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

Yuzheng Wang†∗, Zhaoyu Chen†∗, Dingkang Yang†, Pinxue Guo†
Kaixun Jiang†, Wenqiang Zhang†, Lizhe Qi††

1Shanghai Engineering Research Center of AI & Robotics, Academy for Engineering & Technology, Fudan University
2Engineering 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.
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
we try to improve the efficiency of knowledge transfer and endowment of the Knowledge from the Teacher.

(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.

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.

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 $f_t(\hat{x})$ should be close to the one-hot labels $y$, e.g., minimize them with the following cross-entropy loss:

$$L_{cls} = CE(f_t(\hat{x}), y),$$

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

In the above process, the teacher provides a more prominent target logit or less varied wrong logits. We argue that
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 $p : \mathcal{X} \to [0, 1]$:

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

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:

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

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 $\text{Con}_t$ and the class with the highest confidence $c$ as $\text{Con}_t, c = \arg \max_i f_t(x)$. The student’s confidence for class $c$ can be directly obtained as $\text{Con}_s$. The interactive temperature $\tilde{\tau}$ can be calculated as:

$$\tilde{\tau} = \max \left\{ \frac{1}{bs} \sum_{i=1}^{bs} \text{Con}_t - \text{Con}_s: C, 1 \right\},$$

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:

$$L_{KD} = \sum_{\tilde{x} \in \mathcal{X}} KL(f_t(\tilde{x}; \theta_t), f_s(\tilde{x}; \theta_s; \tilde{\tau}),$$

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

**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:

$$L_{gen} = \lambda \cdot L_{ids} + (1 - \lambda) \cdot L_{adv},$$

where $\lambda$ 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, $\lambda$ adaptively reduces to avoid blindly pursuing similarity. Specifically, $\lambda$ is calculated as:

$$\lambda = \frac{1}{C \cdot \frac{1}{T_g} \sum_{j=1}^{T_g} C_{o_t}}.$$

The average confidence $\frac{1}{T_g} \sum_{j=1}^{T_g} C_{o_t}$ is greater than or equal to the randomly expected value $\frac{1}{T_g}$ 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).

**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 FastDAFL (Chen et al. 2019), ZSKT (Micaelli and Storkey 2019), CW (Zhang et al. 2019), CW$_\infty$ (Carlini and Wagner 2017) and AutoAttack (AA) (Croce and Hein 2020). The maximum perturbation is set as $\epsilon = 8/255$. The perturbation steps for PGD$_\delta$, PGD$_T$, and CW$_\infty$ 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...
Table 1: Adversarial robustness accuracy (%) on CIFAR-10 and CIFAR-100. The maximum adversarial perturbation $\epsilon$ is 8/255. “Ours*” means only deploying our method during the generation stage. “Ours” means the complete method as Algorithm 1.

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.

Method 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.

<table>
<thead>
<tr>
<th>ID</th>
<th>Settings</th>
<th>Clean</th>
<th>FGSM</th>
<th>PGD₈</th>
<th>PGD₉</th>
<th>CW</th>
<th>AA</th>
</tr>
</thead>
<tbody>
<tr>
<td>(1)</td>
<td>Vanilla</td>
<td>58.35</td>
<td>30.68</td>
<td>28.59</td>
<td>29.75</td>
<td>24.67</td>
<td>23.95</td>
</tr>
<tr>
<td>(2)</td>
<td>w/ ITA (G)</td>
<td>59.16</td>
<td>31.05</td>
<td>28.59</td>
<td>29.90</td>
<td>24.87</td>
<td>24.21</td>
</tr>
<tr>
<td>(3)</td>
<td>w/ ITA (GD)</td>
<td>60.36</td>
<td>33.09</td>
<td>30.97</td>
<td>32.02</td>
<td>26.79</td>
<td>26.24</td>
</tr>
<tr>
<td>(4)</td>
<td>w/ AGB</td>
<td>64.41</td>
<td>36.33</td>
<td>33.53</td>
<td>34.91</td>
<td>30.57</td>
<td>29.92</td>
</tr>
<tr>
<td>(5)</td>
<td>Ours</td>
<td>66.44</td>
<td>38.53</td>
<td>35.94</td>
<td>37.15</td>
<td>32.79</td>
<td>32.14</td>
</tr>
</tbody>
</table>

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.

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.

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 $\lambda$, 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.

Table 4: Other methods with our proposed ITA.

<table>
<thead>
<tr>
<th>Method</th>
<th>Clean</th>
<th>CW</th>
<th>AA</th>
</tr>
</thead>
<tbody>
<tr>
<td>DAFL + ITA</td>
<td>56.26 (+1.28)</td>
<td>24.78 (+1.88)</td>
<td>24.36 (+2.11)</td>
</tr>
<tr>
<td>DFQ + ITA</td>
<td>55.79 (+1.35)</td>
<td>24.12 (+1.75)</td>
<td>23.74 (+2.17)</td>
</tr>
<tr>
<td>ZSKT + ITA</td>
<td>59.93 (+1.85)</td>
<td>29.03 (+1.82)</td>
<td>28.58 (+1.90)</td>
</tr>
</tbody>
</table>

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


5783
Adaptive Denoising...


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.