Out of Thin Air: Exploring Data-Free Adversarial Robustness Distillation

Yuzheng Wang[1*], **Zhaoyu Chen**[1*], **Dingkang Yang**[1], **Pinxue Guo**[1]
Kaixun Jiang[1], **Wenqiang Zhang**[1,2], **Lizhe Qi**[1†]
[1]Shanghai Engineering Research Center of AI & Robotics, Academy for Engineering & Technology, Fudan University
[2]Engineering Research Center of AI & Robotics, Ministry of Education, Academy for Engineering & Technology, Fudan University
{yzwang20, zhaoyuchen20}@fudan.edu.cn

Abstract

Adversarial Robustness Distillation (ARD) is a promising task to solve the issue of limited adversarial robustness of small capacity models while optimizing the expensive computational costs of Adversarial Training (AT). Despite the good robust performance, the existing ARD methods are still impractical to deploy in natural high-security scenes due to these methods rely entirely on original or publicly available data with a similar distribution. In fact, these data are almost always private, specific, and distinctive for scenes that require high robustness. To tackle these issues, we propose a challenging but significant task called Data-Free Adversarial Robustness Distillation (DFARD), which aims to train small, easily deployable, robust models without relying on data. We demonstrate that the challenge lies in the lower upper bound of knowledge transfer information, making it crucial to mining and transferring knowledge more efficiently. Inspired by human education, we design a plug-and-play Interactive Temperature Adjustment (ITA) strategy to improve the efficiency of knowledge transfer and propose an Adaptive Generator Balance (AGB) module to retain more data information. Our method uses adaptive hyperparameters to avoid a large number of parameter tuning, which significantly outperforms the combination of existing techniques. Meanwhile, our method achieves stable and reliable performance on multiple benchmarks.

Introduction

Deep learning has achieved great success in many fields (Devlin et al. 2018; Dosovitskiy et al. 2020; Yang et al. 2023c,a,b,d; Liu et al. 2023b,c,a; Wang et al. 2023c,b). Along with this process, deep learning models are increasingly expected to be deployed in established and emerging artificial intelligence fields. However, high-performance models' large scale and high computational costs (Ramesh et al. 2022) prevent this technology from being applied to mobile devices, driverless cars, and tiny robots. More importantly, many studies have shown that well-trained deep learning models are vulnerable to adversarial examples containing only minor changes (Goodfellow, Shlens, and Szegedy 2014; Chen et al. 2022a, 2023). Therefore, training robust small-capacity models has become the key to breaking the bottleneck.

Various defensive strategies have been proposed (Madry et al. 2017; Jia et al. 2019; Chen et al. 2022b; Wang et al. 2023a) for adversarial robustness. Among them, Adversarial Training (AT) has been considered the most effective approach (Athalye, Carlini, and Wagner 2018; Croce and Hein 2020). By generating adversarial examples, the models can learn robustness knowledge to deal with various adversarial attacks. Therefore, it can significantly improve the robustness of large-capacity models. However, the robustness performance is struggling for small models only relying on AT due to the limited model capacity. Based on this, guided by the pre-trained robust teacher with insights, the robustness of small models is improved. This process is called Adversarial Robustness Distillation (ARD) (Goldblum et al. 2020).

Despite improving the robustness of small models, existing ARD methods are still hard to apply in real-world scenes due to impractical settings. The original training data is of primary concern. Firstly, all existing ARD methods assume that the original training data is available throughout the distillation process (Goldblum et al. 2020; Zhao et al. 2022). In practical application, for scenes with high robustness requirements, the original data is usually private and unavailable (*e.g.*, face data for face recognition system, disease data for medical diagnosis, and financial data for quantitative investment). Secondly, some technologies avoid relying on original private data by using open-world or out-of-domain (OOD) unlabeled data (Fang et al. 2021a). However, these methods rely on a necessary assumption that private data can always be obtained simply from open datasets. Although some methods claim to use OOD data, the performance of these methods degrades drastically when the discrepancy between the unavailable original data and unlabeled data increases. Therefore, good performance extremely depends on broadly similar image patterns between the data domains instead of OOD (Yang et al. 2021). Based on these, existing technologies are still challenging to deploy in high-security robustness scenes. **One question is whether we can efficiently train small, easily deployable, robust models to improve robustness without original private data and specific data with similar patterns**. To explore this question, we propose a novel task called Data-Free Adversarial Robustness Distillation (DFARD). The diagrams are shown in Figure 1. Compared with the existing KD (a) and ARD (b) tasks, our DFARD only uses generated data, which is more general and practical.

[*]Equal contributions

[†]Corresponding author

Copyright © 2024, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

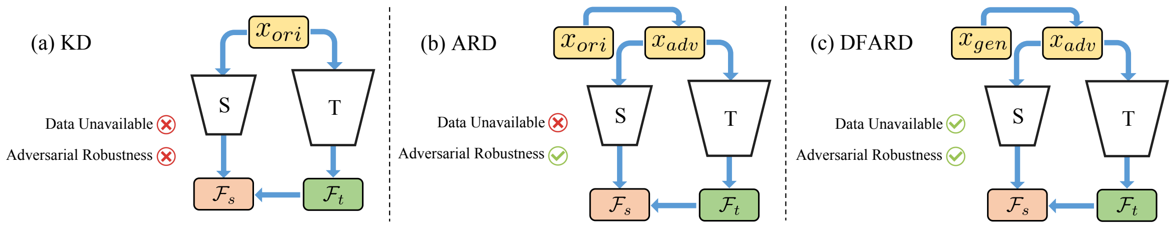

Figure 1: Diagrams of (a) Knowledge Distillation (KD), (b) Adversarial Robustness Distillation (ARD), and (c) Data-Free Adversarial Robustness Distillation (DFARD). S and T represent the student and the teacher network respectively. $\mathcal{F}_s$ and $\mathcal{F}_t$ represent the search spaces. x_{ori} and x_{gen} is the original and generated data. x_{adv} is the adversarial examples.

Considering the knowledge transfer process between the teacher and student networks, we demonstrate that the information upper bound is lower in the DFARD than in existing tasks. While removing the ARD task's dependence on private data, the challenges lie in less effective knowledge transfer and less data knowledge in the generated data.

To tackle the issues, we select the commonly used generator training objectives as a DFARD baseline and optimize it from the following aspects: **1)** To improve the effectiveness of knowledge transfer, we first propose an Interactive Temperature Adjustment (ITA) strategy to help students find more suitable training objectives for each training epoch. **2)** To retain more data information, we then propose an Adaptive Generator Balance (AGB) module to better balance the similarity of the data domains and the information content. In addition, our method uses adaptive hyperparameters to avoid a large number of parameter tuning. Specifically, the primary contributions and experiments are summarized below:

- To our best knowledge, we are the first to propose a novel task named DFARD to apply higher security level application scenes. Further, we theoretically demonstrate the challenges of this new task via the information bound.
- We optimize DFARD to improve the effectiveness of knowledge transfer and retain more data information. A plug-and-play ITA strategy and an AGB module are proposed to gain the simplest combination of generator losses, avoiding complex loss designs and weight balance, significantly reducing parameter tuning costs.
- Experiments show that our DFARD method achieves stable and reliable performance on multiple benchmarks comparing combinations of existing technologies.

Related Work

Data-Free Generation

Data-free generation is proposed to generate substitute data with Generative Adversarial Networks (GANs) or other generation modules. During this process, researchers do not need to access any data, thus being able to deal with data privacy and other data unavailable issues. Chen *et al.* (Chen et al. 2019) first introduce the generator into a data-free generation process to get more vital generation capabilities. To obtain the generated data that the student does not learn well, Micaelli *et al.* (Micaelli and Storkey 2019) introduce the method of adversarial generation. They prompt the generator to generate data with more significant differences between the student's and teacher's predictions so that the shortcomings are made up in the learning process. Choi *et al.* (Choi et al. 2020) add batch categorical entropy into the data-free generation process to promote class balance. To further improve the generation speed, Fang *et al.* (Fang et al. 2022) propose feature sharing to simplify the generation process of each step. To improve generation quality, Bhardwaj *et al.* (Bhardwaj, Suda, and Marculescu 2019) introduce model inversion and use the intermediate layer statistics of the teacher model to restore the original data. Based on this, Yin *et al.* (Yin et al. 2020) introduce adversarial inversion, and Fang *et al.* (Fang et al. 2021b) introduce contrastive learning to enhance the generation quality further.

Adversarial Robustness Distillation

Early adversarial training methods focus on learning directly from adversarial examples to improve model adversarial robustness (Madry et al. 2017; Zhang et al. 2019). However, expanding the adversarial training set leads to increased training costs. More importantly, the robustness improvement of small models is not evident due to the limitation of model capacity. Adversarial robustness distillation is proposed to address these issues. The setting is that both pre-trained robust teacher models and original training data are available. Goldblum *et al.* (Goldblum et al. 2020) first propose the concept of adversarial robustness distillation. They show that improving the robustness of small models is feasible without additional training costs. Zi *et al.* (Zi et al. 2021) find that the soft labels given by the teacher are very effective and can significantly improve the robustness performance of the student. Zhu *et al.* (Zhu et al. 2022) find that the teacher's confidence in the student's adversarial examples continues to decline, which may not be able to give the correct guidance. They propose a multi-stage strategy to allow the student to learn independently in later training. Zhao *et al.* (Zhao et al. 2022) utilize multiple teachers to learn from nature and robust scenes separately. Based on this, they try to focus on clean accuracy while improving adversarial robustness.

The Challenges of DFARD

To explore the impact of missing original training data on existing ARD tasks, we start with the effectiveness of knowledge transfer in the distillation process. By analyzing the lipschitzness of the robust model and the properties of the

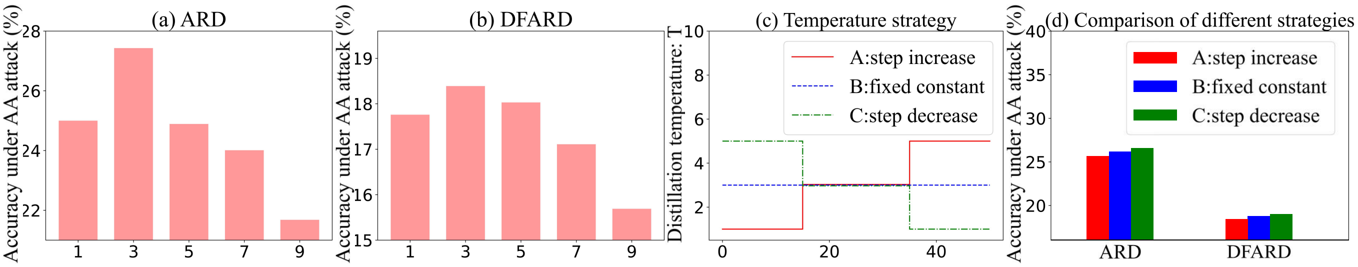

Figure 2: A toy experiment about the effect of different temperatures and simple verification of the easy-to-hard process. (a) and (b) show the impact of using different fixed temperatures for student performance on ARD and DFARD, respectively. (c) shows simple step temperature strategies, which means the trend of varying difficulty changes in the learning objective. (d) shows the performance comparison of students under these strategies.

generated data, we theoretically demonstrate why DFARD is more challenging than KD and ARD tasks. DFARD has a lower information upper limit than KD and ARD in the knowledge transfer process. This conclusion implies that for the DFARD task, more knowledge is needed. Based on this, we try to improve the efficiency of knowledge transfer and ensure higher data information to meet this challenge. Inspired by Human Education and Curriculum Learning (Bengio et al. 2009; Pentina, Sharmanska, and Lampert 2015), we try to look at the knowledge transfer process from the perspective of **1)** the knowledge from the teacher and **2)** the knowledge from the data. Detailed discussions are as follows:

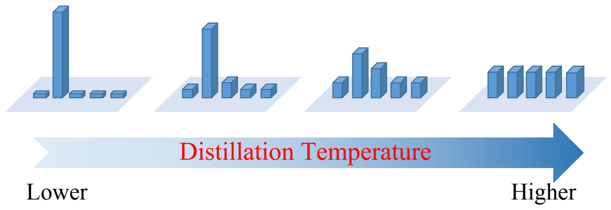

Figure 3: Diagrams of teacher's soft labels. As the temperature increases, the distance between the teacher's and naive student's prediction distributions decreases, and the difficulty of the learning objectives decreases.

The Knowledge from the Teacher. In the process of human education, teachers always teach students simple knowledge in the beginning. With improving students' abilities, more difficult knowledge is gradually covered. This easy-to-hard training process improves the efficiency of knowledge transfer. Inspired by this, we analogize the DFARD task to the complex challenge of the human learning process. The key lies in how to build an easy-to-hard process. In fact, distillation temperature enables the teacher network to provide suitable soft labels to transfer knowledge from the cumbersome model to a small model (Hinton et al. 2015; Romero et al. 2014). The temperature controls the discrepancy between two distributions and represents learning objectives of varying degrees of difficulty (Müller, Kornblith, and Hinton 2019; Li et al. 2022a; Zi et al. 2021; Li et al. 2022b) as shown in Figure 3. Most existing methods ignore the usefulness of the distillation temperature itself, regard it as a fixed hyperparameter, and inefficiently search for optimum. On this basis, they spend several times on computational costs.

To verify that the easy-to-hard process can or cannot improve knowledge transfer efficiency and better deal with DFARD tasks, we conduct a toy experiment as shown in Figure 2. We first verify the effect of different fixed distillation temperatures on two tasks. We train all student models for 50 epochs with or without original data and report the best robustness accuracy under AutoAttack (AA) attack (Croce and Hein 2020). From Figure 2(a) and (b), a general conclusion is that different temperatures have effects on the two tasks. Further, we respectively construct three temperature strategies of step increase, fixed constant, and step decrease to build learning objectives with different difficulties for each epoch (Figure 2(c)). The inflection points of temperature change are at the 15th and 35th epochs. Based on these strategies, we test the robustness performance as shown in Figure 2(d). We find that the strategy of step decrease achieves the best results. As shown in Figure 3, the decrease in temperature means the learning difficulty increases. That is, the easy-to-hard knowledge promotes the student's progress.

The Knowledge from the Data. For human education, apart from teachers, good tutorials are equally important. Generally speaking, tutorials that contain more knowledge can give students more help in the process of learning. More knowledge with more information is crucial. Inspired by this, we analogize the generated data as a medium of knowledge transfer to the tutorials. The generated data with different information content may also help students differently. Existing all data-free generation methods set a fixed generation loss weight to train a generator and constrain the teacher's confidence in the generated data (Chen et al. 2019; Yin et al. 2020; Choi et al. 2020; Fang et al. 2021b, 2022). To obtain generated data that is closer to the original distribution, the teacher model's predictions $f_t(\hat{\mathbf{x}})$ should be close to the one-hot labels $\mathbf{y}$, *e.g.*, minimize them with the following cross-entropy loss:

$$\mathcal{L}_{cls} = CE(f_t(\hat{\mathbf{x}}), \mathbf{y}), \qquad (1)$$

where $\mathbf{y}$ can be randomly generated labels or pseudo-labels of the teacher. $\hat{\mathbf{x}}$ is synthesized by the generator g through random noise $\mathbf{z}$ and the label $\mathbf{y}$: $\hat{\mathbf{x}} = g(\mathbf{z}, \mathbf{y})$.

In the above process, the teacher provides a more prominent target logit or less varied wrong logits. We argue that

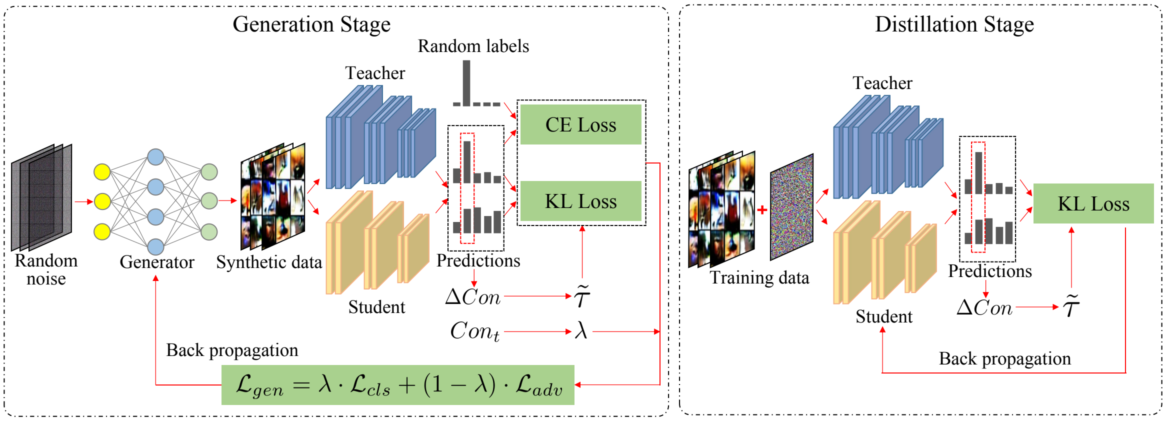

Figure 4: The pipeline of our optimized DFARD method from the most commonly used training baseline. Our method consists of two stages: (1) In the generation stage, we design an Interactive Temperature Adjustment strategy to adjust the temperature $\tilde{\tau}$ according to the student's learning. Simultaneously, we propose an Adaptive Generator Balance module to balance the similarity between data domains and the information content of data. (2) In the distillation stage, we keep the interactive temperature to help the student learn better. Training data represents the generated adversarial examples.

the above process gradually decreases the information content of the teacher's soft labels. The information content is a basic quantity derived from the probability prediction for the generated data of the teacher model. In this paper, we measure the information content by Information Entropy. Some studies have shown that such soft labels reduce information entropy and are not conducive to the knowledge distillation process (Shen et al. 2021; Zhang et al. 2022). We provide a theoretical analysis as follows.

Information Entropy. The entropy of a random variable is the average level of "information" or "uncertainty" (Shannon 1948) inherent to the variable's possible outcomes. Given a discrete random variable X, which takes values in the alphabet $\mathcal{X}$ and is distributed according to $p : \mathcal{X} \to [0, 1]$:

$$H(X) = -\sum_{x \in \mathcal{X}} p(x) \log p(x) = \mathbb{E}[-\log p(X)], \quad (2)$$

where $\sum$ denotes the sum over the variable's possible values. Based on the definition, in the existing process, the teacher's soft label will also change in the existing generation process as "information" decreases and "uncertainty" decreases. Although better distribution similarity between generated and original domains (Chen et al. 2019; Fang et al. 2021b), we think existing methods ignore the information content of data and the teacher's soft labels. A trade-off relationship rather than ignoring one of these two helps improve student performance (Tests and analyses are shown in Table 3).

Data-Free Adversarial Robustness Distillation

According to the above analysis, we clarify the motivation and perform simple analytical tests. In this section, we try to use the adaptive approach to improve the efficiency of knowledge transfer and ensure higher data information while reducing hyperparameter tuning costs. Firstly, we propose an Interactive Temperature Adjustment (ITA) strategy, which dynamically adjusts the distillation temperature according to the training status of the students in the current training epoch. The strategy helps students find the appropriate learning objectives for each training epoch. Secondly, we design an Adaptive Generator Balance (AGB) module to balance similarity and information capacity, avoiding the excessive pursuit of one. The pipeline is shown in Figure 4. The generator is trained via the ITA and AGB methods. Then the student is trained with the ITA strategy. Besides, the detailed training process is shown in Algorithm 1.

Interactive Temperature Adjustment

Our ITA strategy adjusts the teacher's soft label through the interactive distillation temperature $\tilde{\tau}$ so that the confidence gap between the teacher and student models is kept in a suitable range. In the generate stage, the generated data should transfer the information of decision boundary from the teacher model to the student model as effectively as possible (Heo et al. 2019). Unlike previous methods, our generator does not directly synthesize specific data but starts from accessible data for the student to learn. We maximize the predictions of the student and the teacher to find effective generated data via the adversarial generation loss as:

$$\mathcal{L}_{adv} = -KL(f_t(\hat{\mathbf{x}}; \theta_t), f_s(\hat{\mathbf{x}}; \theta_s), \tilde{\tau}), \quad (3)$$

where KL denotes the Kullback-Leibler (KL) divergence loss. Then we calculate the teacher's confidence for the generated data and collect the numerical value of the confidence Con_t and the class with the highest confidence c as $Con_t, c = \arg\max f_t(\hat{\mathbf{x}})$. The student's confidence for class c can be directly obtained as Con_s. The interactive temperature $\tilde{\tau}$ can be calculated as:

$$\tilde{\tau} = max \left\{ \frac{1}{bs} \sum_{i=1}^{bs} |Con_t - Con_s| \cdot C, 1 \right\}, \quad (4)$$

where C is total number of classes and bs denotes the batch size. The calculated absolute value represents the difference between the student's and the teacher's prediction in the current training epoch, thus reflecting the current learning situation. In the early epochs, higher distillation temperatures are set to obtain the generated data that is easier for students to learn. As the student's prediction gets closer to the teacher's, the temperature drops to synthesize more challenging data. Notably, ITA can be combined with other generation methods as a plug-and-play strategy.

Similarly, we consider the effectiveness of knowledge transfer in the distillation stage. As the progress of the students continues to increase the learning difficulty, we define the interactive knowledge distillation loss as:

$$\mathcal{L}_{KD} = \sum_{\hat{\mathbf{x}}' \in \hat{\mathbf{X}}'} KL(f_t(\hat{\mathbf{x}}'; \theta_t), f_s(\hat{\mathbf{x}}'; \theta_s), \tilde{\tau}), \tag{5}$$

where $\hat{\mathbf{x}}'$ is the adversarial examples of the generated data $\hat{\mathbf{x}}$.

Adaptive Generator Balance

For generator training objectives, we choose the common training losses (called Vanilla) to elaborate on our proposed optimization for simplicity and persuasiveness. The proposed AGB module can adjust the weight of the losses to balance the domain similarity and data information content according to the current confidence of the teacher (related to information entropy). We combine Equation (1) and (3) as:

$$\mathcal{L}_{gen} = \lambda \cdot \mathcal{L}_{cls} + (1 - \lambda) \cdot \mathcal{L}_{adv}, \tag{6}$$

where λ is the trade-off parameter. When the teacher's confidence is too high, the information content of the data may be ignored. At this time, λ adaptively reduces to avoid blindly pursuing similarity. Specifically, λ is calculated as:

$$\lambda = \frac{1}{C \cdot \frac{1}{bs} \sum_{i=1}^{bs} Con_t}. \tag{7}$$

The average confidence $\frac{1}{bs} \sum_{i=1}^{bs} Con_t$ is greater than or equal to the randomly expected value $\frac{1}{C}$ for the dataset with the number of classes C. Therefore, it always satisfies $0 < \lambda \leq 1$. With the help of AGB, we can increase the amount of information in the generated data while satisfying the similarity, which helps the student's performance. Simultaneously, we no longer need to try many different weight combinations to test results. Therefore, our method is more simple and more convenient.

Experiments

Experimental Setup

Dataset and Model. We evaluate the proposed DFARD method on 32×32 CIFAR datasets (Krizhevsky, Hinton et al. 2009), which are the most commonly used datasets for testing adversarial robustness. For a fair comparison, we use the same pre-trained WideResNet (WRN) teacher models with (Zi et al. 2021). Furthermore, we evaluate all methods using the student with ResNet-18 (RN-18) (He et al. 2016) and MobileNet-V2 (MN-V2) (Sandler et al. 2018) following existing ARD methods (Zhu et al. 2022; Zi et al. 2021).

Algorithm 1: Training process of our Data-Free Adversarial Robustness Distillation

```
Input: A pre-trained teacher network $f_t$, a generator $g$ with
    parameter $\theta_g$, a student $f_s$ with parameter $\theta_s$, distillation
    epochs $T$, the iterations of generator $g$ in each epoch $Tg$,
    the iterations of student $f_s$ in each epoch $Ts$.
Initialize parameter $\theta_g$ and $\theta_s$
for $i$ in $[1, \dots, T]$ do
    // Generation stage
    for $j$ in $[1, \dots, Tg]$ do
        Randomly sample noises and labels $(\mathbf{z}, \mathbf{y})$
        Synthesize training data $\hat{\mathbf{x}} = g(\mathbf{z}, \mathbf{y})$
        Update generator $g$ through Equation (6)
    end for
    // Distillation stage
    for $j$ in $[1, \dots, Ts]$ do
        Synthesize training data $\hat{\mathbf{x}} = g(\mathbf{z}, \mathbf{y})$
        Generate adversarial examples $\hat{\mathbf{x}} \rightarrow \hat{\mathbf{x}}'$
        Distill the student $f_s$ through Equation (5)
    end for
end for
Output: The student $f_s$ with adversarial robustness.
```

Baselines. We compare our optimized method with different data-free generation methods, including Dream (Bhardwaj, Suda, and Marculescu 2019), DeepInv (Yin et al. 2020), DAFL (Chen et al. 2019), ZSKT (Micaelli and Storkey 2019), DFQ (Choi et al. 2020), CMI (Fang et al. 2021b), and Fast (Fang et al. 2022). We use the same PGD-attack to generate the adversarial examples and the same distillation training loss as ARD (Goldblum et al. 2020) for all methods.

Implementation details. Our proposed method and all others are implemented in PyTorch. All models are trained on RTX 3090 GPUs (Paszke et al. 2019). The students are trained via SGD optimizer with cosine annealing learning rate with an initial value of 0.05, momentum of 0.9, and weight decay of 1e-4. The generators are trained via Adam optimizer with a learning rate of 1e-3, β_1 of 0.5, β_2 of 0.999. The distillation batch size and the synthesis batch size are both 256. The distillation epochs T is 200, the iterations of generator Tg is 1, and the iterations of student Ts is 5. Both the student model and the generator are randomly initialized. A 10-step PGD (PGD-10) with a random start size of 0.001 and step size of $2/255$ is used to generate adversarial samples. The perturbation bounds are set to L_∞ norm $\epsilon = 8/255$.

Attack Evaluation. We evaluate the adversarial robustness with: FGSM (Goodfellow, Shlens, and Szegedy 2014), $\mathrm{PGD_S}$ (Madry et al. 2017), $\mathrm{PGD_T}$ (Zhang et al. 2019), CW_∞ (Carlini and Wagner 2017) and AutoAttack (AA) (Croce and Hein 2020). The maximum perturbation is set as $\epsilon = 8/255$. The perturbation steps for $\mathrm{PGD_S}$, $\mathrm{PGD_T}$ and CW_∞ are 20. In addition, we test the accuracy of the models in normal conditions without adversarial attacks (Clean).

Comparison with Other Methods

To compare the effects of various data-free generation methods, we set the same distillation process as ARD (Goldblum

Model	Method	CIFAR-10						CIFAR-100					
		Attacks Evaluation						Attacks Evaluation					
		Clean	FGSM	PGD_S	PGD_T	CW	AA	Clean	FGSM	PGD_S	PGD_T	CW	AA
RN-18	Dream	**68.26**	34.76	29.72	31.36	27.96	26.70	22.00	10.18	9.52	9.85	7.11	6.68
	DeepInv	64.53	35.18	31.26	32.49	28.77	27.93	40.91	19.46	17.86	18.68	15.27	14.54
	DAFL	54.98	27.04	24.75	25.87	22.90	22.25	41.67	21.42	20.13	20.81	17.96	17.16
	ZSKT	58.08	31.98	29.94	30.92	27.21	26.68	38.91	20.16	18.78	19.41	16.38	15.52
	DFQ	54.44	26.90	24.63	25.78	22.37	21.57	45.24	22.49	20.78	21.61	18.24	17.38
	CMI	53.28	25.78	23.14	23.97	21.03	20.38	45.04	22.78	21.02	21.90	17.90	16.97
	Fast	61.13	31.40	28.01	29.17	26.26	25.42	36.75	18.66	17.72	18.33	15.57	14.77
	Ours*	65.10	36.36	33.47	34.89	30.79	30.06	45.33	24.08	22.71	23.38	19.84	19.00
	Ours	66.44	**38.53**	**35.94**	**37.15**	**32.79**	**32.14**	**46.33**	**24.56**	**22.94**	**23.59**	**20.12**	**19.19**
MN-V2	Dream	**64.95**	32.03	26.09	27.63	23.83	22.28	18.73	9.78	8.96	9.37	6.93	6.33
	DeepInv	59.53	31.76	28.42	29.74	25.86	24.99	37.75	16.94	15.54	16.19	12.65	11.80
	DAFL	47.53	24.51	21.18	22.09	19.50	18.86	40.46	20.63	19.03	19.78	16.54	15.82
	ZSKT	57.02	30.29	27.07	28.25	24.89	24.40	25.16	12.34	11.36	11.78	9.69	9.16
	DFQ	44.25	21.13	19.14	20.07	16.87	16.20	40.26	19.45	17.74	18.44	15.14	14.35
	CMI	44.53	21.34	19.67	19.97	16.25	15.97	40.23	19.76	17.96	18.56	14.86	14.02
	Fast	54.06	28.23	25.69	26.83	23.18	22.42	38.69	18.58	16.77	17.58	14.62	13.75
	Ours*	59.79	32.25	29.25	30.24	26.18	25.56	40.94	21.47	20.18	20.89	17.60	16.82
	Ours	61.16	**34.46**	**31.66**	**32.80**	**28.40**	**27.90**	**41.78**	**22.04**	**20.84**	**21.68**	**17.93**	**17.04**

Table 1: Adversarial robustness accuracy (%) on CIFAR-10 and CIFAR-100. The maximum adversarial perturbation ϵ is 8/255. "Ours*" means only deploying our method during the generation stage. "Ours" means the complete method as Algorithm 1.

Method	Dream	DeepInv	DAFL	ZSKT	DFQ	CMI	Fast	Ours
CIFAR-10	$29.17h * m$	$24.70h * nm$	$4.04h * nm$	$3.03h * n$	$4.09h * nm$	$48.69h * nm$	$6.10h * nm$	$3.28h$
CIFAR-100	$125.14h * m$	$101.28h * nm$	$13.43h * nm$	$5.62h * n$	$13.11h * nm$	$77.99h * nm$	$12.16h * nm$	$5.85h$

Table 2: The synthesis time of various data-free generation methods. We test the specific GPU time on a single RTX 3090 for the entire generation process. h is short for hours, n denotes the distillation temperature hyperparameter tuning times, and m denotes the generator loss weights tuning times.

et al. 2020). Therefore, the difference only lies in the generator loss function of these methods. We select and report the best checkpoint of all methods among all epochs. The best checkpoints are based on the adversarial robustness performance against PGD_T attack. For the computational costs, we compare the synthesis time on the generation stage of different generation methods.

Performance Comparison. The robustness performances of our and other baseline methods are shown in Table 1. Our generation method (**Ours***) achieves better adversarial robustness performance in all baselines. The results demonstrate that our interactive and adaptive approach can be more effective for the challenging DFARD task. For different backbone and dataset combinations, our method improves the average adversarial robustness by 1.98%, 0.55%, 1.69%, and 1.03%, respectively, compared to other best results.

Notably, our method maintains the most stable performance in various settings, while others may perform poorly in some settings. We consider that one reason is our interactive learning objective, which helps to improve students' versatility in different settings. Specifically, Dream (Bhardwaj, Suda, and Marculescu 2019) inverts enough data for normal ARD. However, these data might not be suitable for student learning as the data comes exclusively from teachers. DeepInv (Yin et al. 2020) and CMI (Fang et al. 2021b) excessively pursue distribution similarity between generated and original domains ignoring the information content of data. Fast (Fang et al. 2022) uses a feature-sharing method, but the lack of rich new features in complex datasets leads to performance degradation. In contrast, some early methods (DAFL (Chen et al. 2019), ZSKT (Micaelli and Storkey 2019) and DFQ (Choi et al. 2020)) are more stable and effective, but these methods keep the same teacher predictions throughout the generation process. Therefore, their learning objectives may not meet every epoch for the randomly initialized students. Good results are often inseparable from multiple hyperparameter tuning.

Generation Computational Costs. In addition, we also compare the overall generation computational costs of all methods while considering the hyperparameter tuning calculation costs. The results are shown in Table 2. Other methods must test multiple sets of temperature parameters (denoted by n) or trade-offs between multiple generator losses (denoted by m). Notably, some methods require significantly higher weights tuning times m when using four or more generator losses, *e.g.*, Dream, DeepInv, DFQ, and CMI. Thanks to fully adaptive parameters, our generator's computational costs are significantly lower than most other methods. In summary, our method has the most stable performance while main-

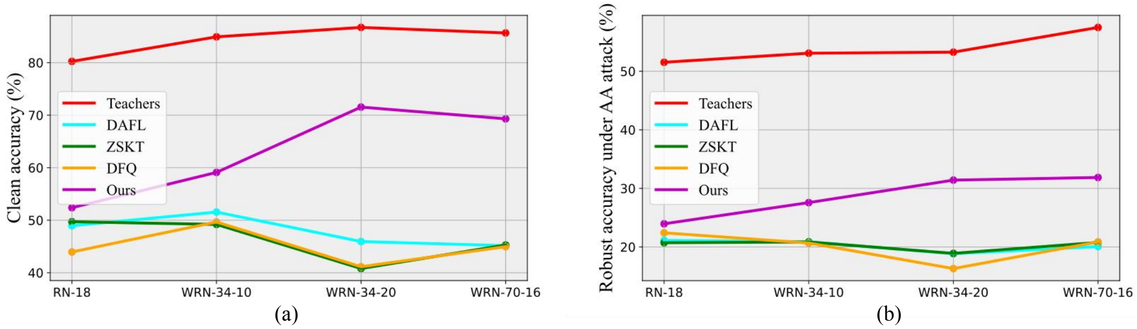

Figure 5: Performance of RN-18 students trained with different teachers. Students train 100 epochs for 4 generative methods. (a) shows clean accuracy, and (b) shows robust accuracy under AA attack.

ID	Settings	Attacks Evaluation					
		Clean	FGSM	PGD$_\mathrm{S}$	PGD$_\mathrm{T}$	CW	AA
(1)	Vanilla	58.35	30.68	28.59	29.75	24.67	23.95
(2)	w/ ITA (G)	59.16	31.05	28.59	29.90	24.87	24.21
(3)	w/ ITA (GD)	60.36	33.09	30.97	32.02	26.79	26.24
(4)	w/ AGB	64.41	36.33	33.53	34.91	30.57	29.92
(5)	**Ours**	**66.44**	**38.53**	**35.94**	**37.15**	**32.79**	**32.14**

Table 3: Ablation study on CIFAR-10. For 'Vanilla', we choose the best hyperparameters (the distillation temperature $\tau = 3$, the loss weight $\lambda = 0.3$). 'G' means that ITA is applied only in the generation stage, and 'GD' means that ITA is applied in both the generation and distillation stages.

taining the advantage of significantly lower generation cost. Therefore, our method is simple, reliable, and convenient.

Adaptability for Different Teachers

The method's adaptability is important to reduce reliance on customized teachers (Zi et al. 2021). To compare the adaptability for different teachers, we select four teachers (RN-18, WRN-34-10, WRN-34-20, WRN-70-16 (Croce et al. 2021)) to train the RN-18 student on CIFAR-10. The results are shown in Figure 5. For the other methods, we find the robust saturation phenomenon. Due to the capacity gap between the teacher and student, the learning objectives provided by larger-scale teachers may not be suitable for small students to learn. However, our method is less susceptible to the gap due to the interactive learning objectives. Our proposed easy-to-hard process alleviates the saturation and has more vital adaptability for teachers with different capacities.

Ablation Study

Impact of Interactive Temperature Adjustment. To thoroughly verify the effectiveness of the proposed Interactive Temperature Adjustment (ITA) strategy, we test it in both the generation and distillation stages. As shown in Table 3(1-3) and (4-5), compared with the best-fixed temperature parameters (w/o ITA), students' performance improves when ITA is applied in the generation stage. Further, when the interactive temperature is deployed in the both generation and distillation stages, the performance improves again. It is worth noting that Table 3 does not reflect hyperparameter tuning time. The best temperature comes from multiple experimental tests. Even so, the student's feedback can dynamically adjust the difficulty of knowledge transfer to improve student performance, which verifies the effectiveness of ITA.

Method	Clean	CW	AA
DAFL + ITA	56.26 (+1.28)	24.78 (+1.88)	24.36 (+2.11)
DFQ + ITA	55.79 (+1.35)	24.12 (+1.75)	23.74 (+2.17)
ZSKT + ITA	59.93 (+1.85)	29.03 (+1.82)	28.58 (+1.90)

Table 4: Other methods with our proposed ITA.

Other Methods with the ITA Strategy. Further, we combine the proposed plug-and-play Interactive Temperature Adjustment (ITA) strategy with three other methods for both the generation and distillation stages. Experiments are carried out on CIFAR-10 with the RN-18 student. The results are shown in Table 4. Compared with the baseline performance in Table 1, for the three methods, both clean and robust accuracy are significantly improved, which proves that our ITA strategy promotes student performance through an easy-to-hard knowledge transfer process.

Impact of Adaptive Generator Balance. Further, we evaluate the effectiveness of the proposed AGB module. The results are shown in Table 3 (4-6). Compared with the best-fixed λ, our AGB module significantly improves student performance for both clean and robust accuracy. At the same time, our adaptive approach omits the weight-tuning costs.

Conclusion

This paper proposes a novel task named Data-Free Adversarial Robustness Distillation (DFARD) to deal with realistic scenes with higher security levels. We demonstrate that the task is more challenging than existing tasks, making combining previous methods less effective. Due to less information available, we try to optimize the new task by improving knowledge transfer efficiency and maintaining higher data information. Our method is simple yet effective, achieving stable performance while maintaining the advantage of significantly lower generation cost. We believe the proposed technique helps apply deep learning techniques to real scenes.

Acknowledgments

This work is supported by Shanghai Municipal Science and Technology Major Project (No.2021SHZDZX0103), the Shanghai Engineering Research Center of AI & Robotics, Fudan University, China, and the Engineering Research Center of AI & Robotics, Ministry of Education, China.

References

Athalye, A.; Carlini, N.; and Wagner, D. 2018. Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples. In *International conference on machine learning*, 274–283. PMLR.

Bengio, Y.; Louradour, J.; Collobert, R.; and Weston, J. 2009. Curriculum learning. In *Proceedings of the 26th annual international conference on machine learning*, 41–48.

Bhardwaj, K.; Suda, N.; and Marculescu, R. 2019. Dream distillation: A data-independent model compression framework. *arXiv preprint arXiv:1905.07072*.

Carlini, N.; and Wagner, D. 2017. Towards evaluating the robustness of neural networks. In *2017 ieee symposium on security and privacy (sp)*, 39–57. Ieee.

Chen, H.; Wang, Y.; Xu, C.; Yang, Z.; Liu, C.; Shi, B.; Xu, C.; Xu, C.; and Tian, Q. 2019. Data-free learning of student networks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 3514–3522.

Chen, Z.; Li, B.; Wu, S.; Jiang, K.; Ding, S.; and Zhang, W. 2023. Content-based Unrestricted Adversarial Attack. *arXiv preprint arXiv:2305.10665*.

Chen, Z.; Li, B.; Wu, S.; Xu, J.; Ding, S.; and Zhang, W. 2022a. Shape matters: deformable patch attack. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part IV*, 529–548. Springer.

Chen, Z.; Li, B.; Xu, J.; Wu, S.; Ding, S.; and Zhang, W. 2022b. Towards practical certifiable patch defense with vision transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 15148–15158.

Choi, Y.; Choi, J.; El-Khamy, M.; and Lee, J. 2020. Data-free network quantization with adversarial knowledge distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 710–711.

Croce, F.; Andriushchenko, M.; Sehwag, V.; Debenedetti, E.; Flammarion, N.; Chiang, M.; Mittal, P.; and Hein, M. 2021. RobustBench: a standardized adversarial robustness benchmark. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

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. PMLR.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Fang, G.; Bao, Y.; Song, J.; Wang, X.; Xie, D.; Shen, C.; and Song, M. 2021a. Mosaicking to distill: Knowledge distillation from out-of-domain data. *Advances in Neural Information Processing Systems*, 34: 11920–11932.

Fang, G.; Mo, K.; Wang, X.; Song, J.; Bei, S.; Zhang, H.; and Song, M. 2022. Up to 100x Faster Data-Free Knowledge Distillation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, 6597–6604.

Fang, G.; Song, J.; Wang, X.; Shen, C.; Wang, X.; and Song, M. 2021b. Contrastive model inversion for data-free knowledge distillation. *arXiv preprint arXiv:2105.08584*.

Goldblum, M.; Fowl, L.; Feizi, S.; and Goldstein, T. 2020. Adversarially robust distillation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, 3996–4003.

Goodfellow, I. J.; Shlens, J.; and Szegedy, C. 2014. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.

Heo, B.; Lee, M.; Yun, S.; and Choi, J. Y. 2019. Knowledge distillation with adversarial samples supporting decision boundary. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, 3771–3778.

Hinton, G.; Vinyals, O.; Dean, J.; et al. 2015. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2(7).

Jia, X.; Wei, X.; Cao, X.; and Foroosh, H. 2019. Comdefend: An efficient image compression model to defend adversarial examples. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 6084–6092.

Krizhevsky, A.; Hinton, G.; et al. 2009. Learning multiple layers of features from tiny images.

Li, X.-C.; Fan, W.-s.; Song, S.; Li, Y.; Zhan, D.-C.; et al. 2022a. Asymmetric Temperature Scaling Makes Larger Networks Teach Well Again. *Advances in neural information processing systems*.

Li, Z.; Li, X.; Yang, L.; Zhao, B.; Song, R.; Luo, L.; Li, J.; and Yang, J. 2022b. Curriculum Temperature for Knowledge Distillation. *arXiv preprint arXiv:2211.16231*.

Liu, S.; Chen, Z.; Liu, Y.; Wang, Y.; Yang, D.; Zhao, Z.; Zhou, Z.; Yi, X.; Li, W.; Zhang, W.; et al. 2023a. Improving generalization in visual reinforcement learning via conflict-aware gradient agreement augmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 23436–23446.

Liu, Y.; Liu, J.; Yang, K.; Ju, B.; Liu, S.; Wang, Y.; Yang, D.; Sun, P.; and Song, L. 2023b. Amp-net: Appearance-motion prototype network assisted automatic video anomaly detection system. *IEEE Transactions on Industrial Informatics*.

Liu, Y.; Xia, Z.; Zhao, M.; Wei, D.; Wang, Y.; Liu, S.; Ju, B.; Fang, G.; Liu, J.; and Song, L. 2023c. Learning causality-inspired representation consistency for video anomaly detection. In *Proceedings of the 31st ACM International Conference on Multimedia*, 203–212.

Madry, A.; Makelov, A.; Schmidt, L.; Tsipras, D.; and Vladu, A. 2017. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*.

Micaelli, P.; and Storkey, A. J. 2019. Zero-shot knowledge transfer via adversarial belief matching. *Advances in Neural Information Processing Systems*, 32.

Müller, R.; Kornblith, S.; and Hinton, G. E. 2019. When does label smoothing help? *Advances in neural information processing systems*, 32.

Paszke, A.; Gross, S.; Massa, F.; Lerer, A.; Bradbury, J.; Chanan, G.; Killeen, T.; Lin, Z.; Gimelshein, N.; Antiga, L.; et al. 2019. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32.

Pentina, A.; Sharmanska, V.; and Lampert, C. H. 2015. Curriculum learning of multiple tasks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 5492–5500.

Ramesh, A.; Dhariwal, P.; Nichol, A.; Chu, C.; and Chen, M. 2022. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*.

Romero, A.; Ballas, N.; Kahou, S. E.; Chassang, A.; Gatta, C.; and Bengio, Y. 2014. Fitnets: Hints for thin deep nets. *arXiv preprint arXiv:1412.6550*.

Sandler, M.; Howard, A.; Zhu, M.; Zhmoginov, A.; and Chen, L.-C. 2018. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 4510–4520.

Shannon, C. E. 1948. A mathematical theory of communication. *The Bell system technical journal*, 27(3): 379–423.

Shen, Z.; Liu, Z.; Xu, D.; Chen, Z.; Cheng, K.-T.; and Savvides, M. 2021. Is label smoothing truly incompatible with knowledge distillation: An empirical study. *arXiv preprint arXiv:2104.00676*.

Wang, Y.; Chen, Z.; Yang, D.; Liu, Y.; Liu, S.; Zhang, W.; and Qi, L. 2023a. Adversarial Contrastive Distillation with Adaptive Denoising. *arXiv preprint arXiv:2302.08764*.

Wang, Y.; Chen, Z.; Zhang, J.; Yang, D.; Ge, Z.; Liu, Y.; Liu, S.; Sun, Y.; Zhang, W.; and Qi, L. 2023b. Sampling to distill: Knowledge transfer from open-world data. *arXiv preprint arXiv:2307.16601*.

Wang, Y.; Ge, Z.; Chen, Z.; Liu, X.; Ma, C.; Sun, Y.; and Qi, L. 2023c. Explicit and implicit knowledge distillation via unlabeled data. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 1–5. IEEE.

Yang, D.; Chen, Z.; Wang, Y.; Wang, S.; Li, M.; Liu, S.; Zhao, X.; Huang, S.; Dong, Z.; Zhai, P.; et al. 2023a. Context De-confounded Emotion Recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 19005–19015.

Yang, D.; Huang, S.; Xu, Z.; Li, Z.; Wang, S.; Li, M.; Wang, Y.; Liu, Y.; Yang, K.; Chen, Z.; et al. 2023b. Aide: A vision-driven multi-view, multi-modal, multi-tasking dataset for assistive driving perception. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 20459–20470.

Yang, D.; Liu, Y.; Huang, C.; Li, M.; Zhao, X.; Wang, Y.; Yang, K.; Wang, Y.; Zhai, P.; and Zhang, L. 2023c. Target and source modality co-reinforcement for emotion understanding from asynchronous multimodal sequences. *Knowledge-Based Systems*, 110370.

Yang, D.; Yang, K.; Wang, Y.; Liu, J.; Xu, Z.; Yin, R.; Zhai, P.; and Zhang, L. 2023d. How2comm: Communication-Efficient and Collaboration-Pragmatic Multi-Agent Perception. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Yang, J.; Zhou, K.; Li, Y.; and Liu, Z. 2021. Generalized out-of-distribution detection: A survey. *arXiv preprint arXiv:2110.11334*.

Yin, H.; Molchanov, P.; Alvarez, J. M.; Li, Z.; Mallya, A.; Hoiem, D.; Jha, N. K.; and Kautz, J. 2020. Dreaming to distill: Data-free knowledge transfer via deepinversion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8715–8724.

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. PMLR.

Zhang, Q.; Cheng, X.; Chen, Y.; and Rao, Z. 2022. Quantifying the knowledge in a DNN to explain knowledge distillation for classification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Zhao, S.; Yu, J.; Sun, Z.; Zhang, B.; and Wei, X. 2022. Enhanced Accuracy and Robustness via Multi-teacher Adversarial Distillation. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part IV*, 585–602.

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 *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 16443–16452.