Enhancing the Adversarial Robustness via Manifold Projection

Zhiting Li, Shibai Yin, Tai-Xiang Jiang*, Yexun Hu, Jia-Mian Wu, Guowei Yang, Guisong Liu

School of Computing and Artificial Intelligence, Southwestern University of Finance and Economics, Chengdu, P.R.China
Kash Institute of Electronics and Information Industry, Kash, P.R.China
Engineering Research Center of Intelligent Finance, Ministry of Education, Chengdu, P.R.China

Abstract

Deep learning has been widely applied to various aspects of computer vision, but the emergence of adversarial attacks raises concerns about its reliability. Adversarial training (AT) is one of the most effective defense methods, which incorporates adversarial examples into the training data. However, AT is typically employed in a discriminative learning manner, i.e., learning the mapping (conditional probability) from samples to labels, it essentially reinforces this mapping without considering the underlying data distribution. It is notable that adversarial examples often deviate from the distribution of normal (clean) samples. Therefore, building upon existing adversarial defense schemes, we propose to further exploit the distribution of normal samples, partly from the generative learning perspective, resulting in a novel robustness enhancement paradigm. We train a simple autoencoder (AE) autoregressively on normal samples to learn their prior distribution, effectively serving as an image manifold. This AE is then used as a manifold projection operator to incorporate the distribution information of normal samples. Specifically, we organically integrate the pretrained AE into the training process of both AT and adversarial distillation (AD), a method aiming at improving the robustness of small models with low capacity. Since the AE captures the distribution of normal samples, it can adaptively pull adversarial examples closer to the normal sample manifold, weakening the attack strength of adversarial samples and easing the learning of mappings from adversarial samples to correct labels. From the Pearson correlation coefficient (PCC) between the statistics on normal and adversarial examples, it's validated that the AE indeed pulls adversarial samples closer to normal samples. Extensive experiments illustrate that our proposed adversarial defense paradigm significantly improves the robustness compared with previous state-of-the-art AT and AD methods.

1 Introduction

Currently, deep neural networks (DNNs) have achieved tremendous success in the field of computer vision (He et al. 2016; Simonyan and Zisserman 2015; Tan and Le 2019; Liu et al. 2021). However, adversarial attacks against DNNs pose significant security threats (Goodfellow, Shlens, and Szegedy 2015; Moosavi-Dezfooli, Fawzi, and Frossard 2016; Zhao, Dua, and Singh 2018; Xiao et al. 2018; Poursaeed et al. 2018; Dia, Barshan, and Babanezhad 2019; Liu et al. 2022), particularly in fields such as autonomous driving and facial recognition (Kong et al. 2020; Sun et al. 2022; Zheng et al. 2024; Li et al. 2023). These attacks exploit the vulnerabilities of DNNs by introducing perturbations that are imperceptible to humans but can drastically alter the model's predictions.

Adversarial training (AT) has emerged as one of the most effective defense methods against such attacks (Goodfellow, Shlens, and Szegedy 2015; Madry et al. 2018; Wong, Rice, and Kolter 2020; Andriushchenko and Flammarion 2020). AT incorporates adversarial examples into the training data to enhance the model's robustness. However, AT is typically employed in a discriminative learning manner, focusing on learning the mapping (conditional probability) from samples to labels (Ng and Jordan 2001; Krizhevsky, Sutskever, and Hinton 2012; He et al. 2016; Sandler et al. 2018). This approach primarily aims to reinforce the mapping from input samples to their corresponding labels, without considering the underlying distribution of normal samples. Consequently, AT often requires high-capacity models, making it less suitable for edge computing scenarios where computational resources are limited (Madry et al. 2018; Xie and Yuille 2020; Zhang et al. 2021).

To address the limitations of AT, adversarial distillation (AD) has been proposed, where robustness is distilled from an adversarially pretrained model (teacher model) to a student model in a teacher-student architecture (Goldblum et al. 2020; Zi et al. 2021; Huang et al. 2023). AD aims to improve the robustness of small, low-capacity models. However, existing AD methods are limited to aligning the prediction outputs of teacher and student models, neglecting the underlying data distribution. If there is a significant capacity gap between the teacher and student models or the teacher model's performance is suboptimal, the effectiveness of adversarial distillation may be compromised (Cho and Hariharan 2019; Huang et al. 2022). Both AT and AD reinforce the sample-to-label mapping without considering the underlying data distribution. Moreover, adversarial examples often deviate from the distribution of normal samples, leading to suboptimal robustness (Pang et al. 2022; Yu, Chen, and Gan 2023).

In this paper, we propose a novel adversarial defense

*Corresponding author. (Email: taixiangjiang@gmail.com)

paradigm by integrating manifold projection via an autoencoder (AE) into existing AT and AD methods. To achieve this, we train the AE autoregressively on normal samples to learn their prior distribution. The pretrained AE then serves as a manifold projection operator, incorporating the distribution information of normal samples into the training process of AT and AD methods. By capturing the prior distribution of normal samples, our approach integrates aspects of generative learning into the adversarial defense framework. This enables the AE to adaptively pull adversarial examples closer to the manifold of clean examples, weakening their attack strength and easing the learning of mappings from adversarial samples to correct labels. This approach allows us to train more robust models for AT, which can then be used as teacher models. For AD, the manifold projection enables better knowledge distillation from larger teacher models to smaller student models, overcoming the capacity gap and enhancing the adversarial robustness of student models.

Our main contributions can be summarized as:

• We propose a novel adversarial defense paradigm that exploits the distribution of normal samples. We train an autoencoder on normal samples, to learn this distribution. It can be found that integrating the manifold projection via this AE into adversarial defense methods makes adversarial examples approach the manifold of clean examples.

• We use this pretrained AE as a manifold projection operator and organically integrate it into the training process of both adversarial training (AT) and adversarial distillation (AD) methods. For AT with the manifold projection, the robustness of the model is significantly enhanced. For AD with the manifold projection, the student models can more effectively inherit the adversarial robustness from teacher models. Through this, we propose novel AT and AD methods.

• Extensive experiments illustrate that our proposed adversarial defense paradigm significantly improves the robustness compared with previous state-of-the-art AT and AD methods. Through extensive ablation experiments, we validate the effectiveness of the AE and also discover its benefits in terms of robust fairness and resistance to transfer-based attacks.

2 Related Work

Research has illustrated that DNNs are vulnerable to adversarial attacks (Goodfellow, Shlens, and Szegedy 2015; Ilyas et al. 2018; Mahmood, Mahmood, and Van Dijk 2021). Various adversarial defense methods have been developed to mitigate these vulnerabilities, with adversarial training (AT) and adversarial distillation (AD) being two of the most prominent approaches. This section provides an overview of adversarial attacks, AT, and AD methods.

2.1 Adversarial Attacks

Adversarial attacks can be broadly categorized into two main types: white-box and black-box attacks. In white-box attacks, attackers exploit gradient information from the target models to craft adversarial examples. Prominent methods include the Fast Gradient Sign Method (FGSM) (Goodfellow, Shlens, and Szegedy 2015), Projected Gradient Descent (PGD) (Madry et al. 2018), Carlini-Wagner attacks (CW) (Carlini and Wagner 2017), and AutoAttack (AA) (Croce and Hein 2020). In contrast, black-box attacks either transfer adversarial examples from surrogate models (Papernot, McDaniel, and Goodfellow 2016; Tramèr et al. 2017; Wang et al. 2023) or rely on querying the target model to identify vulnerabilities (Ilyas et al. 2018; Brendel, Rauber, and Bethge 2018; Guo et al. 2019).

2.2 Adversarial Training

Adversarial training (AT) is one of the most effective defenses against adversarial attacks. It strengthens model robustness by incorporating adversarial examples during training. Initially introduced with FGSM (Goodfellow, Shlens, and Szegedy 2015), AT was later enhanced by PGD-AT (Madry et al. 2018), which uses iterative PGD-generated adversarial examples. TRADES (Zhang et al. 2019) further refined AT by balancing robustness and clean accuracy.

For a typical image classification task, assuming the input data point $(\mathbf{x}_i, y_i)$ follows the joint data distribution $p_d(\mathbf{x}_i, y_i)$, the PGD-AT objective can be formulated as:

$$\mathbb{E}_{p_d(\mathbf{x})}\, \ell(F(\mathbf{x}^*), p_d(y|\mathbf{x})), \tag{1}$$

where $p_d(y|\mathbf{x})$ represents the ground-true labels commonly provided by the dataset, $F(\cdot)$ denoted the model parameterized by θ_F, ℓ refers to the Cross-Entropy (CE) loss, a standard choice in supervised learning, and $\mathbf{x}^*$, the result obtained from the inner optimization, is obtained as

$$\mathbf{x}^* = \mathbf{x} + \arg\max_{\|\delta\|_p \le \epsilon} \ell(F(\mathbf{x}+\delta), p_d(y|\mathbf{x})), \tag{2}$$

with δ being the perturbation contained by the L_p-norm ϵ.

2.3 Adversarial Distillation

Adversarial Distillation (AD) extends knowledge distillation (KD) (Hinton, Vinyals, and Dean 2015) to enhance the robustness of smaller models by transferring knowledge from larger and more robust teacher models. Unlike KD, AD prioritizes both clean accuracy and adversarial robustness. Adversarial Robust Distillation (ARD) (Goldblum et al. 2020) combines AT with KD as:

$$\mathbb{E}_{p_d(\mathbf{x})}\left[(1-\alpha)\ell(S(\mathbf{x}), p_d(y|\mathbf{x})) + \alpha\tau^2 \mathrm{KL}(S^\tau(\mathbf{x}^*), T^\tau(\mathbf{x}))\right],$$

where, $T(\cdot)$ and $S(\cdot)$ represent teacher and student models, parameterized by θ_T and θ_S, respectively, $\mathrm{KL}(\cdot\,,\cdot)$ denotes the Kullback-Leibler divergence, τ is the softmax temperature, and $\mathbf{x}^*$ comes from Eq. (2) with $F(\cdot)$ replaced by $S(\cdot)$.

RSLAD (Zi et al. 2021) improves ARD by generating adversarial examples using robust soft labels from the teacher model. Introspective Adversarial Distillation (IAD) (Zhu et al. 2022) addresses the issue of unreliable teacher models at specific data points. Adaptive Adversarial Distillation (AdaAD) (Huang et al. 2023) further enhances AD by dynamically aligning student and teacher predictions, with the inner optimization formulated as:

$$\mathbf{x}^* = \mathbf{x} + \arg\max_{\|\delta\|_p \le \epsilon} \mathrm{KL}(S(\mathbf{x}+\delta), T(\mathbf{x}+\delta)), \tag{3}$$

and the overall objective is

$$\mathbb{E}_{p_d(\mathbf{x})}\left[(1-\alpha)\mathrm{KL}(S(\mathbf{x}), T(\mathbf{x})) + \alpha\mathrm{KL}(S(\mathbf{x}^*), T(\mathbf{x}^*))\right].$$

AdaAD is currently among the most effective AD methods for adversarial defense.

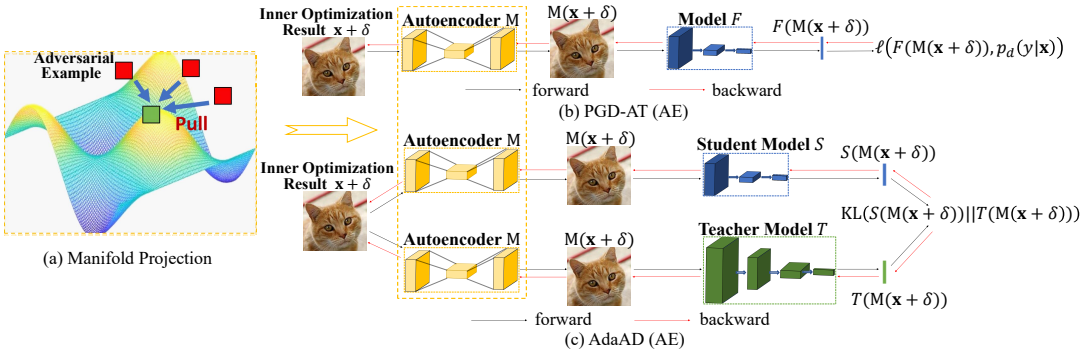

Figure 1: (a) A symbolic illustration showing how the manifold projection, implemented through an autoencoder, can "correct" adversarial examples (red squares) by pulling them toward clean examples (green squares) within a clean sample manifold. (b) The inner optimization process for PGD-AT (AE), where the autoencoder projects adversarial examples closer to the manifold of clean examples within training. (c) The inner optimization process for AdaAD (AE), illustrating how both student and teacher models interact with the autoencoder to improve robustness during adversarial distillation.

3 Methodology

Reforming adversarial examples via autoencoders (AEs) for robustness can be traced back to MagNet (Meng and Chen 2017), where the AE merely acts as an image pre-processor. In this work, we propose to organically integrate manifold projection via AEs into existing Adversarial Training (AT) and Adversarial Distillation (AD) methods, resulting in a novel robustness enhancement paradigm. We begin by introducing the concept of manifold projection via an autoencoder and then illustrate how it can be effectively incorporated into AT and AD methods.

3.1 Manifold Projection via Autoencoder

We train the autoencoder $\mathtt{M} : \mathbb{X}_{\text{train}} \rightarrow \mathbb{X}_{\text{train}}$ to minimize reconstruction error autoregressively on the clean training dataset $\mathbb{X}_{\text{train}}$. The loss function is defined using the simple mean squared error as:

$$L(\mathbb{X}_{\text{train}}) = 1/|\mathbb{X}_{\text{train}}| \sum_{\mathbf{x}\in\mathbb{X}_{\text{train}}} \|\mathbf{x} - \mathtt{M}(\mathbf{x})\|_2. \quad (4)$$

The AE is expected to weaken the attack strength of adversarial samples by pulling them close to the normal samples. Then, we integrate the pretrained AE into the training process of both AT and AD, as shown in Fig. 1-(b) and (c).

To delve deeper into the impact of integrating manifold projection on model robustness, we conducted a statistic analysis using 1,000 mini-batches from the CIFAR-100 training set. The inner optimization results are obtained during the training process and the adversarial examples are generated under a PGD-10 attack for both PGD-AT and AdaAD, with and without manifold projection. We then calculate the mean statistics from the final batch normalization layer of ResNet18 for each batch. To quantify distributional differences from clean examples, we compute the Pearson correlation coefficient (PCC) between the statistics of clean examples and those of both inner optimization results and adversarial examples. Fig. 2 presents the frequency histograms of distributional differences across 1,000 mini-batches for PGD-AT with and without manifold projection. The frequency histograms for the AD method AdaAD with

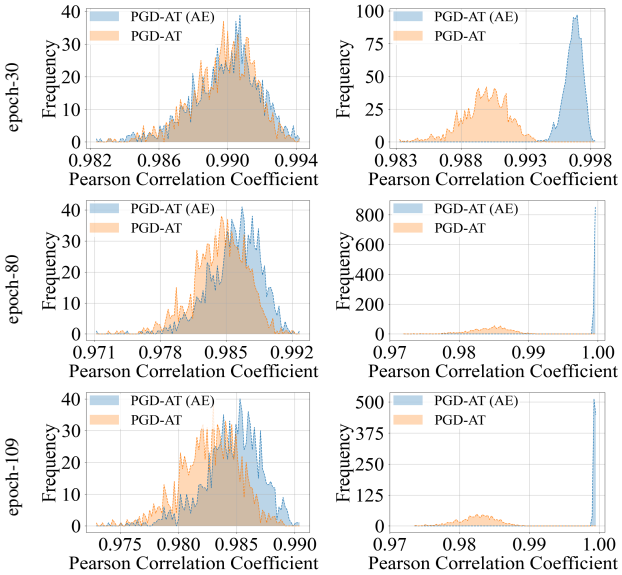

Figure 2: Frequency histograms of the PCC between batch normalization (BN) statistics for clean and adversarial images. Higher PCC values indicate smaller distributional differences relative to clean images. The histograms in the first column illustrate the distributional differences in the mean statistics from the inner optimization results of the model, while the second column shows the distributional differences of adversarial examples generated by the PGD attack.

and without manifold projection are provided in Supplementary Material (SM)[1].

For PGD-AT with manifold projection, it can be observed that the PCC values between clean and adversarial example statistics generated by the PGD attack gradually approach 1. This suggests that the manifold projection pulls adversarial

[1]https://github.com/TaiXiangJiang/Enhancing-the-Adversarial-Robustness-via-Manifold-Projection/

examples more closely with the clean sample manifold. Fig. 1-(a) symbolically visualizes this alignment in a 3-D sample space. Consequently, manifold projection significantly weakens the attack strength of adversarial examples, making it easier to learn mappings from adversarial examples to correct labels.

To further conceptualize this effect, consider a metaphor: the model is like a *martial artist*, and adversarial examples are *opponents* exploiting the artist's *weaknesses*. Traditional adversarial training is akin to the artist training specifically to counter these attacks. The *manifold projection* acts as a *shield*, enhancing the artist's defense. By integrating this projection into the defense strategy, we create a scenario where the artist trains specifically against *spear-wielding opponents* (adversarial examples) using the shield (manifold projection). Once the opponents lose their spear, their attacks can be easily deflected. Thus, PGD-AT with manifold projection produces a more *robust* model, which can also serve as a *teacher model* in AD.

3.2 AT with Manifold Projection

In this part, we use PGD-AT as an example to introduce how to integrate manifold projection into adversarial training. We propose incorporating manifold projection into the robust framework of PGD-AT, with the autoencoder $\mathtt{M}$ positioned as a crucial, yet subtle, component preceding the model.

The autoencoder $\mathtt{M}$ is trained by minimizing Eq. (4) and is used to perform the manifold projection. Once $\mathtt{M}$ is trained, adversarial examples are generated during the inner optimization phase of AT. Mathematically, given a model $F(\cdot)$, parameterized by θ_F, an input $\mathbf{x}$, and a perturbation size ϵ, the inner optimization aims to find the "support" point $\mathbf{x}^*$ nearing the neighborhood of $\mathbf{x}$. This point maximizes the prediction discrepancy between the classifier and the ground-truth labels under the manifold projection and is formulated as:

$$\mathbf{x}^* = \mathbf{x} + \arg\max_{\|\delta\|_p \leq \epsilon} \ell(F(\mathtt{M}(\mathbf{x}+\delta)), p_d(y|\mathbf{x})). \quad (5)$$

The Cross-Entropy loss is adopted to measure the discrepancy between the model's output probabilities and the ground-truth labels. Similar to traditional AT methods, a projected gradient descent strategy is employed to obtain the "support" point $\mathbf{x}^*$ for training. Subsequently, the outer optimization then seeks to minimize the upper bound of the prediction discrepancy under the manifold projection, defined as:

$$\arg\min_{\theta_F} \ell(F(\mathtt{M}(\mathbf{x}^*)), p_d(y|\mathbf{x})). \quad (6)$$

Fig. 1-(b) illustrates the inner optimization process of PGD-AT with the manifold projection. As discussed in Sect. 3.1, the manifold projection can significantly weaken the attack strength of adversarial samples generated by adversarial attacks.

3.3 AD with Manifold Projection

In this part, we use AdaAD as an example to introduce adversarial distillation with manifold projection. AdaAD aims to reduce the prediction discrepancy between student and teacher models by achieving the highest degree of point-to-point alignment. However, Cho and Hariharan (Cho and Hariharan 2019) found that the student often struggles to fully mimic the teacher, indicating a mismatch between their capacities. Therefore, an exact match via KL divergence may be overly ambitious and challenging, given the model capacity discrepancy between student and teacher models (Huang et al. 2022). Integrating the manifold projection into the AdaAD method can enhance the distillation process and improve the student model's learning ability, as described in **SM**. Fig. 1-(c) exhibits the inner optimization process of AdaAD with manifold projection.

To begin with, we train the autoencoder $\mathtt{M}(\cdot)$ on the clean samples where the loss function is calculated by Eq. (4). Then, as detailed in Sect. 3.2, we obtain the robust teacher model $T(\cdot)$ by adversarial training with the manifold projection. Given a student model $S(\cdot)$ parameterized by θ_S, an input $\mathbf{x}$, and a perturbation size ϵ, our inner optimization seeks to find $\mathbf{x}^*$ within the neighborhood of $\mathbf{x}$, which maximizes the prediction discrepancy between the student and teacher models with the manifold projection. This optimization can be formulated as

$$\mathbf{x}^* = \mathbf{x} + \arg\max_{\|\delta\|_p \leq \epsilon} \mathrm{KL}(S(\mathtt{M}(\mathbf{x}+\delta)), T(\mathtt{M}(\mathbf{x}+\delta))), \quad (7)$$

Then, the outer optimization is to minimize the upper bound of prediction discrepancy with the manifold projection to perform adversarial distillation, defined as

$$\begin{aligned}\arg\min_{\theta_S} [&(1-\alpha)\mathrm{KL}(S(\mathtt{M}(\mathbf{x})), T(\mathtt{M}(\mathbf{x})))+ \\ &\alpha\mathrm{KL}(S(\mathtt{M}(\mathbf{x}^*)), T(\mathtt{M}(\mathbf{x}^*)))].\end{aligned} \quad (8)$$

We hypothesize that the teacher model, with its larger model capacity compared to the student model, will generate more informative adversarial examples that are challenging for the student model. Therefore, we propose 'Difficult Adversarial Distillation' (DAD) which uses the teacher model's adversarial examples as the searching result $\mathbf{x}^*$ of the inner optimization. The searching result $\mathbf{x}^*$ is formulated as

$$\mathbf{x}^* = \mathbf{x} + \arg\max_{\|\delta\|_p \leq \epsilon} \mathrm{KL}(T(\mathtt{M}(\mathbf{x}+\delta)), T(\mathtt{M}(\mathbf{x}))). \quad (9)$$

Meanwhile, the outer optimization of DAD remains unchanged as defined in Eq. (8), aiming to enforce the student model to learn from the teacher model's difficult adversarial examples. Considering that the adversarial robustness of the student model generally arises very slightly in the latter half of the training epochs, we recommend combining DAD in the training every ten batches of data to further enhance the robustness of the student model.

3.4 Preventive and Remedial Measures

Both adversarial training and distillation with manifold projection, as depicted in Sect. 3.2 and Sect. 3.3, have a notable limitation: *the autoencoder must remain undetected by the adversary*. If the autoencoder is detected, the adversary would "*wield a spear*" and target both the autoencoder and the classifier as a combined unit to generate adversarial examples, significantly weakening the autoencoder's effectiveness. Therefore, preventive and remedial measures are essential.

Layer	Operations	# Kernel	Kernel Size
1	Convolutional (Sigmoid)	3	3×3
2	Convolutional (Sigmoid)	3	3×3
3	Convolutional (Sigmoid)	3	3×3

Table 1: The autoencoder architecture used for CIFAR-10 and CIFAR-100 datasets.

Preventive Measure We diversify our defense by creating a large number of autoencoders with random initialization as an interference group. We randomly select one of these autoencoders for subsequent training, while the others serve to interfere with the adversary. Assuming the adversary cannot predict which autoencoder we will select, and that successful adversarial examples trained on one autoencoder have a low probability of succeeding on others, the adversary would need to train their adversarial examples to be effective against all autoencoders in the interference group.

Remedial Measure We create a reserve group of autoencoders trained on the same clean dataset as the selected autoencoder. If the adversary discovers the chosen autoencoder, we can replace it with the best-performing autoencoder from the reserve group, without retraining the classifier.

By implementing both preventive and remedial measures, we ensure that we have sufficient time to update the autoencoder and classifier while maintaining the original robustness as much as possible, even if the adversary discovers the autoencoder. Although these measures are engineering-oriented, as we will show in the experimental part, they are implementation-friendly and effective, thereby significantly enhancing the practicality of our paradigm.

4 Experimental Evaluations

Experimental Setup We evaluate the effectiveness of our proposed adversarial defense paradigm using three benchmark image datasets: CIFAR-10, CIFAR-100[2] (Alex 2009), and Tiny ImageNet (Le and Yang 2015). In all cases, the image pixel values are normalized to the range [0,1]. Our approach integrates manifold projection via an autoencoder into the robust paradigms of two common adversarial training (AT) methods—PGD-AT and TRADES—and several representative adversarial distillation (AD) methods, including ARD, RSLAD, and AdaAD, as well as a knowledge distillation (KD) method. The architecture of the autoencoder used for CIFAR-10 and CIFAR-100 is detailed in Table 1. For Tiny ImageNet, we employ a simplified U-Net (Ronneberger, Fischer, and Brox 2015) consisting of one downsampling and one upsampling block as the autoencoder. Our methods are compared against the original versions that do not utilize manifold projection. The details of the student and teacher networks are available in **SM**.

Implementation Details The networks are trained using the Stochastic Gradient Descent (SGD) optimizer with an initial learning rate of 0.1, momentum of 0.9, and weight decay of 5×10^{-4}. Unless otherwise specified, for PGD-AT, we train for 110 epochs, reducing the learning rate by a factor of 10 at the 100th and 105th epochs. For TRADES[3] and the other AD methods, we train for 200 epochs, with learning rate reductions at the 100th and 150th epochs. The inner optimization involves 10 iterations with a step size of $2/255$, and the total perturbation bound $\epsilon = 8/255$ under the L_∞ constraint. For CIFAR-10, the distillation temperature τ is set to 30 in all distillation methods, with $\alpha = 5/6$ in RSLAD, and $\alpha = 1.0$ in KD, ARD, and AdaAD. For CIFAR-100 and Tiny ImageNet, we set τ = 5 in all distillation methods, with $\alpha = 0.95$ in KD, $\alpha = 5/6$ in RSLAD, and $\alpha = 1.0$ in ARD and AdaAD. The α = 0.95 in the ARD with manifold projection for CIFAR-100 differs from the baseline method ARD; aside from this, the parameters in our proposed methods strictly follow the settings of their respective baseline methods.

Evaluation Metrics We assess performance using two metrics: natural (clean) accuracy on normal test samples and robust accuracy on adversarial test samples. Four representative adversarial attacks are considered: FGSM, PGD, CW_2 (constrained by the ℓ_2 norm), and AutoAttack (AA). The maximum perturbation for evaluation is set as $\epsilon = 8/255$ for all datasets, and the balance constant in CW is set to 0.1, following (Huang et al. 2023). Unless stated otherwise, results are reported from the checkpoint with the highest PGD-10 accuracy.

4.1 Results on CIFAR-10 and Tiny ImageNet

Tables 2 and 3 respectively present the recognition accuracy of models trained using our proposed AT and AD methods with manifold projection, alongside their original counterparts, under various adversarial attacks. The results indicate that all of our methods significantly outperform the original AT and AD approaches across all attack types. Notably, our methods with manifold projection show the most substantial improvements under the AutoAttack (AA), which is recognized as the most powerful among the four evaluated attacks. For instance, our proposed PGD-AT (AE) and AdaAD (AE) achieve improvements of up to 33.75% and 34.00% on CIFAR-10, respectively. These findings affirm that integrating manifold projection via an autoencoder into the existing AT and AD paradigms is both effective and broadly applicable for enhancing model robustness. Additionally, since AA includes query-based attacks, the significant improvements in robust accuracy under AA suggest that our methods with manifold projection maintain reliability even when confronted with query-based attacks. Furthermore, integrating manifold projection into AD methods reduces the robustness gap between teacher and student models, thereby accelerating the distillation process and enhancing the student model's learning capability. Specifically, for AdaAD (AE) on CIFAR-10, the gap in AA accuracy between ResNet-18 and WRN-34-20 decreases from 6.58% to 0.66%.

[2]Results on CIFAR-100 are provided in **SM**.

[3]For Tiny ImageNet, we train for 110 epochs in TRADES with and without manifold projection.

Model	ResNet-18					MobileNetV2				
Method	Clean	FGSM	PGD	CW	AA	Clean	FGSM	PGD	CW	AA
PGD-AT (Madry et al. 2018)	83.75	58.72	53.51	77.77	48.45	77.42	53.46	49.66	72.53	44.34
PGD-AT (AE)	82.34	79.21	82.19	82.31	82.20	75.18	72.26	74.22	75.08	74.69
△	-1.41	+20.49	+28.68	+4.54	+33.75	-2.24	+18.80	+24.56	+2.55	+30.35
TRADES (Zhang et al. 2019)	83.03	59.11	53.60	76.99	49.82	80.03	55.28	51.05	75.59	46.34
TRADES (AE)	81.78	79.63	81.41	81.77	81.78	78.04	75.22	77.06	78.02	77.59
△	-1.25	+20.52	+27.81	+4.78	+31.96	-1.99	+19.94	+26.01	+2.43	+31.25
KD	87.82	41.24	9.39	67.31	1.86	72.78	19.90	2.74	17.77	0.05
KD (AE)	**88.46**	64.52	68.10	**87.39**	77.74	**84.92**	60.20	74.46	84.00	79.35
△	+0.64	+23.28	+58.71	+20.08	+75.88	+12.14	+40.30	+71.72	+66.23	+79.30
ARD (Goldblum et al. 2020)	83.35	59.25	54.56	78.60	49.83	80.27	56.14	52.42	76.09	47.74
ARD (AE)	83.24	79.74	82.80	83.10	82.95	82.06	79.19	81.39	82.00	81.75
△	-0.11	+20.49	+28.24	+4.50	+33.12	+1.79	+23.05	+28.97	+5.91	+34.01
RSLAD (Zi et al. 2021)	84.08	59.74	54.76	79.19	49.92	81.67	56.19	52.10	76.67	46.95
RSLAD (AE)	84.27	80.70	83.70	84.2	84.03	82.82	79.16	81.85	82.73	82.54
△	+0.19	+20.96	+28.94	+5.01	+34.11	+1.15	+22.97	+29.75	+6.06	+35.59
AdaAD (Huang et al. 2023)	85.45	61.03	56.50	81.22	51.15	84.05	57.87	53.31	79.56	48.14
AdaAD (AE)	85.27	**82.45**	**85.06**	85.25	**85.15**	84.36	**81.05**	**83.48**	**84.32**	**84.08**
△	-0.18	+21.42	+28.56	+4.03	+34.00	+0.31	+23.18	+30.17	+4.76	+35.94

Table 2: Model robustness measured by classification accuracy (%) under various adversarial attacks on the CIFAR-10 dataset. The highest accuracy for each scenario is boldfaced, while the second-highest (suboptimal) results are underlined.

Model	ResNet-18					MobileNetV2				
Method	Clean	FGSM	PGD	CW	AA	Clean	FGSM	PGD	CW	AA
PGD-AT (Madry et al. 2018)	50.19	26.56	24.34	45.57	19.18	38.93	20.47	19.08	35.14	13.73
PGD-AT (AE)	50.10	30.93	29.80	49.77	33.68	38.70	23.43	22.84	38.34	26.16
△	-0.09	+4.37	+5.46	+4.20	+14.50	-0.23	+2.96	+3.76	+3.20	+12.43
TRADES (Zhang et al. 2019)	50.74	25.60	23.93	45.66	17.97	42.90	19.93	18.82	38.18	13.30
TRADES (AE)	50.79	30.17	29.35	50.15	33.88	42.76	23.42	22.72	42.32	26.78
△	+0.05	+4.57	+5.42	+4.49	+15.91	-0.14	+3.49	+3.90	+4.14	+13.48
RSLAD (Zi et al. 2021)	48.80	26.89	24.26	44.27	18.74	46.01	25.84	23.76	41.96	17.63
RSLAD (AE)	48.16	30.76	30.00	47.77	33.41	46.02	29.27	28.59	45.68	31.49
△	-0.64	+3.87	+5.74	+3.50	+14.67	+0.01	+3.43	+4.83	+3.72	+13.86
AdaAD (Huang et al. 2023)	**53.73**	29.39	26.77	48.81	21.36	50.50	25.11	22.50	45.47	17.29
AdaAD (AE)	53.58	**33.45**	**32.35**	**53.09**	**36.51**	**50.63**	**29.67**	**28.79**	**50.05**	**32.99**
△	-0.15	+4.06	+5.58	+4.28	+15.15	+0.13	+4.56	+6.29	+4.58	+15.70

Table 3: Model robustness measured by classification accuracy (%) under various adversarial attacks on the Tiny ImageNet dataset. The highest accuracy for each scenario is boldfaced, while the second-highest (suboptimal) results are underlined.

4.2 Discussions

In this section, we further explore the effectiveness of our proposed methods with manifold projection from additional perspectives, such as model robust fairness, resistance to transfer-based attacks, and comparison with adversarial purification approaches. Due to space constraints, some discussions are included in SM.

Model Robust Fairness Although AT and AD methods have achieved notable robustness, a significant disparity in class-wise robustness remains in adversarially trained models, with certain classes demonstrating strong robustness while others are notably vulnerable. This disparity raises concerns regarding robustness fairness. Following Li and Liu (Li and Liu 2023), we evaluate robust fairness using average natural accuracy, average robust accuracy, worst-class natural accuracy, and worst-class robust accuracy for some AT and AD methods with and without manifold projection, as shown in Table 4. From Table 4, our methods consistently improve both average robustness and worst-class robustness compared to the original approaches, particularly against PGD and AA attacks. Notably, the improvement in worst-class robustness surpasses that of average robustness, especially against AA. These observations suggest that our methods explicitly address the disparity in class-wise robustness and enhance robust fairness.

Evaluation on Transfer-based Attacks We also evaluate whether our proposed methods with manifold projection,

Method	Clean		PGD		AA	
ResNet-18	Avg	Worst	Avg	Worst	Avg	Worst
TRADES	83.03	64.40	53.60	25.10	49.82	19.50
with AE	81.78	62.80	81.41	62.00	81.78	62.80
△	-1.25	-1.60	+27.81	+36.90	+31.96	+43.30
AdaAD	85.45	70.60	56.50	30.50	51.15	22.10
with AE	85.27	70.60	85.06	71.10	85.15	70.70
△	-0.18	0	+28.56	+40.60	+34.00	+48.60

Table 4: Model robust fairness, measured by classification accuracy (%) under various attacks on the CIFAR-10 dataset. "Avg" and "Worst" denote the average accuracy and the worst-class accuracy, respectively.

Surrogate	ResNet-34			VGG-16		
Method	PGD	CW	AA	PGD	CW	AA
PGD-AT (110)	61.27	82.23	61.94	63.50	**83.36**	67.36
with AE	61.13	81.69	63.96	62.13	82.08	66.61
PGD-AT (200)	61.13	82.13	62.79	63.26	82.91	67.71
with AE	**63.46**	**82.74**	**68.02**	**65.31**	83.22	**71.09**
RSLAD	63.65	82.85	66.17	65.76	83.80	70.90
with AE	**64.08**	83.72	**67.80**	66.05	84.09	71.64
AE + DAD	64.02	**84.04**	67.72	**66.13**	**84.45**	**71.78**
AdaAD	64.27	84.41	66.94	**67.20**	85.13	72.82
with AE	64.24	84.60	68.23	66.42	85.14	72.81
AE + DAD	**64.32**	**85.14**	**68.33**	67.03	**85.51**	**73.19**

Table 5: Classification accuracy (%) under transfer-based *black-box* attacks for ResNet-18 models trained by AT and AD methods with and without the manifold projection, and their variants over CIFAR-10. PGD-AT (110) and PGD-AT (200) are abbreviations of 110 epochs PGD-AT and 200 epochs PGD-AT. DAD stands for difficult adversarial distillation, a variation where more challenging adversarial examples generated by the teacher model are used to improve the student model's robustness.

along with their variants, effectively resist transfer-based attacks. Following the methodology of Huang *et al.* (Huang et al. 2023), we train two surrogate models with different architectures—ResNet-34 and VGG-16—using the PGD-AT method with early stopping. We then generate adversarial examples using these surrogate models to assess the effectiveness of the ResNet-18 model trained with our proposed methods.

As exhibited in Table 5, extending the number of training epochs does not improve robustness against transfer-based attacks for the original PGD-AT method. In contrast, our proposed PGD-AT (AE) significantly enhances model robustness, increasing AA accuracy from 63.96% to 68.02% when tested against the ResNet-34 surrogate model and from 66.61% to 71.09% against the VGG-16 surrogate model. Moreover, when combined with DAD, our proposed RSLAD (AE) and AdaAD (AE) outperform their original methods in most scenarios. Therefore, our proposed AT and AD methods with manifold projection are effective in mitigating transfer-based attacks.

Model	WideResNet-28-10		WideResNet-70-16	
Method	PGD	AA	PGD	AA
DiffPure	46.84	63.60	51.13	66.06
RE-DiffPure	55.82	70.47	56.88	70.31
PGD-AT (AE)	**84.66**	**84.88**	**84.02**	**85.08**

Table 6: Comparison with DiffPure and RE-DiffPure under various adversarial attacks on the CIFAR-10 dataset.

Comparison with Adversarial Purification Adversarial purification, first introduced in Defense-GAN (Samangouei, Kabkab, and Chellappa 2018), is a defense strategy that utilizes generative models to remove adversarial perturbations. Building on this concept, Nie *et al.* (Nie et al. 2022) proposed DiffPure, which leverages diffusion models for adversarial purification. RE-DiffPure (Lee and Kim 2023) improves the robustness of DiffPure by incorporating a gradual noise-scheduling strategy. We compare our proposed PGD-AT (AE) with DiffPure and RE-DiffPure on CIFAR-10. As exhibited in Table 6, the robustness of models trained with PGD-AT (AE) significantly outperforms that of both DiffPure and RE-DiffPure. Our method differs from those approaches in that the autoencoder is fully integrated into the adversarial training process, directly enhancing the model's learning of robust features. In contrast, DiffPure and similar methods apply purification post-training. This key difference allows our method to achieve superior robustness.

5 Conclusion

In this paper, we proposed a novel adversarial defense paradigm by integrating manifold projection via an autoencoder into the robust frameworks of existing Adversarial Training (AT) and Adversarial Distillation (AD) methods. The manifold projection aligns adversarial samples with the manifold of clean examples, thereby weakening attack strength and simplifying the learning process from adversarial samples to correct labels. Additionally, incorporating manifold projection into AD methods facilitates the distillation process, reducing the complexity of transferring robustness from teacher to student models. Extensive experiments on three benchmark datasets demonstrate that our proposed AT and AD methods with manifold projection significantly outperform previous state-of-the-art methods across various adversarial attacks, highlighting the effectiveness and versatility of our approach in enhancing model robustness.

Limitation A key limitation of our method is the need to keep the autoencoder undetected by adversaries, as detection could undermine its effectiveness. To address this, we proposed preventive measures such as diversifying defenses with randomly initialized autoencoders and remedial measures like maintaining a reserve of autoencoders for quick replacement if necessary. These strategies ensure that our approach remains robust and practical, even against adaptive adversaries.

Acknowledgements

This work was supported in part by the Sichuan Science and Technology Program under Grant 2024ZYD0147, in part by the Natural Science Foundation of Xinjiang Uygur Autonomous Region under Grant 2024D01A18, in part by the National Natural Science Foundation of China (NSFC) under Grant 62376228, in part by Chengdu Science and Technology Program under Grant 2023 JB00-00016-GX, and in part by the Guanghua Talent Project.

References

Alex, K. 2009. Learning Multiple Layers of Features from Tiny Images. *Master's thesis, University of Tront*.

Andriushchenko, M.; and Flammarion, N. 2020. Understanding and Improving Fast Adversarial Training. In *Advances in Neural Information Processing Systems*, 16048–16059.

Brendel, W.; Rauber, J.; and Bethge, M. 2018. Decision-Based Adversarial Attacks: Reliable Attacks Against Black-Box Machine Learning Models. In *International Conference on Learning Representations*, 1983–1994.

Carlini, N.; and Wagner, D. 2017. Towards Evaluating the Robustness of Neural Networks. In *IEEE Symposium on Security and Privacy*, 39–57.

Cho, J. H.; and Hariharan, B. 2019. On the Efficacy of Knowledge Distillation. In *International Conference on Computer Vision*, 4794–4802.

Croce, F.; and Hein, M. 2020. Reliable Evaluation of Adversarial Robustness with an Ensemble of Diverse Parameter-free Attacks. In *International Conference on Machine Learning*, 2206–2216.

Dia, O. A.; Barshan, E.; and Babanezhad, R. 2019. Semantics Preserving Adversarial Learning. *CoRR*, abs/1903.03905.

Goldblum, M.; Fowl, L.; Feizi, S.; and Goldstein, T. 2020. Adversarially Robust Distillation. In *AAAI Conference on Artificial Intelligence*, 04, 3996–4003.

Goodfellow, I. J.; Shlens, J.; and Szegedy, C. 2015. Explaining and Harnessing Adversarial Examples. In *International Conference on Learning Representations*.

Guo, C.; Gardner, J. R.; You, Y.; Wilson, A. G.; and Weinberger, K. Q. 2019. Simple Black-Box Adversarial Attacks. In *International Conference on Machine Learning*, 2484–2493.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep Residual Learning for Image Recognition. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 770–778.

Hinton, G. E.; Vinyals, O.; and Dean, J. 2015. Distilling the Knowledge in a Neural Network. *CoRR*, abs/1503.02531.

Huang, B.; Chen, M.; Wang, Y.; Lu, J.; Cheng, M.; and Wang, W. 2023. Boosting Accuracy and Robustness of Student Models via Adaptive Adversarial Distillation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 24668–24677.

Huang, T.; You, S.; Wang, F.; Qian, C.; and Xu, C. 2022. Knowledge Distillation from a Stronger Teacher. In *Advances in Neural Information Processing Systems*, 33716–33727.

Ilyas, A.; Engstrom, L.; Athalye, A.; and Lin, J. 2018. Black-Box Adversarial Attacks with Limited Queries and Information. In *International Conference on Machine Learning*, 2137–2146.

Kong, Z.; Guo, J.; Li, A.; and Liu, C. 2020. PhysGAN: Generating Physical-World-Resilient Adversarial Examples for Autonomous Driving. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 14242–14251.

Krizhevsky, A.; Sutskever, I.; and Hinton, G. E. 2012. ImageNet Classification with Deep Convolutional Neural Networks. In *Advances in Neural Information Processing Systems*, 1106–1114.

Le, Y.; and Yang, X. 2015. Tiny ImageNet Visual Recognition Challenge. *CS 231N*, 7(7): 3.

Lee, M.; and Kim, D. 2023. Robust Evaluation of Diffusion-Based Adversarial Purification. In *International Conference on Computer Vision*, 134–144.

Li, B.; and Liu, W. 2023. WAT: Improve the Worst-Class Robustness in Adversarial Training. In *AAAI Conference on Artificial Intelligence*, 14982–14990.

Li, Z.; Yin, B.; Yao, T.; Guo, J.; Ding, S.; Chen, S.; and Liu, C. 2023. Sibling-Attack: Rethinking Transferable Adversarial Attacks against Face Recognition. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 24626–24637.

Liu, Y.; Cheng, Y.; Gao, L.; Liu, X.; Zhang, Q.; and Song, J. 2022. Practical Evaluation of Adversarial Robustness via Adaptive Auto Attack. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 15084–15093.

Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; and Guo, B. 2021. Swin Transformer: Hierarchical Vision Transformer Using Shifted Windows. In *International Conference on Computer Vision*, 9992–10002.

Madry, A.; Makelov, A.; Schmidt, L.; Tsipras, D.; and Vladu, A. 2018. Towards Deep Learning Models Resistant to Adversarial Attacks. In *International Conference on Learning Representations*, 4138–4160.

Mahmood, K.; Mahmood, R.; and Van Dijk, M. 2021. On the Robustness of Vision Transformers to Adversarial Examples. In *International Conference on Computer Vision*, 7838–7847.

Meng, D.; and Chen, H. 2017. MagNet: a Two-Pronged Defense against Adversarial Examples. In *ACM SIGSAC Conference on Computer and Communications Security*, 135–147.

Moosavi-Dezfooli, S.; Fawzi, A.; and Frossard, P. 2016. DeepFool: A Simple and Accurate Method to Fool Deep Neural Networks. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2574–2582.

Ng, A. Y.; and Jordan, M. I. 2001. On Discriminative VS. Generative Classifiers: A Comparison of Logistic Regression and Naive Bayes. In *Advances in Neural Information Processing Systems*, 841–848.

Nie, W.; Guo, B.; Huang, Y.; Xiao, C.; Vahdat, A.; and Anandkumar, A. 2022. Diffusion Models for Adversarial Purification. In *International Conference on Machine Learning*, 16805–16827.

Pang, T.; Lin, M.; Yang, X.; Zhu, J.; and Yan, S. 2022. Robustness and Accuracy Could Be Reconcilable by (Proper) Definition. In *International Conference on Machine Learning*, 17258–17277.

Papernot, N.; McDaniel, P. D.; and Goodfellow, I. J. 2016. Transferability in Machine Learning: from Phenomena to Black-Box Attacks Using Adversarial Samples. *CoRR*, abs/1605.07277.

Poursaeed, O.; Katsman, I.; Gao, B.; and Belongie, S. 2018. Generative Adversarial Perturbations. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4422–4431.

Ronneberger, O.; Fischer, P.; and Brox, T. 2015. U-Net: Convolutional Networks for Biomedical Image Segmentation. In *Medical Image Computing and Computer-Assisted Intervention*, 234–241.

Samangouei, P.; Kabkab, M.; and Chellappa, R. 2018. Defense-GAN: Protecting Classifiers Against Adversarial Attacks Using Generative Models. In *International Conference on Learning Representations*, 4385–4401.

Sandler, M.; Howard, A.; Zhu, M.; Zhmoginov, A.; and Chen, L.-C. 2018. MobileNetV2: Inverted Residuals and Linear Bottlenecks. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4510–4520.

Simonyan, K.; and Zisserman, A. 2015. Very Deep Convolutional Networks for Large-Scale Image Recognition. In *International Conference on Learning Representations*.

Sun, Q.; Yao, X.; Rao, A. A.; Yu, B.; and Hu, S. 2022. Counteracting Adversarial Attacks in Autonomous Driving. *Transactions on Computer-Aided Design of Integrated Circuits and Systems*, 41(12): 5193–5206.

Tan, M.; and Le, Q. V. 2019. EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks. In *International Conference on Machine Learning*, 6105–6114.

Tramèr, F.; Papernot, N.; Goodfellow, I. J.; Boneh, D.; and McDaniel, P. D. 2017. The Space of Transferable Adversarial Examples. *CoRR*, abs/1704.03453.

Wang, Z.; Yang, H.; Feng, Y.; Sun, P.; Guo, H.; Zhang, Z.; and Ren, K. 2023. Towards Transferable Targeted Adversarial Examples. In *IEEE/CVF Conference on Computer Cision and Pattern Recognition*, 20534–20543.

Wong, E.; Rice, L.; and Kolter, J. Z. 2020. Fast Is Better Than Free: Revisiting Adversarial Training. In *International Conference on Learning Representations*, 5849–5865.

Xiao, C.; Li, B.; Zhu, J.; He, W.; Liu, M.; and Song, D. 2018. Generating Adversarial Examples with Adversarial Networks. In *International Joint Conference on Artificial Intelligence*, 3905–3911.

Xie, C.; and Yuille, A. 2020. Intriguing Properties of Adversarial Training at Scale. In *International Conference on Learning Representations*, 4799–4812.

Yu, C.; Chen, T.; and Gan, Z. 2023. Adversarial Amendment Is the Only Force Capable of Transforming an Enemy into a Friend. In *International Joint Conference on Artificial Intelligence*, 4522–4530.

Zhang, H.; Yu, Y.; Jiao, J.; Xing, E.; El Ghaoui, L.; and Jordan, M. 2019. Theoretically Principled Trade-off between Robustness and Accuracy. In *International Conference on Machine Learning*, 7472–7482.

Zhang, J.; Zhu, J.; Niu, G.; Han, B.; Sugiyama, M.; and Kankanhalli, M. 2021. Geometry-aware Instance-reweighted Adversarial Training. In *International Conference on Learning Representations*.

Zhao, Z.; Dua, D.; and Singh, S. 2018. Generating Natural Adversarial Examples. In *International Conference on Learning Representations*, 689–703.

Zheng, J.; Lin, C.; Sun, J.; Zhao, Z.; Li, Q.; and Shen, C. 2024. Physical 3D Adversarial Attacks against Monocular Depth Estimation in Autonomous Driving. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 24452–24461.

Zhu, J.; Yao, J.; Han, B.; Zhang, J.; Liu, T.; Niu, G.; Zhou, J.; Xu, J.; and Yang, H. 2022. Reliable Adversarial Distillation with Unreliable Teachers. In *International Conference on Learning Representations*.

Zi, B.; Zhao, S.; Ma, X.; and Jiang, Y.-G. 2021. Revisiting Adversarial Robustness Distillation: Robust Soft Labels Make Student Better. In *International Conference on Computer Vision*, 16443–16452.