the-art methods in section 5.

Finally, Section 6 concludes the paper.

2. Related work

2.1. *Adversarial Attacks*

Based on adversary's knowledge to the model, adversarial attacks can be grouped into *white-box attacks, black-box attacks*. In white-box setting, one assumes the attackers possess perfect knowledge about the target model, including the architecture, parameters, and gradient of the loss w.r.t. the input. Most methods adopt the gradient information of the target model to launch adversarial attacks under the white-box setting. For example, Fast Gradient Sign Method (FGSM) [10] generates an adversarial example by taking a single step within a small distance ϵ along the loss function's gradient direction, Project Gradient Descent (PGD) [24] extends FGSM by iteratively taking multiple small gradient steps and projecting the generated adversarial example onto the ϵ-sphere around the clean sample at each step, and Carlini and Wagner Attack (C&W) [3] reformulates the constrained loss into an Lagrangian form and adopts Adam [15] for optimization. However, white box attacks are almost unrealistic in real applications because the model structure and parameters are usually hidden from the attackers.

Based on the different level of adversary's knowledge to the model, black-box attacks can be further grouped into three scenarios, including score-based, decision-based and transfer-based attack. Score-based black-box attacks can acquire the output probabilities by querying the target model, and the gradient can be estimated through queries. For example, Zeroth Order Optimization (ZOO) [4] estimates the gradient by finite differences and then adopts C&W attacks based on the estimated gradients. Decision-based black-box attacks can only solely rely on the predicted classes of the queries, this setting is more challenging since the target model only provides discrete hard-label predictions. [5] formulates the hard-label black-box attack as a real-valued optimization problem which can be solved by any zeroth order optimization algorithm. Transfer-based black-box attacks require the least knowledge of the target model which are based on the transferability of adversarial examples [31]. Transfer-based black-box attacks are the topic we study in this work where we apply white-box attacks on surrogate models to find adversarial examples that are then transferred to black-box target models. In this setting, the most important aspect is to improve the transferability of adversarial examples so that transfer-based black-box attacks can be made more effective in real world scenarios. Many works have been proposed to this direction. Momentum Iterative Method (MIM) [7] integrates a momentum term into the gradient calculation which stabilize the update direction and boost the transferability by a large extent. The Diverse Inputs Method (DIM) [40] applies the gradient of the randomly resized and padded input for transferable adversarial example generation. Advance gradient calculation and data augmentation are the key parts in creating adversarial examples with high transferablity, in this work, we are going further along this direction by proposing elastic momentum and random erasure to boost the adversarial transferability by a large margin.

2.2. *Defend against Adversarial Attacks*

Due to the threat of adversarial examples, extensive research efforts have been put on building robust models to defend various against adversarial attacks. There are roughly three lines of research direction on adversarial robustness. The first one is adversarial training [10, 24, 32], which inject the generated adversarial samples into the training data to help model discriminate adversarial examples. For example, [24] proposes to augment the training data with adversarial examples crafted by PGD attack which remains the state-of-the-art defense to date. While adversarial training is promising, it is computationally expensive and hard to scale to large-scale datasets [17].

The second line of defenses proceeds by input transformation. Specifically, this kind of approaches firstly preprocess the input images to rectify adversarial perturbations without reducing the classification accuracy on benign images. The input transformation methods include random resizing and padding [39], JPEG compression [9], bit-depth reduction [41], total variance minimization [11], autoencoder-based denoising [19], and so on. However, this kind of defenses can cause shattered gradients or vanishing/exploding gradients, which can be evaded by adaptive attacks [1].

The last category is certified defenses, which are mathematical provably robust to the worst-case attacks under some assumptions. The motivation of certified defenses is to end the long-standing arms race between adversarial defenders and attackers. Recent certified defenses [6, 42] have been made scalable to ImageNet, showing the applicability of this type of defenses.

Besides above, model ensemble is another effective defense strategy in practice which leverage the outputs from an ensemble of individual models [21, 27]. Model ensemble can be integrated with the above defenses such as ensemble adversarial training [32] which greatly boost the robustness of adversarial training.

3. Preliminaries

Let x be a benign image, y the corresponding true label and $f(x;\theta)$ the classifier with parameters θ and which outputs the prediction result. Let $J(x, y; \theta)$ denote the loss function (e.g., cross-entropy loss) of the classifier f. We define an adversarial attack as finding an adversarial example x^{adv} that satisfies $\|x^{adv} - x\|_p \leq \epsilon$ but incurs misclassification to the model, i.e., $f(x;\theta) \neq f(x^{adv};\theta)$. Here $\|\cdot\|_p$ denotes p-norm and we consider $p = \infty$ in this paper to be consistent with previous works. Mathematically, given a benign (clean) example x, we seek to find an AE x^{adv} as the solution to the following constrained optimization problem:

$$\arg\max_{x^{adv}} J(x^{adv}, y; \theta), \quad \text{s.t. } \left\|x^{adv} - x\right\|_\infty \leq \epsilon \tag{1}$$

As mentioned earlier, gradient-based methods have been shown to be the most effective to solve the above problem and we focus on this category of methods, for which the representative ones are described below.

Fast Gradient Sign Method (FGSM). FGSM [10] is the first gradient-based attack which crafts an adversarial example x^{adv} by attempting to maximize the loss function

$J(x^{adv}, y; \theta)$ with a one-step update:

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

where $\nabla_x J(x, y; \theta)$ is the gradient of loss function with respect to x, and $\text{sign}(\cdot)$ denotes the sign function.

Iterative Fast Gradient Sign Method (I-FGSM). I-FGSM [16] extends FGSM to an iterative version:

$$\begin{aligned} x_{t+1}^{adv} &= x_t^{adv} + \alpha \cdot \text{sign}(\nabla_{x_t^{adv}} J(x_t^{adv}, y; \theta)), \\ x_0^{adv} &= x, \end{aligned} \tag{3}$$

where $\alpha = \epsilon/T$ is a small step size and T is the number of iterations.

Momentum Iterative Fast Gradient Sign Method (MI-FGSM). MI-FGSM [7] integrates a momentum term into I-FGSM and achieves much better transferability:

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

where $g_0 = 0$ and μ is a decay factor.

Nesterov Iterative Fast Gradient Sign Method (NI-FGSM) [20] integrates Nesterov's accelerated gradient (NAG) [26] into the iterative attack method, by replacing x_t^{adv} in (4) with $\tilde{x}_t^{adv}$ which is defined as

$$\tilde{x}_t^{adv} = x_t^{adv} + \alpha \cdot \mu \cdot g_t \tag{5}$$

Other notable methods for boosting gradient-based adversarial attacks are described below as well, some of which are also image transformation methods (DIM, TIM, SIM) and we will use as our baselines in addition to the above. Diverse Inputs Method (DIM) [40] applies random resizing and padding with a given probability to the inputs and uses the transformed images for gradient calculation. Translation-Invariant Attacks Method (TIM) [8] generates more transferable AE by optimizing the perturbation over a set of translated images. Scale-Invariant attack Method (SIM) [20] uses scaling rather than translation as the data augmentation technique. Variance Tuning Gradient-based attack (VNI-FGSM) [35] considers the gradient variance of the previous iteration to tune the current gradient to stabilize the update direction. The Admix Attack method (Admix) [36] mixes the input image with a small portion of another randomly selected image and calculates the gradient based on an ensemble of scaled copies of mixed image. Adam Iterative Fast Gradient Tanh Method (AI-FGTM) [44] replace the momentum algorithm and the sign function with Adam and the tanh function which boost the indistinguishability and transferability of adversarial examples. Object based Diverse Input (ODI) [2] draws an adversarial image on a 3D object which effectively diversifies the input by leveraging an ensemble of multiple source objects and randomizing viewing conditions.

Despite some improvement, the above prior works have limited success rate on unknown target models especially when target models have protection or defense in place [6].

4. Proposed Method

4.1. *Elastic Momentum*

We make two key observations. First, the main reason why integrating momentum benefits AE computation is because the momentum essentially combines several steps of (potentially discounted) gradients together to help stablize gradient descent and obtain a more robust direction of convergence. Second, the reason why Nesterov's accelerated gradient can benefit it even further is because it computes the gradients based on an estimated next-step AE, rather than the last-step AE, which speeds up the training.

Thus, our basic idea is as follows. First, generalize the prediction of next-step AE, by allocating a *flexible look-ahead horizon* for computing an estimated *future AE*. Next, compute the gradient using that future AE to obtain a more *far-sighted* momentum, which *accelerates* the convergence (with reduced number of iterations) and thereby prevents overfitting.

Formally, an AE x^{adv} is computed iteratively as follows:

$$x_t^{em} = x_t^{adv} + \alpha \cdot \sigma \cdot g_t, \tag{6}$$

$$g_{t+1} = \mu \cdot g_t + \frac{\nabla_{x_t^{em}} J(x_t^{em}, y; \theta)}{\|\nabla_{x_t^{em}} J(x_t^{em}, y; \theta)\|_1}, \tag{7}$$

$$x_{t+1}^{adv} = x_t^{adv} + \alpha \cdot \text{sign}(g_{t+1}). \tag{8}$$

The momentum term g accumulates previous gradients with a decay factor μ, while the gradient is not computed based on the current AE x_t^{adv} but a future AE x_t^{em} estimated over a look-ahead horizon. The parameter σ is critical: although μ has to be a value extremely close to 1, as experimentally shown by [7], σ is independent of μ (as opposed to NI-FGSM) and could take a value much larger than 1, which essentially means that we can use g_t to approximate g_{t+1}, g_{t+2}, ... and tune the length of this look-ahead horizon to achieve the best transferability. For this reason, we call the momentum term g an elastic momentum (EM). Fig. 1 illustrates our method EM as compared to NI-FGSM.

Our approach also generalizes MI-FGSM and NI-FGSM which can be viewed as special cases of ours: When $\sigma = 0$, we obtain the momentum iterative method MI-FGSM; when $\sigma = \mu$, we obtain Nesterov's momentum method NI-FGSM. Note, however, that we typically do *not* use these σ values in order to achieve acceleration and thus better performance. In fact, our method gives us flexibility to control the converging process via σ, in order to reach a local optimum before hitting the overfitting region, thereby obtaining better AE transferability.

4.2. *Using Random Erasure in AE Generation*

Previous work [40] has demonstrated that random transformations of input images such as random resizing and random padding could boost the transferability of adversarial examples. However, what specific type of transformation is better remains an open question. In our work, we hypothesize that partial occlusion would make the resulting AE more transferable, and the rationale is as follows. A classification model usually examines dif-

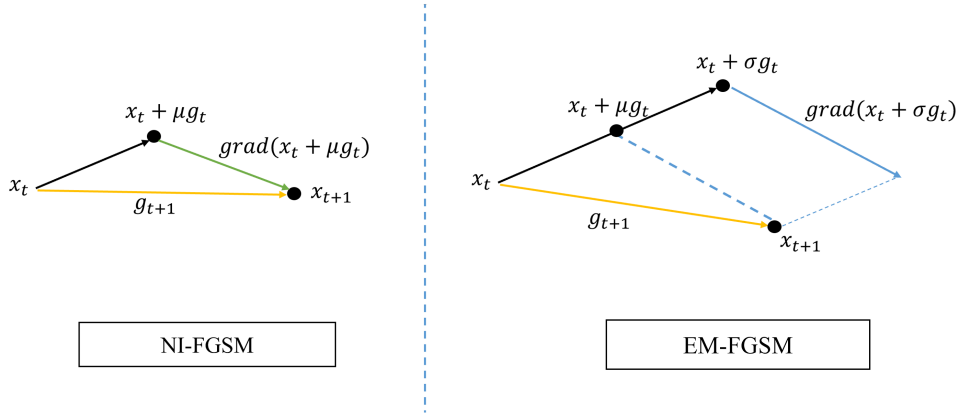


Fig. 1: Illustration of EM as compared to NI-FGSM.

ferent regions of an image to recognize its category, which is why a white-box attack could achieve near 100% attack success rate whereas black-box transferred AE are much less likely to fool target models since those models tend to ignore the adversarial regions. However, when an image is partially occluded, a model will classify it based on the overall object structure. Thus, if we use occluded adversarial images during AE generation, the AE generation process will make the non-occluded region of the object structure adversarial, and as a result, the generated AE will be more transferable and more likely to fool other target models. Similar techniques are also proposed in [34, 43] as a generic data augmentation technique for deep learning to address data insufficiency which bring benefits to the task of image classification, object detection and person re-identification. In this paper, however, we apply RE to AE generation which has never been explored before. In addition, we identify that RE is the most suitable candidate for AE transferability through our comparison with many other data augmentation techniques such as translation, scaling, rotation, resizing, padding, weighting, and even a nearest neighbor method that we created on our own.

Given an image I with width W and height H, we apply RE by randomly selecting a rectangle region I_e in I and removes the pixels in the region I_e. This region is determined as follows. Denoting by S_e the area of the region I_e, we randomly generate an erasure ratio s in the range $[0, s_h]$ where $s_h < 1$, and use $s = \frac{S_e}{S}$ to determine the value of S_e, where S is the area of the input image I, i.e., $S = W \times H$. Now, denote the aspect ratio of I_e by r_e. The height and width of I_e are therefore determined by $H_e = \sqrt{S_e \times r_e}$ and $W_e = \sqrt{\frac{S_e}{r_e}}$, respectively. To determine the location of I_e, we randomly pick a point $\mathcal{P} = (x_e, y_e) \in I$, until $x_e + W_e \leq W$ and $y_e + H_e \leq H$, upon which we finalize the coordinates of the erasure region $I_e = (x_e, y_e, x_e + W_e, y_e + H_e)$. An example is given in Figure 2.

To remove the pixels in the region I_e, there are three typical choices, namely using 0s, 1s and random noise, to fill the region. Our experiments show that they do not make a notable difference in performance. Hence, we adopt random noise in this paper.

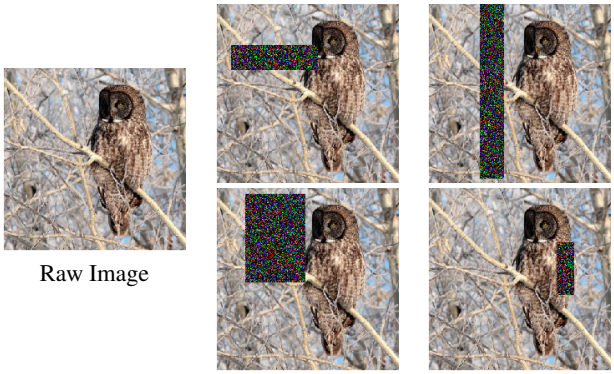


Fig. 2: Example: Applying RE to an raw image to generate four resulting images.

Now, we reformulate the original objective function (1) by incorporating RE, as

$$\begin{aligned} \arg\max_{x^{adv}} \frac{1}{m}\sum_{i=0}^{m} J(RE_i(x^{adv}), y; \theta), \\ \text{s.t.} \left\|x^{adv} - x\right\|_\infty \leq \epsilon, \end{aligned} \tag{9}$$

where m is the number of erasure copies, and $i = 0$ represents the input image without erasure.

Thus, our AE crafting process, integrated RE, is as follows. At each iteration t, with probability p, we apply RE to the input image x_t^{adv} to generate a collection of m erased images, and compute their losses and the average gradient $\frac{1}{m}\sum_{i=0}^{m} \nabla_x J(RE_i(x^{adv}), y; \theta)$, which will be used to compute the momentum g. With probability $1-p$, we keep the input image x^{adv} intact.

To incorporate RE into EM, however, the above x^{adv} needs to be replaced by x^{em} defined by (6). Therefore, our final proposed method is formulated as

$$x_t^{em} = x_t^{adv} + \alpha \cdot \sigma \cdot g_t, \tag{10}$$

$$g_{t+1} = \mu \cdot g_t + \frac{\frac{1}{m}\sum_{i=0}^{m} \nabla_x J(RE_i(x_t^{em}), y; \theta)}{\|\frac{1}{m}\sum_{i=0}^{m} \nabla_x J(RE_i(x_t^{em}), y; \theta)\|_1}, \tag{11}$$

$$x_{t+1}^{adv} = x_t^{adv} + \alpha \cdot \text{sign}(g_{t+1}). \tag{12}$$

Algorithm 1 summarizes our proposed method, where $RE(x; p)$ denotes that we apply RE with probability p.

5. Experiments

This section reports the evaluation of our proposed method, EM-RE-FGSM.

5.1. *Experiment Setup*

Dataset. We use an image dataset [28] which is a curated portion of ImageNet and is widely used such as by [20, 35]. This dataset randomly selects one clean and correctly classified

Algorithm 1 Proposed Method: EM-RE-FGSM

Input: A clean example x with ground-truth label y; a classifier f with loss function J;
Input: Perturbation size ϵ; maximum iterations T; decay factor μ; look-ahead parameter σ; number of random erasure copies m; random erasure probability p.
Output: An adversarial example x^{adv}

$\alpha = \epsilon/T$
$g_0 = 0; x_0^{adv} = x$
for $t = 0$ to $T - 1$ **do**
  Compute $x_t^{em} = x_t^{adv} + \alpha \cdot \sigma \cdot g_t$
  $g = 0$
  for $i = 0$ to $m - 1$ **do**
    Compute gradient $\nabla_x J(RE_i(x_t^{em}; p), y; \theta)$
    Update $g = g + \nabla_x J(RE_i(x_t^{em}; p), y; \theta)$
  Average momentum as $g = \frac{g}{m}$
  Update g_{t+1} as $g_{t+1} = \mu \cdot g_t + \frac{g}{\|g\|_1}$
  Update $x_{t+1}^{adv} = x_t^{adv} + \alpha \cdot \text{sign}(g_{t+1})$
return $x^{adv} = x_T^{adv}$

Source model	Attack	Inc-v3	Inc-v4	IncRes-v2	Res-101	Inc-v3$_{ens3}$	Inc-v3$_{ens4}$	IncRes-v2$_{ens1}$
Inc-v3	MI-FGSM	**100.0***	43.6	42.4	35.7	13.1	12.8	**6.2**
	NI-FGSM	**100.0***	51.7	50.3	41.3	**13.5**	**13.2**	6.0
	EM-FGSM	**100.0***	**55.0**	**52.7**	**44.6**	11.4	11.5	5.5
Inc-v4	MI-FGSM	56.3	99.7*	46.6	41.0	**16.3**	**14.8**	**7.5**
	NI-FGSM	63.1	**100.0***	51.8	45.8	15.4	13.6	6.7
	EM-FGSM	**66.9**	**100.0***	**54.4**	**47.6**	14.7	12.4	6.8
IncRes-v2	MI-FGSM	60.7	51.1	97.9*	46.8	**21.2**	**16.0**	**11.9**
	NI-FGSM	62.8	54.7	99.1*	46.0	20.0	15.1	9.6
	EM-FGSM	**65.2**	**56.2**	**99.2***	**48.7**	18.6	13.1	7.8
Res-101	MI-FGSM	58.1	51.6	50.5	99.3*	23.9	**21.5**	**12.7**
	NI-FGSM	65.6	58.3	57.0	**99.4***	**24.5**	21.4	11.7
	EM-FGSM	**65.7**	**60.9**	**61.1**	99.3*	20.8	17.6	10.0

Table 1: The attack success rates (ASR) (%) on seven target models in the **single-source-model** setting, **using EM alone**. The AE are generated using a single source model Inc-v3, Inc-v4, IncRes-v2, or Res-101. '*' indicates white-box attack.

images from each of the 1,000 categories of the ILSVRC 2012 validation dataset, and thus contains 1,000 good images with each of the size $299 \times 299 \times 3$.

Models to attack. We first consider four widely used state-of-the-art DNNs, namely Inception-v3 (Inc-v3) [30], Inception-v4 (Inc-v4) [29], Inception-Resnet-v2 (IncRes-v2) [29], and Resnet-v2-101 (Res-101) [13]. In addition, to increase the difficulty level, we also include three *adversarially trained* DNNs (and thus are more robust to AE), namely Inc-v3$_{ens3}$, Inc-v3$_{ens4}$ and IncRes-v2$_{ens1}$ [33]. The first two models are Inc-v3 trained on

Source model	Attack	Inc-v3	Inc-v4	IncRes-v2	Res-101	Inc-v3$_{ens3}$	Inc-v3$_{ens4}$	IncRes-v2$_{ens1}$
Inc-v3	DIM	99.0*	64.3	60.9	53.2	19.9	18.3	9.3
	TIM	100.0*	48.8	43.6	39.5	24.8	21.3	13.2
	SIM	100.0*	69.4	67.3	62.7	32.5	30.7	17.3
	RE	**100.0***	**71.1**	**68.7**	**64.3**	**33.1**	**31.6**	**19.0**
Inc-v4	DIM	72.9	97.4*	65.1	56.5	20.2	21.1	11.6
	TIM	58.6	99.6*	46.5	42.3	26.2	23.4	17.2
	SIM	80.6	99.6*	74.2	68.8	47.8	44.8	29.1
	RE	**82.3**	**99.8***	**76.3**	**71.5**	**49.6**	**45.9**	**31.4**
IncRes-v2	DIM	70.1	63.4	93.5*	58.7	30.9	23.9	17.7
	TIM	62.2	55.4	97.4*	50.5	32.8	27.6	23.3
	SIM	84.7	81.1	99.0*	76.4	56.3	48.3	42.8
	RE	**86.2**	**83.3**	**99.4***	**78.7**	**59.1**	**50.6**	**46.2**
Res-101	DIM	75.8	69.5	70.0	98.0*	35.7	31.6	19.9
	TIM	59.3	52.1	51.8	99.3*	35.4	31.3	23.1
	SIM	75.2	68.9	69.0	99.7*	43.7	38.5	26.3
	RE	**78.2**	**71.5**	**72.8**	**99.8***	**45.2**	**39.8**	**28.7**

Table 2: The attack success rates (ASR) (%) on seven target models in the **single-source-model** setting, **using RE alone**. The AE are generated using a single source model Inc-v3, Inc-v4, IncRes-v2, or Res-101. '*' indicates white-box attack.

AE generated from an ensemble of 3 and 4 other pretrained DNNs, respectively, and the last is IncRes-v2 trained on AE generated from a single pretrained DNN (it is still called an "ensemble" in [33] so we have adopted the same naming convention).

Defenses to attack. To further increase the difficulty level, we also consider *defense mechanisms* in our evaluation. As pointed out by [6], many existing attacks underperform or even fail when target models are armed with defense mechanisms. Therefore, we select nine state-of-the-art advanced defenses: the top-3 winners in the NeurIPS defense strategy competition and 6 recently proposed defense methods. The first group consists of HGD (rank-1) [19], R&P (rank-2) [39], and NIPS-r3 (rank-3), and the second group consits of Bit-Red [41], JPEG [11], FD [23], ComDefend [14], RS [6] and NRP [25]. These 9 defense methods have been integrated into their respective DNNs.

Baseline AE-generation methods. We compare our method with five recently proposed attack methods, namely MI-FGSM [7], NI-FGSM [20], Diverse Inputs Method (DIM) [40], Translation-Invariant attack Method (TIM) [8], and Scale-Invariant attack Method (SIM) [20]. The first two are momentum-based attacks and the other three are image-transformation based attacks.

Versatile as a "plug-in". As mentioned in 1, our method can be applied to any gradient-based attack method to form a new, stronger attack. We demonstrate this by integrating our method with DIM, TIM, and SIM, respectively, as well as all of them three combined together, to obtain two more attacks and include them in our evaluation as well.

Attack setup. We normalize image pixel values in $[-1, 1]$, and set the number of iterations $T = 10$, the maximum perturbation $\epsilon = 16/255$ as in [7]. For parameters related to EM, we set the decay factor $\mu = 1$ following [7] and the look-ahead parameter $\sigma = 2$ as indicated by our ablation study. For parameters related to RE, we set $s_h = 0.4$ and $r_e = 0.3$ following [43], the number of erasure copies $m = 5$ and probability $p = 0.5$. We

Model	Attack	Inc-v3	Inc-v4	IncRes-v2	Res-101	Inc-v3$_{ens3}$	Inc-v3$_{ens4}$	IncRes-v2$_{ens1}$
Inc-v3	DIM	99.0*	64.3	60.9	53.2	19.9	18.3	9.3
	EM-RE-DIM	**99.6***	**71.5**	**69.7**	**62.8**	**30.2**	**30.0**	**19.4**
Inc-v4	DIM	72.9	97.4*	65.1	56.5	20.2	21.1	11.6
	EM-RE-DIM	**81.1**	**98.6***	**75.0**	**64.2**	**28.9**	**30.1**	**18.6**
IncRes-v2	DIM	70.1	63.4	93.5*	58.7	30.9	23.9	17.7
	EM-RE-DIM	**78.9**	**72.1**	**98.9***	**65.4**	**37.9**	**31.5**	**25.6**
Res-101	DIM	75.8	69.5	70.0	98.0*	35.7	31.6	19.9
	EM-RE-DIM	**82.6**	**76.8**	**77.4**	**99.3***	**44.2**	**37.6**	**27.7**

(a) Comparison with DIM

Model	Attack	Inc-v3	Inc-v4	IncRes-v2	Res-101	Inc-v3$_{ens3}$	Inc-v3$_{ens4}$	IncRes-v2$_{ens1}$
Inc-v3	TIM	100.0*	48.8	43.6	39.5	24.8	21.3	13.2
	EM-RE-TIM	**100.0***	**56.9**	**51.2**	**46.5**	**31.1**	**28.5**	**25.9**
Inc-v4	TIM	58.6	99.6*	46.5	42.3	26.2	23.4	17.2
	EM-RE-TIM	**66.8**	**99.9***	**56.8**	**50.9**	**35.6**	**32.4**	**19.5**
IncRes-v2	TIM	62.2	55.4	97.4*	50.5	32.8	27.6	23.3
	EM-RE-TIM	**70.7**	**64.8**	**98.9***	**58.9**	**40.6**	**35.9**	**31.6**
Res-101	TIM	59.3	52.1	51.8	99.3*	35.4	31.3	23.1
	EM-RE-TIM	**66.9**	**61.3**	**59.9**	**99.8***	**41.9**	**39.6**	**30.9**

(b) Comparison with TIM

Model	Attack	Inc-v3	Inc-v4	IncRes-v2	Res-101	Inc-v3$_{ens3}$	Inc-v3$_{ens4}$	IncRes-v2$_{ens1}$
Inc-v3	SIM	100.0*	69.4	67.3	62.7	32.5	30.7	17.3
	EM-RE-SIM	**100.0***	**77.0**	**75.9**	**69.1**	**39.5**	**39.3**	**20.4**
Inc-v4	SIM	80.6	99.6*	74.2	68.8	47.8	44.8	29.1
	EM-RE-SIM	**86.9**	**99.8***	**80.3**	**76.0**	**56.1**	**52.6**	**38.9**
IncRes-v2	SIM	84.7	81.1	99.0*	76.4	56.3	48.3	42.8
	EM-RE-SIM	**88.6**	**86.9**	99.7*	**82.1**	**64.3**	**56.2**	**51.6**
Res-101	SIM	75.2	68.9	69.0	99.7*	43.7	38.5	26.3
	EM-RE-SIM	**84.9**	**75.9**	**76.9**	**99.8***	**52.6**	**49.3**	**35.2**

(c) Comparison with SIM

Table 3: ASR (%) on seven target models in the **single-source-model** setting, **using both EM and RE**. '*' indicates white-box attack.

Attack	Inc-v3	Inc-v4	IncRes-v2	Res-101	Inc-v3$_{ens3}$	Inc-v3$_{ens4}$	IncRes-v2$_{ens1}$
MI-FGSM	99.9*	98.2*	95.3*	99.9*	39.4	35.3	24.2
NI-FGSM	99.8*	99.8*	**98.9***	99.8*	41.0	33.5	23.1
EM-FGSM	**99.9***	**99.8***	98.4*	**99.9***	**43.6**	**36.1**	**25.9**

Table 4: ASR (%) on seven target models in the **ensemble-source-model** setting, **using EM alone**. The source model is the ensemble of {Inc-v3, Inc-v4, IncRes-v2, Res-101}. '*' indicates white-box attack.

use *attack success rate* (ASR) as our evaluation metric, which is the misclassification rate of a classifier when test samples are AE (we have verified that all the benign images are classified correctly in all the cases).

5.2. *Experimental Results*

Attack	Inc-v3	Inc-v4	IncRes-v2	Res-101	Inc-v3$_{ens3}$	Inc-v3$_{ens4}$	IncRes-v2$_{ens1}$
DIM	99.4*	97.4*	94.9*	99.8*	58.1	51.1	34.9
EM-RE-DIM	**99.7***	**99.1***	**97.5***	**99.8***	**64.3**	**59.7**	**41.6**
TIM	99.7*	98.9*	97.7*	99.9*	62.2	56.8	48.0
EM-RE-TIM	**99.9***	**99.3***	**98.9***	**100.0***	**68.9**	**64.1**	**56.4**
SIM	99.7*	99.0*	97.6*	**100.0***	78.8	73.9	59.5
EM-RE-SIM	**99.8***	**99.3***	**98.4***	**100.0***	**84.3**	**79.5**	**66.8**
Composite	99.6*	98.9*	97.8*	99.7*	91.1	90.3	86.8
EM-RE-Composite	**99.8***	**99.3***	**98.4***	**99.8***	**92.3**	**91.6**	**88.6**

Table 5: ASR (%) on seven target models in the **ensemble-source-model** setting, using both EM and RE. The source model is the ensemble of {Inc-v3, Inc-v4, IncRes-v2, Res-101}. *Composite* model is the combination of DIM, TIM, and SIM. '*' indicates white-box attack.

5.2.1. *Single source model*

In this section, we evaluate the case that AE are trained on a single source model and then used to attack multiple target models. We test four source models: Inc-v3, Inv-v4, IncRes-v2, and Res-101, and the target models are these four as well as the three ensemble models, i.e., Inc-v3$_{ens3}$, Inc-v3$_{ens4}$ and IncRes-v2$_{ens1}$.

Using EM alone (without RE). We first evaluate the EM approach only, without using RE. The results are presented in Table 1. First, under white-box attacks (source model is also the target model), all the methods achieve close to 100% ASR as expected. These mean that, although our method focuses on improving black-box performance (due to transferability), we do not sacrifice any white-box performance either. Second, let us look at black-box attacks (target model is different from the source model), which are more important since they particularly reflect the *transferability* of AE. We see that EM achieves the higher ASR in about 60% of the cases while MI-FGSM and NI-FGSM perform the best in about 30% and 10% of the cases, respectively. Note that we have not activated RE yet. Taking a closer look, one can observe that the cases where our proposed EM method outperforms MI-FGSM and NI-FGSM are normally trained models, and the cases in which it does not (but still keeps a comparable performance) are those three adversarially trained models. The reason behind this is that, for normally trained models, EM achieves better optimum in the constrained iterative steps and hence demonstrates better transferability; but on the other hand, the three ensemble adversarially trained models, i.e., Inc-v3$_{ens3}$, Inc-v3$_{ens4}$ and IncRes-v2$_{ens1}$, augmented their training data with AE crafted on other static pre-trained models, and hence were trained to resist transferable AE, making black-box attacks ineffective. Therefore, to achieve higher ASE against such adversarially trained models, we need to increase the diversity of perturbations in AE, which precisely moti-

vated our introduction of our second technique, Ransom Erasure (RE). By combining with RE, our method achieves much higher ASR against ensemble adversarially trained models, as shown later.

To offer a visual intuition, we also give some example AE images generated by all these methods, in Fig. 3. It shows that all the adversarial images are very similar to the original raw image as perceived by human eyes.

Using RE alone (without EM). We then evaluate the RE approach only, without using EM. The results are presented in Table 2. The results indicate that in all the cases our proposed RE consistently outperforms DIM, TIM and SIM by a large margin, which means RE yields higher transferablity on all the black-box models while maintaining high attack success rates on the white-box setting. For instance, if we craft adversarial examples on IncRes-v2 model where our white-box attack achieves 99.4% success rate, RE yields 78.7% ASR on Res-101 which is a black-box setting; in comparison, TIM only achieves an ASR of 97.4% and 50.5%, respectively, in the same two settings. This set of results validate the effectiveness of our proposed RE method.

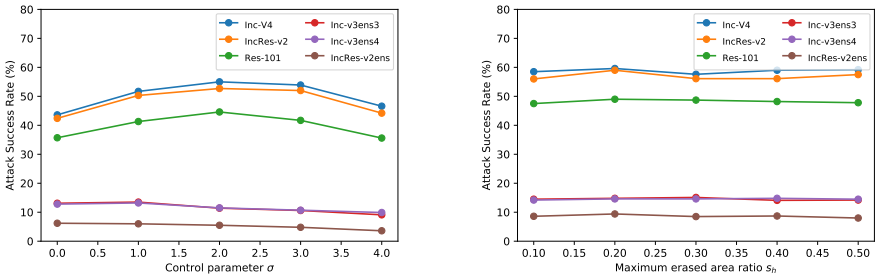
Raw Image MI-FGSM NI-FGSM EM-FGSM

Fig. 3: Adversarial images crafted by MI-FGSM [7], NI-FGSM [20] and our EM approach on the Inc-v3 model [30] with the maximum perturbation $\epsilon = 16/255$.

Using both EM and RE. Next, we add both EM and RE into the evaluation and present the results in Table 3. To demonstrate that our proposed method is versatile in that it can be applied to any gradient-based attacks to form new attacks, we apply it to DIM [40], TIM [8] and SIM [20]. Table 3 shows that our method constantly achieves the highest ASR in all the $4 \times 7 \times 3 = 84$ cases, including the 12 white-box and the 72 black-box attack settings. The wining margin is remarkable too, mostly between 20-50%. This set of results demonstrate the superior transferability of our proposed EM-RE method.

5.2.2. *Ensemble source model*

Crafting AE on an ensemble of models has been shown to be effective to improve AE transferability [7,22]. In this section, we evaluate the performance over an ensemble model of four: Inc-v3, Inc-v4, IncRes-v2 and Res-101, by averaging their logit outputs when calculating the gradients [7]. The results of using EM alone are summarized in Table 4. We observe that EM achieves the highest ASR in all the black-box attack scenarios.



(a) Impact of σ on ASR.

(b) Impact of s_h on ASR.

Fig. 4: Ablation study on σ (look-ahead horizon) and s_h (max. erasure area). The source model is Inc-v3 and the 6 target models under attack are indicated by the legend.

Source	Attack	HGD	R&P	NIPS-r3	Bit-Red	JPEG	FD	ComDefend	RS	NRP	Average
Inc-v3	MI-Composite	56.6	44.9	52.5	36.2	77.3	60.0	80.1	40.3	29.3	53.0
	NI-Composite	50.4	39.4	47.4	34.3	76.0	58.6	77.7	36.9	24.8	49.5
	EM-RE-Composite	**59.6**	**48.3**	**55.9**	**39.6**	**81.1**	**65.5**	**82.3**	**45.4**	**33.1**	**56.8**
Ensemble	MI-Composite	91.0	87.7	89.0	75.9	94.2	88.8	95.1	68.1	76.1	85.1
	NI-Composite	91.3	85.6	89.0	72.3	95.9	89.5	95.4	63.2	69.5	83.5
	EM-RE-Composite	**92.9**	**89.6**	**91.8**	**79.3**	**96.9**	**92.4**	**96.4**	**74.3**	**80.1**	**88.2**

Table 6: ASR (%) on 9 advanced defense mechanisms. *Composite* refers to the combination of DIM, TIM, and SIM.

Next, we apply both EM and RE to DIM, TIM and SIM, respectively, to form three new models. In addition, we create a new attack *Composite* by combining DIM, TIM and SIM together which forms the strongest baseline. On top of that, we apply EM and RE to *Composite* to obtain an enhanced attack using our method. We evaluate these 8 attacks and report their performance in Table 5. It shows that our proposed method again yields the best ASR in all the white-box and black-box attacks (4×7 cases), outperforming all the baselines by up to 17.5%.

5.2.3. *Attacking Advanced Defense Mechanisms*

Although our proposed method exhibits superior performance on both regularly and adversarially trained deep models, there is still a question left as to whether it will perform well against models that are protected by more sophisticated mechanisms. As pointed out by [6], many existing attacks underperform or even fail when target models have additional defense mechanisms. Motivated by this, we select 9 advanced defense mechanisms to attack, as described in our experiment setup, for the purpose of a more thorough evaluation.

We use Inc-v3 and the ensemble of {Inc-v3, Inc-v4, IncRes-v2, Res-101} as the source models to train AE, and attack the above 9 advanced defense mechanisms. We further

create more baseline attacks by combining MI and NI respectively with the *Composite* (MI and NI do have this similar "plug-in" kind of advantage as our method, but most other methods in the literature do not have). The results are given in Table 6. In this case, there is no white-box attack and all attacks are black-box. We observe that our proposed EM-RE approach is the best performer in all the scenarios, with a substantial winning margin, up to 25.4%.

5.2.4. *Ablation Study on Hyper-parameters*

We also conduct ablation experiments to study the impact of the hyper-parameters on the performance of our approach. Two key parameters are σ which determines the look-ahead horizon in EM, and s_h which determines the maximum erasure area in RE. In this ablation study, the source model is chosen to be Inc-v3 and the generated AE are then used to attack the other six models, and hence all the attacks are black-box.

The results as shown in Fig. 4, where we vary σ from 0.0 to 4.0 with step size 1.0, and vary s_h from 0.1 to 0.5 with step size 0.1. The perturbation $\epsilon = 16/255$ and the number of iterations $T = 10$. The results indicate that the best ASR is achieved at $\sigma = 2.0$, yet is *insensitive* to the choice of s_h (which is a good thing since it implies robustness of our erasure). Therefore, we have chosen $\sigma = 2.0$ and $s_h = 0.4$ in our experiments.

6. Conclusion

In this paper, we propose a new black-box approach of crafting transferable adversarial examples (AE) to attack deep learning based image classifiers. As such deep models are increasingly being deployed in autonomous driving, medical diagnosis, and many other computer vision applications, studying this topic plays an important role in deepening our understanding of AI security. Our proposed method consists of a gradient-based elastic momentum (EM) technique, and a random erasure (RE) data augmentation technique. EM introduces a flexible look-ahead horizon to estimate future momentum during AE computation, which speeds up the process of finding local optima and thus prevents hitting the overfitting region. RE creates an ensemble of transformed images that increases the diversity of perturbations and helps stablize gradient updates, which optimize the adversarial perturbations. We have performed extensive experiments to evaluate our proposed EM-RE method by attacking 7 modern deep learning classifiers and 9 advanced defense mechanisms, in comparison with 5 recently proposed baseline methods (and an additional Composite method). The results demonstrate superior transferability of the adversarial examples generated by our proposed method for black-box attacks.

References

1. Anish Athalye, Nicholas Carlini, and David Wagner. Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples. *International Conference on Machine Learning*, 2018.
2. Junyoung Byun, Seungju Cho, Myung-Joon Kwon, Hee-Seon Kim, and Changick Kim. Improving the transferability of targeted adversarial examples through object-based diverse input. In

Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 15244–15253, 2022.

3. Nicholas Carlini and David Wagner. Towards evaluating the robustness of neural networks. In *2017 ieee symposium on security and privacy (sp)*, pages 39–57. IEEE, 2017.
4. Pin-Yu Chen, Huan Zhang, Yash Sharma, Jinfeng Yi, and Cho-Jui Hsieh. Zoo: Zeroth order optimization based black-box attacks to deep neural networks without training substitute models. In *Proceedings of the 10th ACM workshop on artificial intelligence and security*, pages 15–26, 2017.
5. Minhao Cheng, Thong Le, Pin-Yu Chen, Jinfeng Yi, Huan Zhang, and Cho-Jui Hsieh. Query-efficient hard-label black-box attack: An optimization-based approach. *arXiv preprint arXiv:1807.04457*, 2018.
6. Jeremy M Cohen, Elan Rosenfeld, and J Zico Kolter. Certified adversarial robustness via randomized smoothing. *International Conference on Machine Learning*, 2019.
7. 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.
8. 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.
9. Gintare Karolina Dziugaite, Zoubin Ghahramani, and Daniel M Roy. A study of the effect of jpg compression on adversarial images. *arXiv preprint arXiv:1608.00853*, 2016.
10. Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *International Conference on Learning Representations*, 2015.
11. Chuan Guo, Mayank Rana, Moustapha Cisse, and Laurens Van Der Maaten. Countering adversarial images using input transformations. *International Conference on Learning Representations*, 2018.
12. Yiwen Guo, Qizhang Li, and Hao Chen. Backpropagating linearly improves transferability of adversarial examples. In *NeurIPS*, 2020.
13. 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.
14. Xiaojun Jia, Xingxing Wei, Xiaochun Cao, and Hassan Foroosh. Comdefend: An efficient image compression model to defend adversarial examples. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6084–6092, 2019.
15. Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.
16. Alexey Kurakin, Ian Goodfellow, and Samy Bengio. Adversarial examples in the physical world. *International Conference on Learning Representations, Workshop Track Proceedings*, 2017.
17. Alexey Kurakin, Ian Goodfellow, and Samy Bengio. Adversarial machine learning at scale. *International Conference on Learning Representations*, 2017.
18. Yingwei Li, Song Bai, Yuyin Zhou, Cihang Xie, Zhishuai Zhang, and Alan Yuille. Learning transferable adversarial examples via ghost networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 11458–11465, 2020.
19. Fangzhou Liao, Ming Liang, Yinpeng Dong, Tianyu Pang, Xiaolin Hu, and Jun Zhu. Defense against adversarial attacks using high-level representation guided denoiser. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1778–1787, 2018.
20. 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.
21. Xuanqing Liu, Minhao Cheng, Huan Zhang, and Cho-Jui Hsieh. Towards robust neural networks

via random self-ensemble. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 369–385, 2018.
22. Yanpei Liu, Xinyun Chen, Chang Liu, and Dawn Song. Delving into transferable adversarial examples and black-box attacks. *International Conference on Learning Representations*, 2017.
23. Zihao Liu, Qi Liu, Tao Liu, Nuo Xu, Xue Lin, Yanzhi Wang, and Wujie Wen. Feature distillation: Dnn-oriented jpeg compression against adversarial examples. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 860–868. IEEE, 2019.
24. Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. *International Conference on Learning Representations*, 2018.
25. 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.
26. Yurii Nesterov. A method for unconstrained convex minimization problem with the rate of convergence o(1/kˆ2). *Doklady AN USSR*, 269:543–547, 1983.
27. Tianyu Pang, Kun Xu, Chao Du, Ning Chen, and Jun Zhu. Improving adversarial robustness via promoting ensemble diversity. In *International Conference on Machine Learning*, pages 4970–4979. PMLR, 2019.
28. Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3):211–252, 2015.
29. Christian Szegedy, Sergey Ioffe, Vincent Vanhoucke, and Alex Alemi. Inception-v4, inception-resnet and the impact of residual connections on learning. *AAAI Conference on Artificial Intelligence*, 2017.
30. Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826, 2016.
31. Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. *International Conference on Learning Representations*, 2014.
32. Florian Tramèr, Alexey Kurakin, Nicolas Papernot, Ian Goodfellow, Dan Boneh, and Patrick McDaniel. Ensemble adversarial training: Attacks and defenses. *arXiv preprint arXiv:1705.07204*, 2017.
33. 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.
34. Xiaolong Wang, Abhinav Shrivastava, and Abhinav Gupta. A-fast-rcnn: Hard positive generation via adversary for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2606–2615, 2017.
35. Xiaosen Wang and Kun He. Enhancing the transferability of adversarial attacks through variance tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1924–1933, 2021.
36. Xiaosen Wang, Xuanran He, Jingdong Wang, and Kun He. Admix: Enhancing the transferability of adversarial attacks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 16158–16167, October 2021.
37. Xiaosen Wang, Jiadong Lin, Han Hu, Jingdong Wang, and Kun He. Boosting adversarial transferability through enhanced momentum. *arXiv preprint arXiv:2103.10609*, 2021.
38. Dongxian Wu, Yisen Wang, Shu-Tao Xia, James Bailey, and Xingjun Ma. Skip connections matter: On the transferability of adversarial examples generated with resnets. In *International Conference on Learning Representations*, 2019.

39. Cihang Xie, Jianyu Wang, Zhishuai Zhang, Zhou Ren, and Alan Yuille. Mitigating adversarial effects through randomization. *International Conference on Learning Representations*, 2018.
40. Cihang Xie, Zhishuai Zhang, Yuyin Zhou, Song Bai, Jianyu Wang, Zhou Ren, and Alan L Yuille. Improving transferability of adversarial examples with input diversity. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2730–2739, 2019.
41. Weilin Xu, David Evans, and Yanjun Qi. Feature squeezing: Detecting adversarial examples in deep neural networks. *Network and Distributed System Security Symposium*, 2018.
42. Yuchen Zhang and Percy Liang. Defending against whitebox adversarial attacks via randomized discretization. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 684–693. PMLR, 2019.
43. Zhun Zhong, Liang Zheng, Guoliang Kang, Shaozi Li, and Yi Yang. Random erasing data augmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 13001–13008, 2020.
44. Junhua Zou, Yexin Duan, Boyu Li, Wu Zhang, Yu Pan, and Zhisong Pan. Making adversarial examples more transferable and indistinguishable. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 3662–3670, 2022.
45. Junhua Zou, Zhisong Pan, Junyang Qiu, Xin Liu, Ting Rui, and Wei Li. Improving the transferability of adversarial examples with resized-diverse-inputs, diversity-ensemble and region fitting. In *European Conference on Computer Vision*, pages 563–579. Springer, 2020.