

Adversarial Attack via Dual-Stage Network Erosion

Yexin Duan^{1,2}, Junhua Zou², Xingyu Zhou², Wu Zhang², Jin Zhang¹ and Zhisong Pan^{2*}

¹Zhenjiang Campus, Army Military Transportation University, Zhenjiang, China

²Army Engineering University, Nanjing, China

Abstract

Deep neural networks are vulnerable to adversarial examples, which can fool deep models by adding subtle perturbations. Although existing attacks have achieved promising results, it still leaves a long way to go for generating transferable adversarial examples under the black-box setting. To this end, this paper proposes to improve the transferability of adversarial examples, and applies dual-stage feature-level perturbations to an existing model to implicitly create a set of diverse models. Then these models are fused by the longitudinal ensemble during the iterations. The proposed method is termed Dual-Stage Network Erosion (DSNE). We conduct comprehensive experiments both on non-residual and residual networks, and obtain more transferable adversarial examples with the computational cost similar to the state-of-the-art method. In particular, for the residual networks, the transferability of the adversarial examples can be significantly improved by biasing the residual block information to the skip connections. Our work provides new insights into the architectural vulnerability of neural networks and presents new challenges to the robustness of neural networks.

1 Introduction

Deep neural networks (DNNs) have shown compelling accuracy in the field of visual tasks. However, it has been found that DNNs are vulnerable to adversarial examples, which are input examples perturbed by imperceptible perturbations, which are carefully crafted but can fool the networks into making wrong predictions [Szegedy *et al.*, 2014; Goodfellow *et al.*, 2015].

The adversarial examples can be generated by white-box or black-box attacks. Since the internal information of the target model is usually not accessible, the black-box attacks remain a challenge. There are two main types of black-box methods, the query-based and the transfer-based. The query-based methods [Chen *et al.*, 2017; Brendel *et al.*, 2018] use

queries to obtain the information of the target model so as to estimate the decision boundary, which makes the black-box attacks almost white-box attacks. However, they require a large number of queries, which would be impractical in real-world applications. It has been found that the adversarial examples can transfer, that is, the examples generated for one model under the white-box setting can successfully attack other unknown models [Szegedy *et al.*, 2014; Liu *et al.*, 2017]. Hence, the transferability of adversarial examples can be leveraged to conduct black-box attacks.

Many techniques have been proposed to improve the transferability of adversarial examples, such as integrating the momentum term into the iterative process [Dong *et al.*, 2018], applying random transformations to the input [Xie *et al.*, 2019] and optimizing a perturbation over a set of translated images [Dong *et al.*, 2019]. The standard model ensemble method [Liu *et al.*, 2017; Dong *et al.*, 2018] average the outputs (*e.g.*, logits) of multiple models to improve the adversarial attacks, which prevents adversarial examples from overfitting to a specific model. These methods are either based on algorithm improvement, data augmentation or model input-output modification to improve the adversarial attacks, without considering the internal structural characteristics of the model.

Recently, methods have been proposed to consider the model internal structures and parameters, such as Ghost Networks (GN) [Li *et al.*, 2020], which explores network parameter perturbations to potentially create a set of diverse models, and fuses these models by longitudinal ensemble. As illustrated in Fig. 1, the standard ensemble requires averaging the outputs of different models. For the longitudinal ensemble, a set of diverse virtual models (*e.g.*, $\{M_{11}, M_{12}, \dots, M_{1N}\}$) can be obtained from a base model (*e.g.*, M_1) by randomizing the perturbation during iterations of adversarial attack. GN improves the adversarial attacks and generates adversarial examples efficiently. However, the results in black-box attacks still leave a lot of room for improvement.

Motivated by the above discussion, in this paper, we propose a Dual-Stage Network Erosion (DSNE) method, which makes the network parameters more diversified to further improve the adversarial attacks. By imposing dual-stage erosion (feature-level perturbations) on the internal structures and parameters of the base networks on-the-fly, the forward and back propagation of the information flow would be modified,

*Corresponding author

and multiple virtual models with similar decision boundaries are generated (“virtual” means that the generated models are not stored or trained). We call this operation “model augmentation”. Then these diversified virtual models are fused by the longitudinal ensemble during the iterations, which can alleviate the overfitting problem of iterative attacks, and the resultant adversarial examples are more likely to transfer across models.

Combining the proposed DSNE with any method (e.g., momentum iterative method [Dong *et al.*, 2018]), we obtain more transferable adversarial examples with computation complexity similar to the baseline method. And the longitudinal ensemble can be easily combined with standard ensemble to further improve the transferability of adversarial examples. In addition, for the non-residual networks, more diversified virtual models are generated through the dual-stage network erosion, which enhances the effectiveness of transfer attacks. In particular, for the residual networks, since the classification performance improvement mainly comes from the skip connections, we adjust the role of skip connections in attacks. We find that the attack success rates significantly improved if the networks bias towards the skip connections. This indicates that the skip connections can expose more transferable information, which is beneficial for the adversarial examples to cross the decision boundaries.

In summary, our main contributions are as follows:

- The proposed Dual-Stage Network Erosion (DSNE) method can generate more diverse virtual models and greatly improve the transferability of adversarial examples.
- We find that the transferability of the resultant adversarial examples can be significantly enhanced by making the output of residual blocks of the residual network biased towards the skip connections.
- We conduct extensive experiments both on normally trained models and robustly trained defense models, and the results demonstrate that our method can improve the black-box attacks with almost no extra computational cost.
- The proposed dual-stage erosion method has wide compatibility, which can be imposed on both non-residual and residual networks, and can also be combined with different attack methods.

2 Related work

Let x be a clean input that can be correctly classified by a classifier $c(\cdot)$ as label y . An adversarial example x^* can be obtained by adding imperceptible perturbations to x , which may fool the classifier, i.e., $c(x^*) \neq y$. For L_∞ norm constraint, the allowed perturbation should be smaller than a threshold ϵ as $\|x^* - x\|_\infty \leq \epsilon$. The attack objective is to maximize the cross-entropy loss function

$$J(x^*, y; \theta) = -\mathbb{1}_y \cdot \log(\text{softmax}(l(x^*))), \quad (1)$$

where θ denotes the network parameters, $-\mathbb{1}_y$ is the one-hot encoding of label y , and $l(x^*)$ is the classification logits of x^* ,

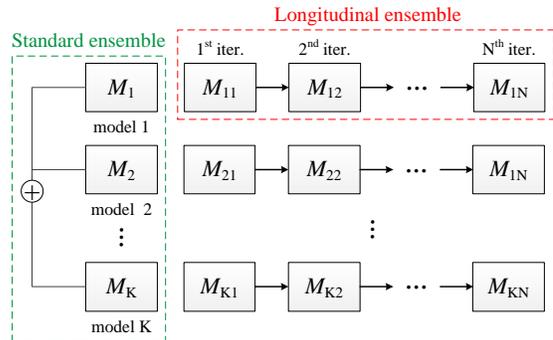


Figure 1: The illustration of standard ensemble and longitudinal ensemble.

thus the adversarial deep learning problem can be expressed as

$$\operatorname{argmax}_{x^*} J(x^*, y; \theta), \quad \text{s.t. } \|x^* - x\|_\infty \leq \epsilon. \quad (2)$$

Iterative Fast Gradient Sign Method(I-FGSM). I-FGSM [Kurakin *et al.*, 2017a] performs attack iteratively with a small step size. It initializes an adversarial example $x_0^* = x$ and the update equation is

$$x_{t+1}^* = \text{Clip}_x^\epsilon \{x_t^* + \alpha \text{sign}(\nabla_{x_t^*} J(x_t^*, y; \theta))\}, \quad (3)$$

where t is the t -th iteration and α is the step size. $\text{sign}(\cdot)$ is the sign function. $\text{Clip}_x^\epsilon \{x'\}$ function performs per-pixel clipping of the image x' , it can be expressed as $\min\{255, x + \epsilon, \max\{0, x - \epsilon, x'\}\}$, so the result will be constrained within ϵ -ball of the original image x .

Momentum Iterative Fast Gradient Sign Method (MI). MI [Dong *et al.*, 2018] integrates the momentum term into I-FGSM to stabilize gradient update direction and avoid trapping into the local maximum, it can be expressed as

$$g_{t+1} = \mu g_t + \frac{\nabla_{x_t^*} J(x_t^*, y; \theta)}{\|\nabla_{x_t^*} J(x_t^*, y; \theta)\|_1}, \quad (4)$$

$$x_{t+1}^* = \text{Clip}_x^\epsilon \{x_t^* + \alpha \text{sign}(g_{t+1})\}, \quad (5)$$

where g_t accumulates the iterated gradient vector of the loss function with a decay factor μ .

Translation-Invariant Method (TI). TI [Dong *et al.*, 2019] optimizes adversarial examples by convolving the gradient with a pre-defined kernel W , so that the generated adversarial examples would be less sensitive to the discriminative regions of the white-box model being attacked and have higher transferability. TI can be integrated into any gradient-based attack method, the integration of TI into the I-FGSM has the following update rule

$$x_{t+1}^* = \text{Clip}_x^\epsilon \{x_t^* + \alpha \text{sign}(W * \nabla_{x_t^*} J(x_t^*, y; \theta))\}. \quad (6)$$

Ghost Networks (GN). For non-residual networks, GN [Li *et al.*, 2020] generates virtual networks by inserting the dropout layer densely to every block throughout the base network. Let z_l be the activation in the l -th layer, f_l be the function that satisfies $z_{l+1} = f_l(z_l)$ for the l -th and $(l + 1)$ -th

layer, after applying dropout erosion, the output of f_l , *i.e.*, $g_l(z_l)$, is

$$g_l(z_l) = f_l\left(\frac{r_l * z_l}{1 - \Lambda_b}\right), \quad r_l \sim \text{Bernoulli}(1 - \Lambda_b), \quad (7)$$

where $*$ denotes an element-wise product and $\text{Bernoulli}(1 - \Lambda_b)$ means the Bernoulli distribution with the probability $p = (1 - \Lambda_b)$ of elements in r_l being 1, *i.e.*, p indicates the probability that z_l is preserved. Λ_b is defined as the magnitude of erosion, larger Λ_b implies a heavier erosion on the source network, and vice versa.

For the networks with residual blocks, GN applies randomized modulating scalar λ_l to the l -th residual block (see Fig. 3 (b)) by

$$z_{l+1} = \lambda_l z_l + f_l(z_l, w_l), \quad \lambda_l \sim U[1 - \Lambda_u, 1 + \Lambda_u], \quad (8)$$

where λ_l is subject to uniform distribution, z_l and z_{l+1} are the input and output of the l -th residual block with the weights w_l , $f(\cdot)$ denotes the residual function. To keep the expected input of z_l consisted after skip connection erosion, the mean of the uniform distribution is set to 1.

3 Methodology

GN explores network erosion to learn transferable adversarial examples, which can be applied both to single-model and multi-model attacks, and is compatible with various model structures and attack methods. However, there are several limitations: (1) GN generates a virtual model pool based on one-stage erosion to improve the transferability of adversarial examples, but the diversity of the network parameters is insufficient; (2) GN analyses the effect of erosion magnitude on classification accuracy, but does not analyse the effect on transferable attack performance, leading to inaccurate erosion magnitude and relatively low black-box attack success rates; (3) For ResNet-like networks, GN treats the skip connections (with an expected value of 1 for uniform distribution erosion) and residual modules equally. However, the main reason for the advanced performance of ResNet-like networks is the skip connections with the implementation of identity mapping, which can improve the information flow during forward and backward propagation, and enhance training efficiency and reduce test error [Srivastava *et al.*, 2015; Huang *et al.*, 2016; Veit *et al.*, 2016]. Therefore, for the parallel structure of the skip connection and the residual module in a residual block, the skip connection should be made to transfer more information, so as to improve the transferability of adversarial examples.

To address these issues, firstly, the proposed DSNE method obtains more diversified networks by imposing dual-stage erosion on the base network, which further alleviates the overfitting phenomenon of iterative attack; secondly, DSNE optimizes the erosion magnitude for different networks according to the attack effect; thirdly, DSNE makes the output of each residual block biased towards the skip connection to mitigate the reduction of transferability information flow.

In the following sections, we provide the detailed description of our DSNE method. The concept of model augmentation is proposed to introduce the principle of model diversification in Sec. 3.1, then we introduce the dual-stage network erosion for non-residual and residual networks in Sec. 3.2 and Sec. 3.3, respectively. The effect of erosion magnitude is analyzed in Sec. 4.2, and comprehensive experiments are conducted for single-model and multi-model attacks in Sec. 4.3 and Sec. 4.4, respectively.

3.1 Model augmentation

Leveraging the transferability to attack is to generate adversarial examples under the white-box setting, and then use these examples to attack the unknown models. Traditional iterative attacks may easily overfit the parameters of the attacked white-box model, and thus making the generated adversarial examples rarely transfer to other models.

Different from the common methods, such as algorithm improvement [Dong *et al.*, 2018; Dong *et al.*, 2019], data augmentation [Xie *et al.*, 2019] and standard model ensemble [Liu *et al.*, 2017; Dong *et al.*, 2018; Dong *et al.*, 2019], this paper alleviates the overfitting phenomenon by directly applying small parameter erosion $E(\cdot)$ to diversify the model, which satisfied $J(x, y; E(\theta)) \approx J(x, y; \theta)$ for any clean input x , by doing so, we derive a new model, and we call such derivation of models as **model augmentation**. Therefore, the constrained optimization problem in Eq. (2) can be rewritten as

$$\operatorname{argmax}_{x^*} J(x^*, y; E(\theta)), \quad \text{s.t. } \|x^* - x\|_\infty \leq \epsilon. \quad (9)$$

Due to the randomness of parameter erosion, each iteration will generate a new virtual model with similar decision boundaries, and then these multiple models generated at each iteration will be fused by the implicit longitudinal ensemble, making the resultant adversarial examples more transferable. The computation cost of the longitudinal ensemble attack is similar to that of base model iteration attack because network erosion requires little computation.

3.2 Non-residual network erosion

For non-residual networks, to make the network parameters more diversified, the proposed DSNE method combines dropout and uniform distribution erosion, and the output of the l -th layer can be rewritten as

$$g_l(z_l) = f_l\left(\frac{r_l * \lambda_l z_l}{1 - \Lambda_b}\right), \quad (10)$$

where λ_l is drawn from the uniform distribution $U[1 - \Lambda_u, 1 + \Lambda_u]$, and r_l is drawn from the Bernoulli distribution $\text{Bernoulli}(1 - \Lambda_b)$.

After applying the dual-stage erosion, the gradient of a loss function J with respect to input z_0 in back-propagation from the L -th layer can be expressed as

$$\frac{\partial J}{\partial z_0} = \frac{\partial J}{\partial z_L} \prod_{l=0}^L \left(\frac{r_l}{1 - \Lambda_b} * \lambda_l \frac{\partial}{\partial z_l} f_l \left(\frac{r_l * \lambda_l z_l}{1 - \Lambda_b} \right) \right). \quad (11)$$

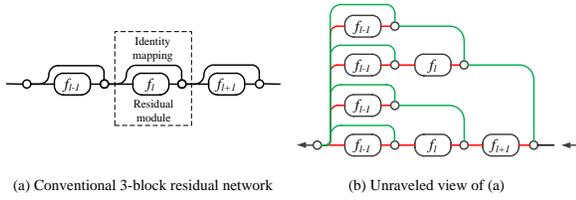


Figure 2: Conventional view (a) and unraveled view (b) of the residual network. Circular nodes denote junction point. The backpropagation paths of identity mapping and residual mapping are shown in green and red color, respectively.

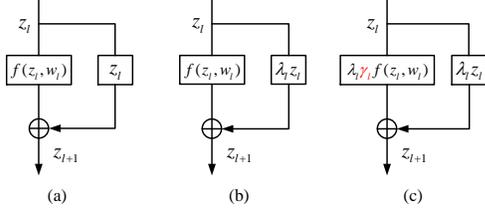


Figure 3: An illustration of (a) an original residual block, (b) the block after skip connection erosion and (c) the block after the dual-stage erosion.

3.3 Residual network erosion

Research [Srivastava *et al.*, 2015; Huang *et al.*, 2016; Veit *et al.*, 2016] demonstrates that the identity mapping helps to learn to proceed in very deep networks, and there is some redundancy in the paths of the residual network, which shows that the random discard of some residual layers has little impact on the testing results. These techniques are mainly used to improve the training efficiency and testing accuracy of the residual networks. However, our work is to study attacking networks to improve the transferability of the adversarial examples.

Residual networks [He *et al.*, 2016a; He *et al.*, 2016b] are neural networks in which each layer consists of two subterms: an identity skip connection mapping and a residual module mapping. With z_l as the input, the output of the $(l + 1)$ -th block is recursively defined as

$$z_{l+1} = z_l + f_l(z_l, w_l). \quad (12)$$

Consider a 3-block residual network, from input z_0 to z_3 , by expanding the recursion into the exponential number of nested items, we can make the structure of the residual network apparent, and obtain an unraveled view of the residual network [Veit *et al.*, 2016]. Omitting the weights for clarity, the output can be expanded as

$$\begin{aligned} z_3 &= z_2 + f_2(z_2) = [z_1 + f_1(z_1)] + f_2(z_1 + f_1(z_1)) \\ &= [z_0 + f_0(z_0) + f_1(z_0 + f_0(z_0))] + f_2(z_0 + f_0(z_0) + \\ &\quad f_1(z_0 + f_0(z_0))). \end{aligned} \quad (13)$$

As shown in Fig. 2, (a) is conventionally display form of the residual network, and (b) is the unraveled view as expressed in Eq. (13). The reduction of residual gradients is accumulated along the backpropagation paths (red paths), while

the identity mappings (green paths) facilitate the information propagation [He *et al.*, 2016b]. Therefore, a bias toward identity mappings may expose more transferable information.

The network parameters are first learned by training the source network from scratch, then we apply dual-stage erosion on the identity mapping and the residual module in the l -th residual block (see Fig. 3 (c)) by

$$z_{l+1} = \lambda_l(z_l + \gamma_l f(z_l, w_l)), \quad (14)$$

where λ_l is drawn from the uniform distribution $U[1 - \Lambda_u, 1 + \Lambda_u]$, γ_l is the bias factor and $0 < \gamma_l \leq 1$, such that the network is initially biased towards the shortcut connections which simply perform identity mapping. By doing so, it helps to improve the transferable information flow during forward and backward propagation, so as to enhance the attack effectiveness and obtain more transferable adversarial examples. It is worth noting that the model is not trained via Eq.(14).

The input of the L -th layer during inference can be written as

$$z_L = \left(\prod_{l=0}^{L-1} \lambda_l \right) z_0 + \sum_{l=0}^{L-1} \left(\prod_{l=0}^{L-1} \lambda_l \right) \gamma_l f(z_l, w_l). \quad (15)$$

The gradient of a loss function J with respect to input z_0 can be expressed as

$$\frac{\partial J}{\partial z_0} = \frac{\partial J}{\partial z_L} \left(\left(\prod_{l=0}^{L-1} \lambda_l \right) + \sum_{l=0}^{L-1} \left(\prod_{l=0}^{L-1} \lambda_l \right) \gamma_l \frac{\partial f(z_l, w_l)}{\partial z_0} \right). \quad (16)$$

The process of generating virtual models for non-residual or residual networks can be described in detail as follows: 1) conduct the uniform distribution erosion on the base network to obtain the perturbed network; 2) conduct the dropout or bias erosion on the perturbed network; 3) repeat step 1) and 2) to independently sample λ , r or γ for N times (N is the iteration number), and obtain a pool of virtual networks $M = \{M_1, M_2, \dots, M_N\}$, which are fused by the implicitly longitudinal ensemble for attacks, *i.e.*, at the i -th iteration, it attacks the virtual model M_i only.

Based on the above analysis, it can be inferred from the gradient of the loss function that a larger magnitude of erosion will have a greater influence on the source network, and deeper networks are influenced more easily according to the product rule. This is consistent with GN.

DSNE is compatible with various attack methods, *e.g.*, combined with MI and TI, we get the TI-MI-DSNE attack, with $x_0^* = x$, it can be written as

$$g_{t+1} = \mu g_t + \frac{W * \frac{\partial J}{\partial z_0}}{\left\| W * \frac{\partial J}{\partial z_0} \right\|_1}, \quad (17)$$

$$x_{t+1}^* = \text{Clip}_x^\epsilon \{x_t^* + \alpha \text{sign}(g_{t+1})\}, \quad (18)$$

where $z_0 = x_t^*$ is the input of the network at the t -th step, and $\frac{\partial J}{\partial z_0}$ for non-residual and residual networks are shown in Eq. (11) and (16), respectively. The TI-MI-DSNE combined with standard ensemble algorithm is summarized in Algorithm 1.

Algorithm 1 TI-MI-DSNE combined with standard ensemble

Input: A clean example x with label y ; K classifiers c_1, c_2, \dots, c_K ; ensemble weights w_1, w_2, \dots, w_K ;

Parameter: Perturbation size ϵ ; iteration number N and momentum decay factor μ ; pre-defined kernel W ; uniform distribution parameter Λ_u , dropout parameter Λ_b and scaling factor γ .

Output: An adversarial example x^* .

- 1: $\alpha = \epsilon/N$;
 - 2: $g_0 = 0$; $x_0^* = x$;
 - 3: **for** $t = 0$ to $N - 1$ **do**
 - 4: Input x_t^* and output the logits of K classifiers: $l_k(x_t^*), k = 1, 2, \dots, K$;
 - 5: Fuse the logits as $l(x_t^*) = \sum_{k=1}^K w_k(l_k(x_t^*))$;
 - 6: Get the cross-entropy loss \mathcal{J} based on $l(x_t^*)$ and Eq. (1);
 - 7: Let $z_0 = x_t^*$, for non-residual network and residual network, the gradient $\frac{\partial \mathcal{J}}{\partial z_0}$ is calculated by Eq. (11) and (16), respectively;
 - 8: Update the accumulated gradient g_{t+1} and adversarial example x_{t+1}^* by Eq. (17) and (18), respectively;
 - 9: **end for**
 - 10: **return:** $x^* = x_N^*$.
-

4 Experiments

In this section, we evaluate our method by comparing the transfer attack success rates on the ImageNet dataset [Rusakovsky *et al.*, 2015] through a large number of experiments. We make our codes public at <https://github.com/YeXinD/DSNE>.

4.1 Experimental settings

Source Models. We choose six models: Inception-v3 (Inc-v3) [Szegedy *et al.*, 2016], Inception-v4 (Inc-v4), Inception-ResNet-v2 (IncRes-v2) [Szegedy *et al.*, 2017], ResNet-v2-{50, 101, 152} (Res-{50, 101, 152}) [He *et al.*, 2016b] as the source models.

Target Models. To evaluate the transferability of the adversarial examples generated by the source models, we consider fifteen target models, nine of which are normally trained models: Inc-v3, Inc-v4, IncRes-v2, Res-{50, 101, 152}, Densenet-169 (Dense-169) [Huang *et al.*, 2017], Xception-71 (Xcep-71) [Chollet, 2017], and PNASnet-Large (PNAS) [Liu *et al.*, 2018]. The other six are robustly trained defense models, including three ensemble adversarially trained models: Inc-v3_{ens3}, Inc-v3_{ens4} and IncRes-v2_{ens} [Tramèr *et al.*, 2018], and the top-3 models in the NIPS 2017 Defense Competition: high-level representation guided denoiser (HGD) [Liao *et al.*, 2018], input transformation through random resizing and padding (R&P) [Xie *et al.*, 2018] and rank-3 solution¹ in the NIPS 2017 defense competition (NIPS-r3).

Datasets. It is less meaningful to study the attack success rates if the models cannot correctly classify the original images. Therefore, we randomly choose 5000 images from the

¹<https://github.com/anlhms/nips-2017/tree/master/mmd>

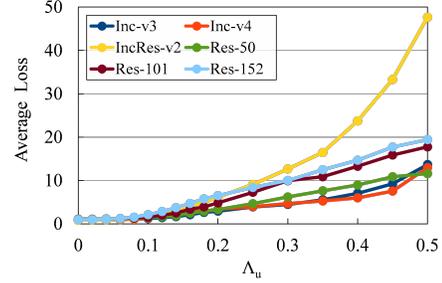


Figure 4: The average losses with different uniform distribution erosion magnitude of the six source models.

ImageNet validation set, and these images are correctly classified by all source models. All these images are resized to $299 \times 299 \times 3$ beforehand.

Baselines. We mainly compare our DSNE method with MI [Dong *et al.*, 2018], TI [Dong *et al.*, 2019] and the corresponding GN [Li *et al.*, 2020] methods. For all attack methods, the iteration number N is set to 10, other hyperparameters are set as in their original papers. We generate untargeted adversarial examples under maximum L_∞ perturbation $\epsilon = 16$ with respect to pixel values in $[0, 255]$.

4.2 Effect of erosion parameters

Due to the important influence of erosion parameters on the generation of strong transferable adversarial examples, a series of ablation experiments are conducted to study the effect of different erosion magnitude.

Uniform distribution parameter Λ_u . Uniform distribution parameter plays an important role in network diversity. We first verify the property of erosion parameter, *i.e.*, the effect of erosion on the classification performance of the model, with $\Lambda_u \in [0, 0.5]$, where $\Lambda_u = 0$ means no erosion on the source network. We input the clean images of the whole ILSVRC2012 validation set into the Inc-v3, Inc-v4, IncRes-v2, Res-50, Res-101 and Res-152, respectively. The average losses over all clean images for models with different erosion magnitude are shown in Fig. 4.

It can be seen that with the increase of the erosion magnitude, the loss increases smoothly. Therefore, $J(x, y; E(\theta)) \approx J(x, y; \theta)$ is satisfied when this erosion magnitude is within a small range, which is consistent with the proposed concept of model augmentation in Sec. 3.1. This rule also applies to the other two erosion parameters.

We then test the transferability with varying $\Lambda_u \in [0, 0.2]$. The larger the Λ_u , the greater the erosion of the source network. The attack results of DSNE combined with MI method against six target models (one white-box and five black-box models) are illustrated in Fig. 5 (a1), (b1), (c1) and Fig. 6 (a1), (b1), (c1). It can be observed that the trends of attack success rates of all black-box attacks against different target models are consistent. Increasing the erosion magnitude Λ_u tends to improve transferability until it exceeds a certain threshold.

For the Inception series networks, all three source models have the highest attack success rates when Λ_u is set to 0.10;

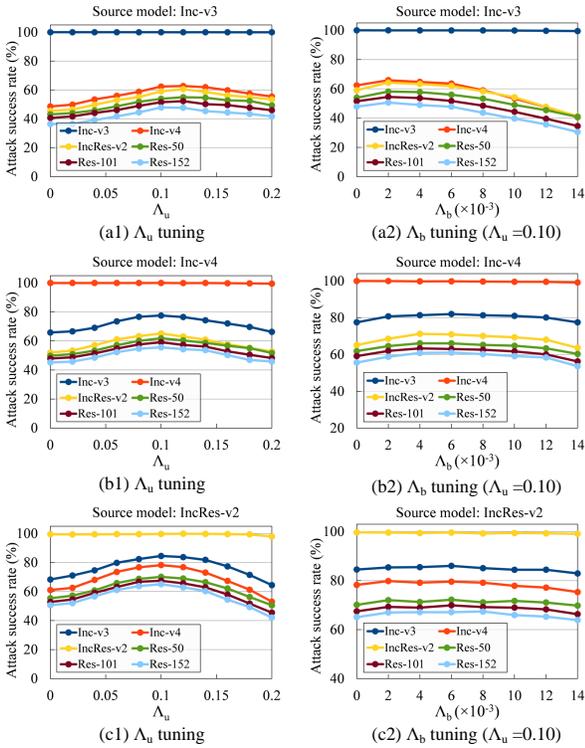


Figure 5: The attack success rates of Inception series networks with different erosion magnitude. The left column shows the Λ_u tuning, the right column shows the Λ_b tuning.

for the ResNet series networks, Λ_u is set to 0.14 for ResNet-50, 0.12 for ResNet-101, and 0.10 for ResNet-152. It can be seen that for deeper networks, the erosion magnitude should be smaller, which is consistent with the previous inference that deeper networks are influenced more easily.

When the enhancement of transferable information brought by the network diversity is greater than the gradient information loss caused by network erosion, the attack success rates will increase. If the erosion magnitude is too large, the gradient information of the virtual networks will be quite different from that of the source network, and the obtained virtual network will not satisfy $J(x, y; E(\theta)) \approx J(x, y; \theta)$, leading to the decrease of the attack success rates.

Dropout parameter Λ_b . For the Inception series networks, after tuning the uniform distribution parameter Λ_u , we test the transferability with varying dropout parameter $\Lambda_b \in [0, 0.014]$, where $\Lambda_b = 0$ means no dropout erosion on the network, and Λ_u is set to 0.10. As shown in Fig. 5 (a2), (b2), (c2), the attack success rates increase until Λ_b is greater than a certain value, 0.002 for Inc-v3, 0.004 for Inc-v4, 0.006 for IncRes-v2. The second stage erosion can make the virtual model more diverse, which further alleviates the overfitting problem and makes the resultant adversarial examples more transferable.

Bias factor γ . For residual networks, after tuning the erosion parameter Λ_u , we investigate the effect of initial bias of the residual block towards identity mapping on transfer attack. We set the range of the bias factor $\gamma \in [0.5, 1.0]$, where

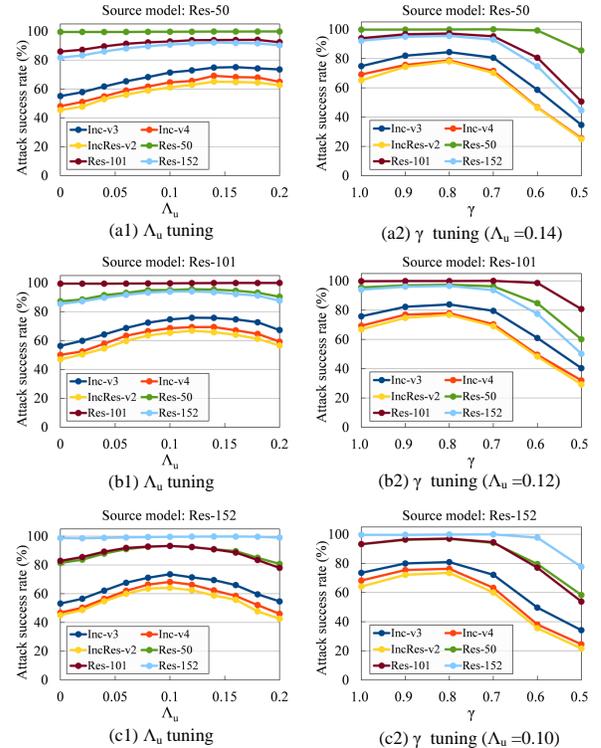


Figure 6: The attack success rates of ResNet series networks with different erosion magnitude. The left column shows the Λ_u tuning, the right column shows the γ tuning.

$\gamma = 1.0$ means no bias in the residual blocks.

Different layers of a neural network learn different levels of features, but the identity mapping can help preserve low-level features and avoid performance degradation when adding more layers, and allow unimpeded information flow across several layers [Srivastava *et al.*, 2015; He *et al.*, 2016b]. While the reduction of residual gradients is accumulated along the backpropagation path, that is, the residual gradients at lower layers will be reduced more times than those at higher layers, the bias towards the identity mapping would help to preserve the low-level features (see Fig. 2 (b) and Eq. (13)) and expose more gradient information, so that the information flow bias towards the identity mapping (by reducing γ) could boost the adversarial attack and improve the transferability of adversarial examples.

As shown in Fig. 6 (a2), (b2), (c2), the trends of the influence of bias factor on transfer attack are consistent. And these three residual networks share the same optimal γ , *e.g.*, $\gamma = 0.8$, which makes it easier to optimize the attack results. When the bias factor is too small, the class-relevant information will be excessively reduced, resulting in the failure of the model to obtain the correct class information and the useful gradient of the loss function, therefore, the attack success rates will decrease.

4.3 Single-model attacks

In this section, we perform adversarial attacks on a single network. We craft adversarial examples on each of the six source

Table 1: The attack success rates (%) against the normally trained models. * indicates the white-box attacks. The adversarial examples are generated on each of the six source models, respectively. The best results are in bold.

| Model | Attack | Inc-v3 | Inc-v4 | IncRes-v2 | Res-50 | Res-101 | Res-152 | Dense-169 | Xcep-71 | PNAS | Time(s) |
|-----------|---------|---------------|---------------|--------------|--------------|--------------|--------------|-------------|-------------|-------------|---------|
| Inc-v3 | MI | 100.0* | 48.6 | 45.5 | 43.1 | 40.7 | 36.5 | 46.2 | 43.0 | 33.3 | 997.0 |
| | MI-GN | 99.8* | 60.4 | 59.1 | 53.5 | 49.7 | 45.5 | 56.0 | 54.7 | 41.5 | 1053.2 |
| | MI-DSNE | 100.0* | 65.9 | 64.2 | 58.2 | 54.5 | 50.7 | 59.8 | 58.4 | 46.8 | 1013.1 |
| Inc-v4 | MI | 65.8 | 100.0* | 52.0 | 49.8 | 47.9 | 45.3 | 59.0 | 56.6 | 49.5 | 1424.6 |
| | MI-GN | 79.6 | 99.3* | 67.5 | 63.2 | 61.4 | 58.6 | 71.7 | 72.4 | 64.9 | 1665.2 |
| | MI-DSNE | 81.4 | 99.8* | 71.3 | 66.1 | 63.3 | 60.9 | 75.2 | 73.9 | 66.9 | 1710.3 |
| IncRes-v2 | MI | 68.3 | 61.1 | 99.5* | 55.4 | 52.8 | 50.6 | 58.8 | 53.2 | 48.5 | 1548.5 |
| | MI-GN | 79.6 | 71.9 | 99.7* | 64.8 | 61.8 | 59.3 | 68.1 | 62.3 | 57.4 | 1744.8 |
| | MI-DSNE | 86.0 | 79.6 | 99.6* | 72.2 | 69.9 | 67.1 | 74.7 | 71.4 | 65.8 | 1833.1 |
| Res-50 | MI | 55.1 | 48.1 | 45.4 | 99.5* | 85.9 | 81.7 | 60.2 | 48.8 | 41.9 | 922.5 |
| | MI-GN | 74.1 | 68.3 | 64.9 | 99.8* | 94.4 | 92.1 | 77.8 | 66.2 | 58.7 | 975.0 |
| | MI-DSNE | 84.4 | 78.8 | 78.1 | 99.9* | 97.1 | 95.5 | 86.1 | 77.5 | 70.8 | 975.1 |
| Res-101 | MI | 56.3 | 50.1 | 47.0 | 87.2 | 99.4* | 85.4 | 61.6 | 51.3 | 44.3 | 1241.1 |
| | MI-GN | 75.5 | 69.1 | 65.1 | 95.6 | 99.8* | 93.9 | 79.4 | 69.3 | 60.1 | 1319.5 |
| | MI-DSNE | 83.9 | 77.9 | 76.8 | 97.4 | 99.9* | 96.6 | 85.9 | 78.4 | 69.2 | 1385.7 |
| Res-152 | MI | 53.2 | 46.8 | 45.1 | 81.4 | 82.9 | 98.7* | 58.5 | 48.7 | 42.3 | 1615.0 |
| | MI-GN | 70.6 | 64.1 | 60.0 | 92.9 | 93.0 | 99.6* | 75.0 | 65.7 | 55.3 | 1688.1 |
| | MI-DSNE | 80.9 | 76.3 | 73.5 | 96.9 | 97.1 | 99.8* | 83.9 | 76.5 | 67.0 | 1812.4 |

Table 2: The black-box attack success rates (%) against the robustly trained defense models. The adversarial examples are generated on each of the six source models, respectively. The best results are in bold.

| Model | Attack | Inc-v3 _{ens3} | Inc-v3 _{ens4} | IncRes-v2 _{ens} | HGD | R&P | NIPS-r3 | Time(s) |
|-----------|------------|------------------------|------------------------|--------------------------|-------------|-------------|-------------|---------|
| Inc-v3 | TI-MI | 30.3 | 28.1 | 20.0 | 20.3 | 17.0 | 21.2 | 1090.2 |
| | TI-MI-GN | 41.5 | 39.0 | 26.6 | 28.8 | 24.4 | 24.9 | 1149.7 |
| | TI-MI-DSNE | 43.7 | 41.2 | 30.5 | 31.4 | 27.4 | 30.3 | 1128.7 |
| Inc-v4 | TI-MI | 32.3 | 31.3 | 23.5 | 24.2 | 21.6 | 24.9 | 1520.2 |
| | TI-MI-GN | 45.8 | 43.5 | 34.2 | 35.8 | 32.5 | 37.6 | 1759.6 |
| | TI-MI-DSNE | 46.2 | 44.7 | 34.6 | 35.4 | 32.5 | 37.2 | 1767.4 |
| IncRes-v2 | TI-MI | 44.0 | 41.2 | 40.2 | 37.2 | 36.2 | 39.2 | 1626.5 |
| | TI-MI-GN | 53.4 | 49.9 | 48.7 | 46.1 | 43.7 | 48.5 | 1779.2 |
| | TI-MI-DSNE | 57.6 | 54.7 | 52.9 | 49.7 | 47.7 | 52.7 | 1919.4 |
| Res-50 | TI-MI | 32.0 | 31.3 | 24.1 | 24.0 | 22.2 | 26.3 | 927.1 |
| | TI-MI-GN | 47.2 | 45.0 | 36.9 | 37.0 | 33.7 | 40.0 | 1036.9 |
| | TI-MI-DSNE | 55.8 | 54.6 | 43.9 | 42.5 | 40.0 | 47.6 | 1012.2 |
| Res-101 | TI-MI | 35.5 | 34.3 | 26.8 | 27.4 | 25.1 | 28.8 | 1269.9 |
| | TI-MI-GN | 48.0 | 46.2 | 37.4 | 38.1 | 35.1 | 41.1 | 1352.7 |
| | TI-MI-DSNE | 56.4 | 56.0 | 44.7 | 41.9 | 40.3 | 47.5 | 1495.0 |
| Res-152 | TI-MI | 34.7 | 33.8 | 27.5 | 27.1 | 25.4 | 29.5 | 1721.0 |
| | TI-MI-GN | 46.4 | 44.5 | 36.0 | 36.1 | 33.8 | 39.4 | 1784.3 |
| | TI-MI-DSNE | 55.5 | 55.4 | 45.4 | 42.9 | 41.7 | 48.5 | 1879.8 |

models and test them on all fifteen target models.

According to the discussion above, we select the optimized erosion parameters for each source model and combine our DSNE method with MI [Dong *et al.*, 2018] method to attack against the nine normally trained models, the comparison of the results are shown in Table 1. Since TI [Dong *et al.*, 2019] method is more effective for the defense models, we combine it to attack six robustly trained defense models, and the results are shown in Table 2.

It can be seen that the black-box attack success rates of the proposed DSNE method are significantly higher than that of the baselines. Especially when the source model is the residual network, the average black-box attack success rates

of our DSNE method is about 7% ~ 10% higher than that of the Ghost Networks (GN) [Li *et al.*, 2020] method.

Note that the generated virtual networks are fused by the longitudinal ensemble, and these virtual models are not stored or trained, thus our attacks require similar time and space complexity to the baselines.

In the last column of each table, we also list the running time as the computational cost of each attack method, each attack is run on an NVIDIA GTX 1080Ti GPU. It can be seen that our proposed DSNE method has similar computational costs to the baseline methods.

We visualize two randomly selected clean images and their corresponding adversarial examples in Fig. 7. All these ad-

Table 3: The attack success rates (%) against the normally trained models, and adversarial examples are generated on an ensemble of three source models. * indicates the white-box attacks. The best results are in bold.

| Model | Attack | Inc-v3 | Inc-v4 | IncRes-v2 | Res-50 | Res-101 | Res-152 | Dense-169 | Xcep-71 | PNAS | Time(s) |
|-----------|------------|---------------|--------------|--------------|--------------|--------------|--------------|-------------|-------------|-------------|---------|
| Inc-v3 | MI | 100.0* | 99.7* | 98.4* | 74.8 | 73.8 | 72.1 | 79.4 | 77.9 | 76.6 | 3087.2 |
| | MI-GN | 99.8* | 98.5* | 99.3* | 86.3 | 85.5 | 83.7 | 89.1 | 89.5 | 86.9 | 3222.1 |
| Inc-v4 | MI-DSNE | 99.9* | 99.3* | 98.9* | 87.1 | 85.9 | 84.7 | 90.2 | 90.1 | 87.1 | 3300.9 |
| | TI-MI | 99.6* | 97.7* | 93.1* | 62.7 | 62.1 | 61.0 | 69.4 | 66.5 | 67.8 | 3090.8 |
| IncRes-v2 | TI-MI-GN | 98.0* | 94.8* | 94.8* | 71.4 | 70.3 | 68.5 | 78.8 | 75.8 | 74.4 | 3212.1 |
| | TI-MI-DSNE | 98.9* | 96.7* | 92.2* | 72.1 | 70.4 | 68.6 | 79.4 | 76.5 | 74.9 | 3340.1 |
| Res-50 | MI | 79.6 | 74.8 | 73.9 | 99.4* | 99.4* | 99.4* | 81.7 | 72.9 | 72.7 | 2707.0 |
| | MI-GN | 91.6 | 89.3 | 87.3 | 99.7* | 99.7* | 99.7* | 93.9 | 88.5 | 86.5 | 2836.3 |
| Res-101 | MI-DSNE | 96.7 | 94.8 | 94.8 | 99.9* | 99.9* | 99.9* | 97.3 | 94.5 | 92.1 | 3189.3 |
| | TI-MI | 64.2 | 58.5 | 57.3 | 99.0* | 99.0* | 98.8* | 63.8 | 56.0 | 59.9 | 2859.8 |
| Res-152 | TI-MI-GN | 74.3 | 68.8 | 65.8 | 98.2* | 98.4* | 98.3* | 74.3 | 66.5 | 68.6 | 2972.2 |
| | TI-MI-DSNE | 78.4 | 73.4 | 70.8 | 98.5* | 98.2* | 98.2* | 77.1 | 70.0 | 71.6 | 3281.4 |

Table 4: The black-box attack success rates (%) against the robustly trained defense models, and adversarial examples are generated on an ensemble of three source models. The best results are in bold.

| Model | Attack | Inc-v3 _{ens3} | Inc-v3 _{ens4} | IncRes-v2 _{ens} | HGD | R&P | NIPS-r3 |
|-----------|------------|------------------------|------------------------|--------------------------|-------------|-------------|-------------|
| Inc-v3 | MI | 35.3 | 30.4 | 18.6 | 22.7 | 18.5 | 28.9 |
| | MI-GN | 44.5 | 39.1 | 23.0 | 23.6 | 23.8 | 37.6 |
| Inc-v4 | MI-DSNE | 44.9 | 38.7 | 23.0 | 22.3 | 23.5 | 36.9 |
| | TI-MI | 61.3 | 59.1 | 53.3 | 56.5 | 50.2 | 54.6 |
| IncRes-v2 | TI-MI-GN | 70.0 | 67.8 | 62.1 | 64.8 | 60.1 | 64.3 |
| | TI-MI-DSNE | 70.4 | 68.5 | 62.6 | 64.9 | 61.0 | 64.7 |
| Res-50 | MI | 39.6 | 33.9 | 21.8 | 33.7 | 22.3 | 32.2 |
| | MI-GN | 51.6 | 44.2 | 28.4 | 42.3 | 28.9 | 41.6 |
| Res-101 | MI-DSNE | 67.9 | 61.2 | 43.2 | 49.2 | 44.6 | 59.2 |
| | TI-MI | 57.4 | 55.0 | 47.6 | 51.3 | 46.1 | 51.1 |
| Res-152 | TI-MI-GN | 67.8 | 65.7 | 58.7 | 60.6 | 56.6 | 62.0 |
| | TI-MI-DSNE | 76.0 | 76.3 | 66.7 | 65.9 | 63.7 | 70.4 |

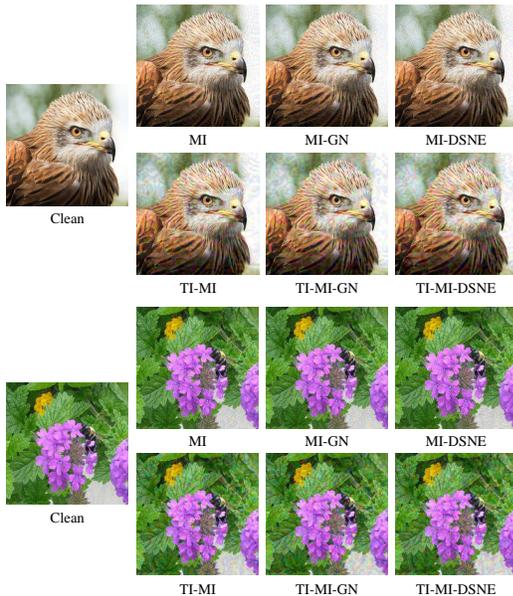


Figure 7: The adversarial examples generated by the proposed DSNE method and other baselines on the Inc-v3 model.

versarial examples are generated on Inc-v3 using different methods with the maximum perturbation $\epsilon = 16$. Although the proposed DSNE method has significantly improved the black-box attack success rates, we can see that the magnitude of adversarial perturbations is almost the same as that of the baselines.

4.4 Multi-model attacks

Research [Liu *et al.*, 2017] demonstrated that attacking different models simultaneously can significantly improve the transferability of adversarial examples, which can also evaluate the robustness of the target models more accurately. We combine the standard ensemble and longitudinal ensemble, *i.e.*, the multi-model attack treats each longitudinal ensemble as a branch of the standard ensemble (seen in Fig. 1).

We attack the Inception series and ResNet series model ensembles, respectively. The success rates against nine normally trained models and six robustly trained models are summarized in Table 3 and Table 4, respectively. Note that the TI method is originally used to attack robustly trained defense models, although here we use it to attack both normally trained and robustly trained models. It can be seen that similar to single-model attacks, our DSNE method can improve the transferability of the resultant adversarial examples significantly.

As shown in Table 3, for the Inception series ensemble, the black-box attack performance of our DSNE method combined with MI is better than other methods. For the ResNet series ensemble, our DSNE method combined with MI consistently outperforms all other methods under both white-box and black-box settings. Compared with the strong baseline, *e.g.*, MI-GN, our MI-DSNE method improves the average black-box attack success rates by a large margin (about 6%). Even only three source models are used, MI-DSNE achieves a high average black-box attack success rate (95.0%), which verifies that the bias towards identity mapping makes the adversarial examples transfer more easily.

In Table 4, for the Inception series ensemble, the DSNE method also shows superior attack performance. In addition, for the ResNet series ensemble, similar to the results of against normally trained models, DSNE combined with TI and MI consistently improves the transferability of the adversarial examples by a large margin, *e.g.* the average attack success rate is about 8% higher than the TI-MI-GN. The results indicate that the structures of the deep networks are still vulnerable and the security of the networks can be enhanced from the structure design.

5 Conclusion

This paper studies enhancing the transferability of adversarial examples by eroding the internal parameters of the source network on-the-fly. First, we adopt the proposed dual-stage network erosion to augment the source models and make the models more diversified, which alleviates the overfitting problem of iterative attacks and makes the generated adversarial examples more transferable. Second, we fuse the generated virtual models by the longitudinal ensemble, which significantly enhances the black-box attack success rates with similar computational consumption. Particularly, for the residual network, we find that when the network is biased towards identity mapping, the transferability of the resultant adversarial examples will be improved significantly, the average attack success rates are about 6% ~ 10% higher than that of the state-of-the-art method under the single-model and multi-model settings. Our work poses new challenges for the application of deep neural networks.

References

- [Brendel *et al.*, 2018] Wieland Brendel, Jonas Rauber, and Matthias Bethge. Decision-based adversarial attacks: Reliable attacks against black-box machine learning models. In *6th International Conference on Learning Representations (ICLR)*, 2018.
- [Chen *et al.*, 2017] 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.
- [Chollet, 2017] François Chollet. Xception: Deep learning with depthwise separable convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*, pages 1251–1258, 2017.
- [Dong *et al.*, 2018] 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 (CVPR)*, pages 9185–9193, 2018.
- [Dong *et al.*, 2019] Yinpeng Dong, Tianyu Pang, Hang Su, and Jun Zhu. Evading defenses to transferable adversarial examples by translation-invariant attacks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4312–4321, 2019.
- [Goodfellow *et al.*, 2015] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In *3rd International Conference on Learning Representations (ICLR)*, 2015.
- [He *et al.*, 2016a] 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 (CVPR)*, pages 770–778, 2016.
- [He *et al.*, 2016b] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Identity mappings in deep residual networks. In *European conference on computer vision (ECCV)*, pages 630–645. Springer, 2016.
- [Huang *et al.*, 2016] Gao Huang, Yu Sun, Zhuang Liu, Daniel Sedra, and Kilian Q Weinberger. Deep networks with stochastic depth. In *European conference on computer vision (ECCV)*, pages 646–661. Springer, 2016.
- [Huang *et al.*, 2017] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*, pages 4700–4708, 2017.
- [Kurakin *et al.*, 2017a] Alexey Kurakin, Ian Goodfellow, and Samy Bengio. Adversarial examples in the physical world. In *5th International Conference on Learning Representations (ICLR)*, 2017a.
- [Li *et al.*, 2020] Yingwei Li, Song Bai, Yuyin Zhou, Cihang Xie, Zhishuai Zhang, and Alan L Yuille. Learning transferable adversarial examples via ghost networks. In *34th AAAI Conference on Artificial Intelligence (AAAI)*, pages 11458–11465, 2020.
- [Liao *et al.*, 2018] 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 (CVPR)*, pages 1778–1787, 2018.
- [Liu *et al.*, 2017] Yanpei Liu, Xinyun Chen, Chang Liu, and Dawn Song. Delving into transferable adversarial examples and black-box attacks. In *5th International Conference on Learning Representations (ICLR)*, 2017.
- [Liu *et al.*, 2018] Chenxi Liu, Barret Zoph, Maxim Neumann, Jonathon Shlens, Wei Hua, Li-Jia Li, Li Fei-Fei, Alan Yuille, Jonathan Huang, and Kevin Murphy. Progressive neural architecture search. In *Proceedings of the*

European Conference on Computer Vision (ECCV), pages 19–34, 2018.

- [Russakovsky *et al.*, 2015] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael S Bernstein, et al. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 115(3):211–252, 2015.
- [Srivastava *et al.*, 2015] Rupesh Kumar Srivastava, Klaus Greff, and Jürgen Schmidhuber. Highway networks. *arXiv preprint arXiv:1505.00387*, 2015.
- [Szegedy *et al.*, 2014] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. In *2nd International Conference on Learning Representations (ICLR)*, 2014.
- [Szegedy *et al.*, 2016] 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 (CVPR)*, pages 2818–2826, 2016.
- [Szegedy *et al.*, 2017] Christian Szegedy, Sergey Ioffe, Vincent Vanhoucke, and Alexander A Alemi. Inception-v4, inception-resnet and the impact of residual connections on learning. In *31st AAAI Conference on Artificial Intelligence (AAAI)*, 2017.
- [Tramèr *et al.*, 2018] Florian Tramèr, Alexey Kurakin, Nicolas Papernot, Ian Goodfellow, Dan Boneh, and Patrick McDaniel. Ensemble adversarial training: Attacks and defenses. In *6th International Conference on Learning Representations (ICLR)*, 2018.
- [Veit *et al.*, 2016] Andreas Veit, Michael J Wilber, and Serge Belongie. Residual networks behave like ensembles of relatively shallow networks. In *Advances in neural information processing systems (NIPS)*, pages 550–558, 2016.
- [Xie *et al.*, 2018] Cihang Xie, Jianyu Wang, Zhishuai Zhang, Zhou Ren, and Alan Yuille. Mitigating adversarial effects through randomization. In *6th International Conference on Learning Representations (ICLR)*, 2018.
- [Xie *et al.*, 2019] Cihang Xie, Zhishuai Zhang, Yuyin Zhou, Song Bai, Jianyu Wang, Zhou Ren, and Alan L Yuille. Improving transferability of adversarial examples with input diversity. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2730–2739, 2019.