

CIARD: Cyclic Iterative Adversarial Robustness Distillation

Liming Lu¹ Shuchao Pang^{1*} Xu Zheng^{2,3*} Xiang Gu¹
Anan Du⁴ Yunhuai Liu⁵ Yongbin Zhou¹

¹Nanjing University of Science and Technology,

²HKUST(GZ), ³INSAIT, Sofia University, St. Kliment Ohridski,

⁴Nanjing University of Industry Technology, ⁵Peking University

{luliming, pangshuchao, eminentguxiang, zhoyongbin}@njjust.edu.cn

zhengxu128@gmail.com, anan.du@niit.edu.cn, yunhuai.liu@pku.edu.cn

Abstract

Adversarial robustness distillation (ARD) aims to transfer both performance and robustness from teacher model to lightweight student model, enabling resilient performance on resource-constrained scenarios. Though existing ARD approaches enhance student model’s robustness, the inevitable by-product leads to the degraded performance on clean examples. We summarize the causes of this problem inherent in existing methods with dual-teacher framework as: ① The divergent optimization objectives of dual-teacher models, i.e., the clean and robust teachers, impede effective knowledge transfer to the student model, and ② The iteratively generated adversarial examples during training lead to performance deterioration of the robust teacher model. To address these challenges, we propose a novel Cyclic Iterative ARD (CIARD) method with two key innovations: ① A multi-teacher framework with contrastive push-loss alignment to resolve conflicts in dual-teacher optimization objectives, and ② Continuous adversarial retraining to maintain dynamic teacher robustness against performance degradation from the varying adversarial examples. Extensive experiments on CIFAR-10, CIFAR-100, and Tiny-ImageNet demonstrate that CIARD achieves remarkable performance with an average 3.53% improvement in adversarial defense rates across various attack scenarios and a 5.87% increase in clean sample accuracy, establishing a new benchmark for balancing model robustness and generalization. Our code is available at <https://github.com/eminentgu/CIARD>.

1. Introduction

In the era of edge computing and real-time applica-

*Corresponding author.

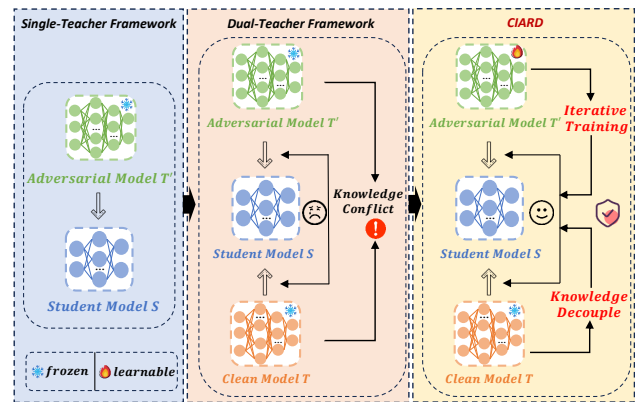


Figure 1. Comparison of single-teacher, dual-teacher and CIARD (ours) distillation frameworks.

tions, the deployment of efficient and robust models on devices with limited resources presents a significant challenge. Knowledge distillation (KD) [12, 20, 35, 42] has become a popular method for compressing large teacher models into smaller, more efficient student models with minimal accuracy loss. However, adversarial attacks [3, 19, 22, 23, 25, 28] pose a significant threat to the deployment of these student models, particularly in edge environments. Therefore, enhancing the robustness of lightweight student models against adversarial attacks is crucial for their application in real-world scenarios such as autonomous driving, image classification, and speech recognition [7, 11, 31].

To address these security concerns, adversarial training (AT) [13, 15–17, 21] has emerged as a promising defense approach. AT enhances model robustness [2, 6, 30, 34] by incorporating both clean and adversarial examples during the training process. However, when applied to knowledge distillation, AT presents two significant challenges including: ① Increasing computational burden for generating adversarial attacks contradicts the model compression effi-

ciency, and ② The robustness-accuracy trade-off problem that models typically sacrifice clean performance for better robustness.

Recently, adversarial robustness distillation (ARD) [8, 14, 36, 38, 44, 45] has emerged as a promising solution. In the context of ARD, the dual-teacher architecture [39] offers powerful solution to the two aforementioned challenges by simultaneously improving both clean sample accuracy and adversarial robustness when guiding the student models. However, two critical challenges remain in the current dual-teacher ARD frameworks including: ① **Conflicting Optimization Objectives**: In conventional dual-teacher distillation frameworks, the two teachers serve distinct purposes where one teacher focuses on clean sample accuracy while the other emphasizes adversarial robustness. This dichotomy often leads to suboptimal knowledge transfer as the student model struggles to reconcile these competing objectives. As shown in Table 1, current methods, such as B-MTARD [41], improve robustness ($\uparrow 0.95\%$) at the cost of a decrease in clean accuracy ($\downarrow 0.17\%$). This cost between clean accuracy and robust accuracy clearly validates the issue, and ② **Degradation of Adversarial Teacher Performance**: Through empirical observations (Fig. 3), we find that as the student model evolves during training, the generated adversarial examples increasingly compromise the performance of the robust teacher model. This degradation significantly impacts the quality of knowledge transfer and the overall robustness of the student model.

To address these limitations of prior dual-teacher ARD-methods [39, 41], we propose a novel cyclic iterative dual-teacher distillation framework, namely Cyclic Iterative Adversarial Robustness Distillation (CIARD), as shown in Figure 1. The CIARD framework resolves the distinct training optimization objectives issue by introducing contrastive push loss alignment. The contrastive push loss effectively decouples clean knowledge from robust knowledge and ensure that student model specializes in learning robust features without interference from clean teacher. Practically, the push loss works by deliberately creating a divergence between the student and clean teacher. By pushing away from the clean teacher’s incorrect predictions, the student effectively absorbs the robust teacher’s specialized knowledge.

After achieving the goal of precisely guide the student’s learning trajectory, we further incorporate an Iterative Teacher Training (ITT) strategy to avoid the performance degradation of the teacher model caused by the iteratively generated adversarial examples during training. In detail, ITT first freezes both teacher models’ parameters at the early training stage to let student have a basic knowledge. Afterwards, ITT iteratively updates the robustness teacher with our proposed continuous adversarial retraining to maintain dynamic teacher robustness against perfor-

Table 1. Performance comparison of MobileNet-V2 under FGSM in two knowledge distillation frameworks on CIFAR-10. Best performance metrics are highlighted in **bold**.

Type	Defense	FGSM [9]	
		Clean (%)	Robust (%)
Single-Teacher	ARD [8]	83.43	57.03
	ABSLD [40]	82.50 \downarrow	58.47 \uparrow
Dual-Teacher	MTARD [39]	89.26	57.84
	B-MTARD [41]	89.09 \downarrow	58.79 \uparrow
	CIARD (Ours)	89.51 \uparrow	59.10 \uparrow

mance degradation from the varying adversarial examples.

Extensive experiments across multiple datasets demonstrate that CIARD significantly outperforms existing ARD methods, achieving substantial improvements in both adversarial robustness and clean sample accuracy. Overall, our contributions are summarized as follows: **(I)** We propose CIARD, a cyclic iterative dual-teacher distillation framework that incorporates a contrastive push loss alignment and an Iterative Teacher Training strategy to effectively guide student model training while maintaining teacher model robustness. **(II)** We introduce a novel optimization strategy, namely Contrastive Push Loss, that resolves conflicting objectives among dual-teacher models, facilitating enhanced knowledge transfer while dynamically adjusting the training process to achieve an optimal balance between robustness and accuracy. **(III)** We introduce the Iterative Teacher Training strategy to dynamically enhance the knowledge supplementation of the adversarial teacher based on the evaluation of the student model. This approach equips the robust teacher a strong capability in both robust and clean knowledge, effectively boosting the student’s performance. **(IV)** Extensive experiments across multiple datasets demonstrate that CIARD significantly outperforms existing ARD methods, achieving substantial improvements in both adversarial robustness and clean sample accuracy.

2. Preliminary & Background

Knowledge Distillation (KD). KD is a widely used method for compressing deep neural networks, aiming to transfer knowledge from a large teacher model to a smaller student model. Given a teacher model $T(\cdot)$ and an in-distribution example $x \sim D$ (from the same distribution as the teacher’s training data), the traditional KD optimization objective is to obtain student model parameters θ that minimize the Kullback-Leibler (KL) divergence loss $\mathcal{L}_{KL}(\cdot)$ between the softmax logits of the teacher and student outputs. The loss function is defined as follows:

$$\arg \min_{\theta_S} (\underbrace{\alpha \text{CE}(S(x), y)}_{\text{Hard Label Loss}} + \underbrace{(\beta \tau^2 \text{KL}(S^\tau(x), T^\tau(x)))}_{\text{Soft Label Loss}}). \quad (1)$$

As seen from the distillation Eq. 1, the loss function typically consists of two components: one is the cross-entropy

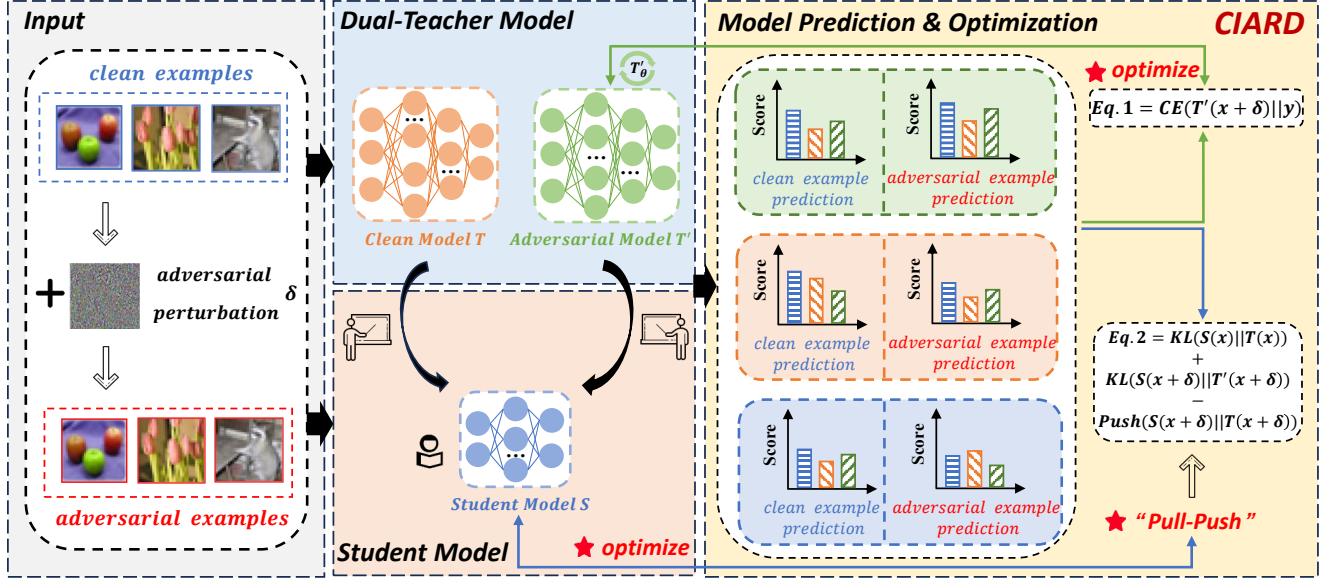


Figure 2. The framework of Cyclic Iterative Adversarial Robustness Distillation (CIARD). Our dual-teacher distillation framework features a continuously updated robust teacher model and a push loss mechanism to guide knowledge transfer, enabling effective balance between adversarial robustness and accuracy in the student model.

loss between the student model and the true label, and the other is the KL divergence between the student model and the teacher model’s soft labels. However, traditional KD focuses solely on the student model’s accuracy on clean examples, without considering its robustness.

Adversarial Training (AT). AT remains one of the most effective methods for defending against adversarial attacks. Its core principle involves training models with both adversarial and clean examples, enabling defense capabilities through exposure to hostile inputs. The key to adversarial training lies in generating perturbations through inner maximization. Extensive research has focused on enhancing the robustness of deep neural networks through AT, by proposing various strategies. For example, Zi et al. [45] employ robust soft labels to improve student model robustness, Zhang et al. [38] introduce TRADES to balance robustness and performance on clean examples, and Wang et al. [32] address misclassified examples through MART. Main factors contributing to robustness include larger models, more data, and the use of KL divergence for the inner maximization. Wu et al.’s [33] study confirm this perspective, demonstrating that AT significantly improves robustness in large models, though the improvements were less substantial for smaller models.

Adversarial Robustness Distillation (ARD). KD does not provide the student model with sufficient robustness, while AT demands a large model capacity. Intuitively, researchers have extensively explored ARD, which enhances student model robustness by combining KD and AT. Revisiting existing ARD methods [24, 43], Robust Soft Label Adversarial Distillation improves robustness by using robust soft la-

bels in inner optimization, highlighting their importance in AT. While most methods focus on prediction output, recent research has explored matching feature layers [27, 29] and input gradients [5] to generate more robust student models. In this paper, we conduct an in-depth analysis of mainstream research achievements related to ARD. RSLAD [45] improves model robustness by introducing robust soft labels, although its accuracy on clean examples still lags behind traditional training. Multi-Teacher Adversarial Robust Distillation (MTARD) [39] leverages a dual-teacher framework and adaptive normalized loss function to achieve a better trade-off between student model’s robustness and accuracy, yet there remains room for optimization to achieve more stable performance balance. Inspired by these studies, we propose a more effective ARD method to enhance the robustness of student models.

3. Methodology

3.1. An Adversarial Training Perspective on KD

As shown in Figure 1, the traditional KD process focuses on enabling the student model to inherit the teacher model’s accuracy without addressing adversarial robustness. However, traditional KD only allows the student model to inherit the accuracy of natural examples, while its robustness against adversarial examples is significantly lower than that of the teacher model. Thus ARD redefines the distillation objective from an adversarial perspective, as follows:

$$\mathbb{E}_{p_d(x)} \left[\underbrace{\alpha \mathcal{L}_{CE}(S(x), y)}_{\text{Clean Example}} + \underbrace{\beta \tau^2 \mathcal{L}_{KL}(S^\tau(x^*), T^\tau(x))}_{\text{Adversarial Example}} \right], \quad (2)$$

where x^* denotes the search result of the inner optimization, which can be expressed in the following form:

$$x^* = x + \arg \max_{\|\delta\|_p \leq \epsilon} \mathcal{L}_{CE}(S(x + \delta), y), \quad (3)$$

this constraint ($\|\delta\|_p \leq \epsilon$) ensures that the perturbation δ is within a specified bound ϵ under the L_p -norm. The L_p -norm measures the magnitude of the perturbation, and ϵ is the maximum allowable perturbation size.

With advancements in ARD research, Zhao et al. [39] design a dynamic training method that can balance the influence of adversarial and non-adversarial teacher models on student models. MTARD extends prior work by incorporating multi-teacher adversarial robustness distillation to guide the adversarial training of lightweight models. The basic min-max optimization framework of MTARD is defined in the following form:

$$\arg \min_{\theta_S} \left[\underbrace{\alpha \mathcal{L}_{KL}(S(x), T_{nat}(x))}_{\text{Nature Example}} + \underbrace{\beta \mathcal{L}_{KL}(S(x^*), T_{adv}(x^*))}_{\text{Adversarial Example}} \right], \quad (4)$$

where x^* is an adversarial example generated from the clean example x as shown below:

$$x^* = \arg \max_{\delta \in \Omega} \text{CE}(S(x + \delta; \theta_S), y), \quad (5)$$

it's worth noting that α and β in the above formulas are weighting factors that sum to 1.

3.2. Framework Overview

In this section, we give framework details of our proposed CIARD, which simultaneously enhances the robustness of student models while preserving their high accuracy on clean examples. As in Figure 2, CIARD achieves high clean accuracy and adversarial robustness by coordinating knowledge transfer between dual teacher models. Given a batch of clean examples $x \in D_{train}$, these inputs first pass through the clean teacher model t to generate softened class probabilities $t(x)$, serving as reference targets for maintaining natural pattern recognition capabilities. Simultaneously, each clean sample undergoes adversarial perturbation through a Projected Gradient Descent (PGD) [23] attack that jointly considers both the student model s and the clean teacher model t :

$$x^* = \text{PGD}_\epsilon(x, s, t) = \arg \max_{\|\delta\|_\infty \leq \epsilon} [\text{KL}(s(x + \delta) \| t(x + \delta))]. \quad (6)$$

This collaborative attack generation strategy produces adversarial examples x^* that challenge both the learning student and the vulnerable clean teacher, ensuring exposure to evolving attack patterns throughout the training process.

At the same time, for each input pair (x, x^*) , the clean teacher processes both the benign examples and adversarial examples simultaneously, generating the corresponding

probability distributions $t(x)$ and $t(x^*)$. Meanwhile, the robust teacher (t') analyzes the adversarial examples at a fixed temperature, producing smoothed outputs $t'(x^*)$, to maintain its pre-trained defensive decision boundary. The student model processes both types of data through separate forward passes simultaneously, generating $s(x)$ for benign inputs and $s(x^*)$ for adversarial examples, thereby establishing a dual behavioral baseline for knowledge distillation. In summary, CIARD's innovation is ultimately manifested in a triple-objective loss function that dynamically balances competing learning goals. The concise knowledge transfer component ($\alpha \text{KL}(s(x), t(x))$) with adaptive temperature initially promotes broad class relationship learning, before gradually sharpening to emphasize discriminative features. Complementarily, the robust knowledge alignment term ($\beta \text{KL}(s(x^*), t'(x^*))$) transplants certified defense mechanisms by enforcing distributional consistency between the student's adversarial responses and the robust teacher's calibrated outputs.

3.3. Robust Specialization: Contrastive Push Loss

The core challenge in adversarial robustness distillation lies in effectively decoupling clean knowledge from robust knowledge, ensuring that student model can specialize in learning robust features without interference from clean teachers. To address this challenge, we propose “**Contrastive Push Loss**”, a novel component designed to enhance the student model's ability to specialize in robustness knowledge. Formally, given an adversarial sample x^* generated through Eq. 6, let $s(x^*)$ and $t(x^*)$ denote the output probability distributions of the student and clean teacher respectively. Based on Eq. 4, for a given input sample x and its adversarial counterpart x^* , the loss function is formulated as:

$$\begin{aligned} \mathcal{L}_{student} = & \underbrace{\alpha \text{KL}(s(x), t(x))}_{\text{Clean Knowledge}} + \underbrace{\beta \text{KL}(s(x^*), t'(x^*))}_{\text{Robust Knowledge}} \\ & - \underbrace{\lambda \text{Push}(s(x^*), t(x^*))}_{\text{Robust Specialization}}, \end{aligned} \quad (7)$$

where s , t , and t' represent the student model, clean teacher, and robust teacher respectively.

Unlike conventional distillation that minimizes this divergence, our formulation explicitly maximizes it through negative weighting in the global loss function (see Eq. 7). This creates a repulsive force that drives the student's adversarial predictions away from the clean teacher's vulnerable patterns while preserving alignment with the robust teacher's guidance through \mathcal{L}_{Robust} .

As shown in Algorithm 1, our proposed push loss works by deliberately creating a divergence between the student model and the clean teacher. This mechanism allows the student to focus exclusively on robust features when processing adversarial examples, minimizing the influence of

potentially misleading robustness knowledge from clean teacher. By pushing away from the clean teacher’s incorrect predictions, the student can more effectively absorb the robust teacher’s specialized knowledge. Meanwhile, the decoupling process also enhances the clean teacher’s robustness against adversarial data generated by the student model during training. As a result, the remaining unpushed data retains more accurate robust knowledge, which in turn further strengthens the robust training process. This approach creates a specialized learning environment where the student model can focus on developing robust features without the confusion introduced by clean but non-robust knowledge. By strategically diverging from the clean teacher in cases where robustness is critical, the student achieves better specialization in handling adversarial examples while maintaining performance on clean data.

3.4. Iterative Teacher Training

The cyclic iterative mechanism in CIARD fundamentally re-imagines the teacher-student relationship through the bidirectional flow of knowledge. Unlike traditional distillation methods where teachers stay unchanged, our approach creates a two-way learning process where both the student model improves and the teacher model develops at the same time. As shown in Figure 2, this mechanism operates in the following way:

$$\mathcal{L}_{adv_teacher} = \text{CE}(t'(x^*), y), \quad (8)$$

where t' represents the robust teacher model, x^* denotes the adversarial examples obtained by Eq. 6, and y is the ground truth label. By optimizing this loss function, the robust teacher model continuously adapts to evolving adversarial attack patterns, maintaining high performance levels throughout the training process. This dynamic adaptation mechanism ensures consistent quality of knowledge transfer and prevents the degradation of the robust teacher’s performance that is commonly observed in traditional fixed-teacher approaches.

In our 300-round model training process, we keep the adversarial teacher model fixed for the first 50 rounds, allowing the student model to maximize its learning of the teacher model’s robustness. After 50 rounds, as the student model continues training, the generated adversarial examples can cause the teacher model’s performance to decline. Therefore, we incorporate the robust teacher model into the training to maintain its performance at a high level. By optimizing this loss function, the robust teacher model continuously adapts to evolving adversarial examples, maintaining high performance levels throughout the training process. This dynamic adaptation mechanism ensures consistent quality of knowledge transfer and prevents the degradation of the robust teacher’s performance that is commonly observed in traditional fixed-teacher approaches.

Algorithm 1: Push Loss Computation

Input: $\mathbf{z}_t \in \mathbb{R}^{B \times C}$ – Teacher logits for a batch of size B and C classes; $\mathbf{z}_s \in \mathbb{R}^{B \times C}$ – Student logits for the same batch; $\mathbf{y} \in \mathbb{R}^B$ – Ground truth labels; Temperature T (default: 4).

Output: \mathcal{L}_{push}

```

1 Procedure:
2 /* Clean Teacher Robustness Evaluation */
3  $\hat{\mathbf{y}}_t \leftarrow \arg \max(\mathbf{z}_t, \dim = 1);$  //  $T_{predict}$ 
4  $J \leftarrow \{i \mid \hat{y}_t^{(i)} \neq y^{(i)}\};$  // Error indices
5 /* Clean Teacher Knowledge Filtering */
6  $\mathbf{z}'_t \leftarrow \{\mathbf{z}_t^{(i)} \mid i \in J\};$  // Select  $T_{logits}$ 
7  $\mathbf{z}'_s \leftarrow \{\mathbf{z}_s^{(i)} \mid i \in J\};$  // Select  $S_{logits}$ 
8 /* Robustness Decoupling & Push Loss */
9  $\tilde{\mathbf{p}}_s \leftarrow \text{softmax}\left(\frac{\mathbf{z}'_s}{T}\right);$  // S-Distribution
10  $\tilde{\mathbf{p}}_t \leftarrow \text{softmax}\left(\frac{\mathbf{z}'_t}{T}\right);$  // T-Distribution
11  $\mathcal{L}_{push} \leftarrow D_{KL}(\tilde{\mathbf{p}}_s \parallel \tilde{\mathbf{p}}_t);$  // KL divergence
12 Return  $\mathcal{L}_{push}$ 

```

3.5. Overall Training Objectives

To sum up, based on the dual-teacher framework for optimizing the student model, our proposed push loss and iterative teacher training effectively enhance the student model’s robustness against adversarial examples while maintaining high performance of the adversarial teacher model throughout the training process, overcoming the performance degradation issues commonly observed in traditional fixed-teacher approaches.

Specifically, for a given dataset x , we first obtain its corresponding adversarial examples x^* generated through PGD that jointly consider both the student model s and the clean teacher model t . We then input both clean examples and adversarial examples into the clean teacher model t , robust teacher model t' , and student model s to obtain the corresponding probability distributions $t(x)$, $t(x^*)$, $t'(x^*)$, $s(x)$, and $s(x^*)$, which serve as the foundation for computing the triple-objective loss function, and optimize student model parameters according to the following objective function:

$$\arg \min_{\theta} \mathcal{L}_{student}, \quad (9)$$

where α , β , and λ balance clean knowledge transfer, robust knowledge acquisition, and robust specialization, respectively. The detailed algorithm description of CIARD can be found in the supplementary files.

Table 2. Performance of Different Teacher Models (ResNet [RN] and WideResNet [WRN]) on CIFAR-10 and CIFAR-100 Datasets.

Dataset	Teacher Model	Type	Clean Acc	FGSM [9]	PGD _{SAT} [23]	PGD _{TRADES} [38]	CW _∞ [4]
CIFAR-10	RN-56	Clean	93.18%	19.18%	0	0	0
	WRN-34-10	Robust	84.92%	60.87%	56.86%	55.30%	53.84%
CIFAR-100	WRN-22-6	Clean	72.55%	25.19%	0	0	0
	WRN-70-16	Robust	63.56%	43.69%	32.24%	30.95%	28.93%

Table 3. White-box Adversarial Robustness of ResNet-18 on CIFAR-10 and CIFAR-100 Datasets. Detailed results from our experiments are presented in the supplementary files. The best results are **bolded**, and the second best results are underlined.

Attack	Defense	CIFAR-10(%)			CIFAR-100(%)		
		Clean	Robust	W-R	Clean	Robust	W-R
FGSM	SAT	84.20	55.59	69.90	56.16	25.88	41.02
	TRADES	83.00	58.35	70.68	57.75	31.36	44.56
	ARD	84.11	58.40	71.26	60.11	33.61	46.86
	RSLAD	83.99	60.41	72.20	58.25	<u>34.73</u>	46.49
	SCORE	84.43	59.84	72.14	56.40	32.94	44.67
	Fair-ARD	83.41	58.91	71.16	57.81	34.39	46.10
	ABSLD	83.21	60.22	71.72	56.77	34.94	45.86
	MTARD	87.36	61.20	74.28	64.30	31.49	47.90
	B-MTARD	88.20	61.42	74.81	65.08	34.21	49.65
	CIARD (Ours)	88.87	61.88	75.38	65.73	34.47	50.10
PGD _{SAT}	SAT	84.20	45.85	65.08	56.16	21.18	38.67
	TRADES	83.00	52.35	67.68	57.75	28.05	42.90
	ARD	84.11	50.93	67.52	60.11	29.40	44.76
	RSLAD	83.99	<u>53.94</u>	68.97	58.25	<u>31.19</u>	44.72
	SCORE	84.43	53.72	69.08	56.40	30.27	43.34
	Fair-ARD	83.41	52.00	67.71	57.81	30.64	44.23
	ABSLD	83.21	54.63	68.92	56.77	32.41	44.59
	MTARD	87.36	50.83	69.05	64.30	24.95	44.63
	B-MTARD	88.20	51.68	69.94	65.08	28.50	46.79
	CIARD (Ours)	88.87	51.70	70.29	65.73	28.05	46.89

Table 4. White-box Adversarial Robustness of MobileNet-V2 on CIFAR-10 and CIFAR-100 Datasets. Detailed results from our experiments are presented in the supplementary files. The best results are **bolded**, and the second best results are underlined.

Attack	Defense	CIFAR-10(%)			CIFAR-100(%)		
		Clean	Robust	W-R	Clean	Robust	W-R
FGSM	SAT	83.87	55.89	69.88	59.19	30.88	45.04
	TRADES	77.95	53.75	65.85	55.41	30.28	42.85
	ARD	83.43	57.03	70.23	60.45	32.77	46.61
	RSLAD	83.20	59.47	71.34	59.01	33.88	46.45
	SCORE	82.32	58.43	70.38	49.38	29.28	39.33
	Fair-ARD	82.65	56.37	69.51	59.18	<u>34.07</u>	46.63
	ABSLD	82.50	58.47	70.49	56.67	33.85	45.26
	MTARD	<u>89.26</u>	57.84	73.55	67.01	32.42	49.72
	B-MTARD	89.09	58.79	73.94	66.13	34.36	50.25
	CIARD (Ours)	89.51	<u>59.10</u>	74.31	<u>66.72</u>	33.56	<u>50.14</u>
PGD _{SAT}	SAT	83.87	46.84	65.36	59.19	25.64	42.42
	TRADES	77.95	49.06	63.51	55.41	23.33	39.37
	ARD	83.43	49.50	66.47	60.45	28.69	44.57
	RSLAD	83.20	<u>53.25</u>	68.23	59.01	<u>30.19</u>	44.60
	SCORE	82.32	53.42	67.87	49.38	27.03	38.21
	Fair-ARD	82.65	50.50	66.58	59.18	30.15	44.67
	ABSLD	82.50	52.98	67.74	56.67	31.28	43.98
	MTARD	<u>89.26</u>	44.16	66.71	67.01	25.14	46.08
	B-MTARD	89.09	47.56	<u>68.33</u>	66.13	28.47	47.30
	CIARD (Ours)	89.51	47.67	68.59	<u>66.72</u>	27.02	<u>46.87</u>

4. Experiments

4.1. Experimental Setup

Datasets & Models. We use the CIFAR-10 [18] and CIFAR-100 datasets, following Zhao et al. [39], to evaluate the experimental results of CIARD and other ARD methods. For model architecture, the student model uses ResNet-18 [11] and MobileNet-V2 [26]. The teacher models are categorized into a clean teacher and an adversarial teacher. The clean teacher uses ResNet-56 for CIFAR-10 and WideResNet-22-6 [37] for CIFAR-100. The adversarial teacher uses WideResNet-34-10 trained with TRADES [38] for CIFAR-10 and WideResNet-70-16 provided by Gowal et al [10], for CIFAR-100. The performance of these pre-trained teacher models is summarized in Table 2.

Implementation Details. Student models are trained for 300 epochs using an SGD optimizer (momentum 0.9, weight decay $2e-4$), with a learning rate following a cosine decay schedule from 0.1 to $1e-5$. The adversarial teacher is frozen for the first 50 epochs and then iteratively updated

using SGD with a low learning rate of $1e-5$. For robust training, adversarial examples are generated via a 10-step PGD attack with a step size of $2/255$ and an L_∞ bound of $\epsilon = 8/255$. The push loss temperature is set to 4. All experiments use a batch size of 64 and standard data augmentation (random cropping and horizontal flipping). Training was conducted in PyTorch on two NVIDIA RTX 4090 GPUs.

Evaluation Metrics. To evaluate model performance, we measure natural accuracy on clean test examples and robust accuracy on adversarial test examples. The evaluation protocol encompasses multiple attack methods: FGSM [9], PGD_{SAT} [23], PGD_{TRADES} [38], CW_∞ [4], and Square Attack [1], all configured with a maximum perturbation limit of $\epsilon = 8/255$. Both PGD_{SAT} and PGD_{TRADES} are implemented with 20 steps and a step size of $2/255$, while CW_∞ employs 30 steps. For the query-based attack, we set the number of queries for the Square Attack to 100.

Table 5. Black-box Adversarial Robustness of ResNet-18 on CIFAR-10 and CIFAR-100 Datasets Under Various Attack Methods. Performance is measured using Weighted Robustness (W-R) and accuracy metrics. Additional experimental details are provided in the supplementary materials.

Attack	Defense	CIFAR-10(%)			CIFAR-100(%)		
		Clean	Robust	W-R	Clean	Robust	W-R
PGD _{TRADES}	SAT	84.20	64.74	74.52	56.16	38.10	47.13
	TRADES	83.00	63.61	73.31	57.75	38.20	47.98
	ARD	84.11	63.50	73.81	60.11	39.53	49.82
	RSLAD	83.99	63.96	73.98	58.25	39.93	49.09
	MTARD	87.36	65.26	76.31	64.30	41.46	52.88
	B-MTARD	88.20	65.29	76.75	65.08	42.11	53.60
	CIARD (Ours)	88.87	66.28	77.58	65.73	42.29	54.01
CW _∞	SAT	84.20	63.84	74.02	56.16	39.42	47.79
	TRADES	83.00	62.83	72.92	57.75	38.63	48.19
	ARD	84.11	62.86	73.49	60.11	38.85	49.48
	RSLAD	83.99	63.05	73.52	58.25	39.67	48.96
	MTARD	87.36	64.58	75.97	64.30	41.18	52.74
	B-MTARD	88.20	64.64	76.42	65.08	41.35	53.22
	CIARD (Ours)	88.87	64.79	76.83	65.73	41.44	53.59
SA [1]	SAT	84.20	72.48	78.34	56.16	40.05	48.11
	TRADES	83.00	72.49	77.75	57.75	42.51	50.13
	ARD	84.11	74.60	79.36	60.11	47.20	53.66
	RSLAD	83.99	72.47	78.23	58.25	45.32	51.79
	MTARD	87.36	78.58	82.97	64.30	48.13	56.22
	B-MTARD	88.20	79.82	84.01	65.08	49.40	57.24
	CIARD (Ours)	88.87	80.03	84.45	65.73	49.76	57.75

4.2. Effectiveness of CIARD

Adversarial attacks can be categorized into two types based on the threat model: white-box and black-box attacks. In a white-box setting, the attacker has complete access to the deep learning model, including its architecture and parameters. In contrast, a black-box setup only allows the attacker to access the model’s output. Due to the length of the article, more detailed experimental results can be found in the supplementary files.

White-box Robustness of Student Models. To evaluate robustness against white-box attacks, we test the student models ResNet-18 and MobileNet-V2 on CIFAR-10 and CIFAR-100 using four attack methods: FGSM, PGD_{SAT}, PGD_{TRADES}, and CW_∞. As shown in Tables 3 and 4, the precision of CIARD W-Robust on CIFAR-10 and CIFAR-100 outperforms other ARD methods in most cases. For ResNet-18, the model weight robustness is improved by up to 0.57% and 0.45% based on the CIFAR-10 and CIFAR-100 datasets. Furthermore, the cyclic iteration mechanism in CIARD improves the accuracy of the student model’s classification, reaching 88.87% and 65.73% in CIFAR-10 and CIFAR-100, respectively. For MobileNet-V2, the results are similar, particularly under FGSM attack. Tests based on the CIFAR-10 dataset show an improvement in weighted robustness of 0.37% and 0.26% compared to the

Table 6. Black-box Adversarial Robustness of MobileNet-V2 on CIFAR-10 and CIFAR-100 Datasets Under Various Attack Methods. Performance is measured using Weighted Robustness (W-R) and accuracy metrics. Additional experimental details are provided in the supplementary materials.

Attack	Defense	CIFAR-10(%)			CIFAR-100(%)		
		Clean	Robust	W-R	Clean	Robust	W-R
PGD _{TRADES}	SAT	83.87	64.66	74.27	59.19	40.70	49.95
	TRADES	77.95	61.04	69.50	55.41	37.76	46.59
	ARD	83.43	63.28	73.36	60.45	39.15	49.80
	RSLAD	83.20	64.33	73.77	59.01	40.32	49.67
	MTARD	<u>89.26</u>	<u>66.30</u>	<u>77.78</u>	67.01	43.23	55.12
	B-MTARD	89.09	66.47	77.78	66.13	42.67	54.40
	CIARD (Ours)	89.51	66.66	78.09	66.72	42.70	54.71
CW _∞	SAT	83.87	64.24	74.06	59.19	40.97	50.08
	TRADES	77.95	60.66	69.31	55.41	38.02	46.72
	ARD	83.43	62.83	73.13	60.45	38.53	49.49
	RSLAD	83.20	63.45	73.33	59.01	39.92	49.47
	MTARD	<u>89.26</u>	<u>65.68</u>	<u>77.47</u>	67.01	42.92	54.97
	B-MTARD	89.09	65.96	77.53	66.13	42.04	54.09
	CIARD (Ours)	89.51	66.12	77.82	66.72	42.85	54.79
SA [1]	SAT	83.87	73.01	78.44	59.19	44.51	51.85
	TRADES	77.95	67.43	62.69	55.41	40.71	48.06
	ARD	83.43	73.26	78.35	60.45	46.95	53.70
	RSLAD	83.20	73.07	78.14	59.01	45.66	52.34
	MTARD	<u>89.26</u>	<u>79.13</u>	<u>84.20</u>	67.01	50.64	58.83
	B-MTARD	89.09	79.61	84.35	66.13	50.83	58.48
	CIARD (Ours)	89.51	80.01	84.76	66.72	50.85	58.79

best benchmark method.

Black-box Robustness of Student Models. We also perform black-box evaluations, incorporating both transfer-based and query-based [1] methods, to assess the student model’s robustness in environments more akin to real-world scenarios. The parameter configurations for the student and teacher models are consistent with those used in the white-box evaluations.

Using CIFAR-10 and CIFAR-100, we evaluate the defensive capabilities of CIARD and other methods against black-box attacks on ResNet-18 and MobileNet-V2, focusing on both transfer-based and query-based attacks. For transfer-based attacks, we use adversarial teachers (WideResNet-34-10 and WideResNet-70-16) to generate adversarial examples for PGD_{TRADES} and CW_∞ attacks. For query-based attacks, we employ Square Attack (SA). We select the best checkpoints for the baseline model and MTARD based on weighted robustness accuracy. Tables 5 and 6 indicate that CIARD generally exhibits stronger resilience against the three types of black-box attacks. Notably, against PGD_{TRADES} attacks, the student model achieve weighted robustness accuracy in CIFAR-10 that was 0.83% and 0.41% higher than the second-best methods, respectively. In summary, the experimental results clearly demonstrate the superior performance of CIARD in defend-

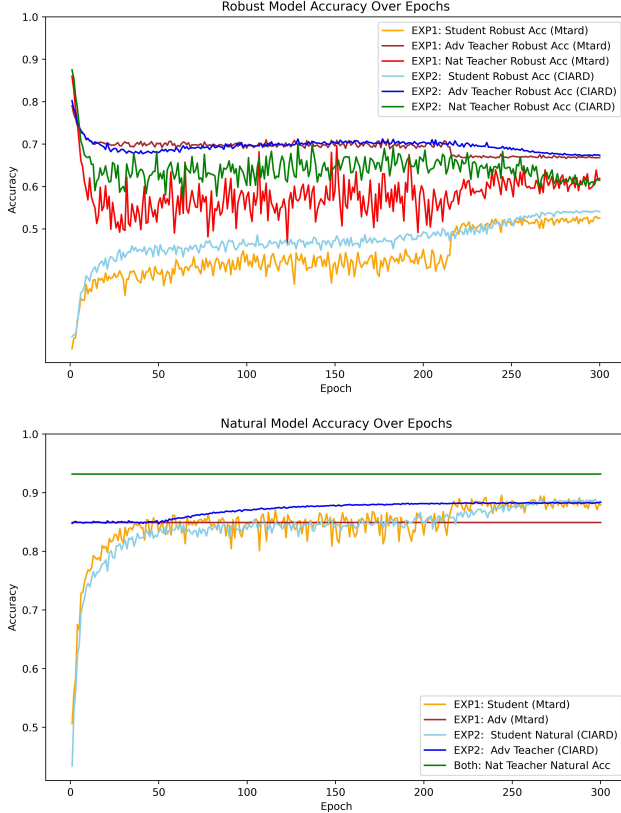


Figure 3. Training curves of ResNet-18 on CIFAR-10 over 300 epochs. Up: Robust accuracy comparison of MTARD/CIARD. Down: Classification accuracy comparison of MTARD/CIARD.

ing against both white-box and black-box attacks. Based on experimental results, ARD, RSLAD, MTARD, and CIARD outperform SAT and TRADES, demonstrating that ARD methods are more effective in enhancing lightweight model performance compared to traditional approaches.

4.3. Ablation Studies

Quantitative Analysis of CIARD. To better understand the contribution of each component in our proposed method, we conduct ablation studies on both Iterative Teacher Training (ITT) and Push loss. Table 7 presents the performance comparison on PGD adversarial accuracy ($\text{PGD}_{\text{TRADES}}$), clean accuracy ($\text{Clean}_{\text{acc}}$), and weighted accuracy (W-acc).

The effects of ITT. From the results, we observe that incorporating ITT improves both clean accuracy and adversarial robustness compared to the baseline model (without ITT and Push loss). Specifically, ITT alone increases clean accuracy by 0.87% and adversarial robustness by 0.19%. This improvement demonstrates the effectiveness of our implicit teacher training strategy in enhancing model generalization.

The effects of contrastive push loss. We further incorporate the push loss into our framework, we achieve the best performance across all metrics. The full model achieves

Table 7. ResNet-18 Results on CIFAR-10. Push Loss refers to Contrastive Push Loss in 3.3, while ITT represents the Iterative Teacher Training in 3.4.

Method	$\text{PGD}_{\text{TRADES}}$	$\text{Clean}_{\text{acc}}$	W-acc
w/o ITT w/o Push loss	53.84	87.76	70.80
w/ ITT w/o Push loss	54.03	88.63	71.30
w/ ITT w/ Push loss	54.54	88.86	71.70

54.54% on $\text{PGD}_{\text{TRADES}}$, 88.86% on $\text{Clean}_{\text{acc}}$, and 71.70% on W-acc, showing consistent improvements over both the baseline and the model with only ITT. The Push loss contributes an additional 0.51% gain in adversarial robustness and 0.23% in clean accuracy. The effectiveness of our contrastive push loss can be attributed to its decoupling design, which helps separate the decision boundaries between clean and adversarial examples. As shown in our prior analysis, this decoupling mechanism allows the clean teacher to maintain higher robustness throughout the training process. Meanwhile, the robust teacher’s capability continuously improves as training progresses, benefiting from the knowledge transfer facilitated by the push loss.

Qualitative Analysis of CIARD. In addition to the above experimental results, line chart 3 also validates our design choices and confirms that ITT and Push loss are both key components of our framework, each contributing to overall performance in terms of robustness and accuracy. Despite the clean teacher being frozen to preserve its high clean accuracy, its robust accuracy has significantly improved during training, demonstrating the effectiveness of our decoupling mechanism with push loss, as mentioned in 3.3. Moreover, both the robust and clean accuracy of our robust teacher have improved and remained consistently high, further validating the contribution of our Iterative Teacher Training method.

5. Conclusion

In this paper, we propose the Cyclic Iterative Adversarial Robustness Distillation (CIARD) method to address the conflict between the teacher optimization objective and performance degradation in adversarial robustness distillation. Our framework introduces: ① a contrastive push-pull alignment mechanism to resolve the objective conflict by regulating the distance between the student adversarial response and the teacher clean features; and ② dynamic adversarial retraining to preserve the teacher robustness via parameter freezing and adaptive knowledge adjustment. Extensive experiments on CIFAR-10/100 and Tiny-ImageNet demonstrate the effectiveness of CIARD. The method sets a new benchmark for compact adversarial models and is planned to be extended to various attack scenarios and resource-constrained deployment in the future.

6. Acknowledgment

Liming Lu, Xiang Gu and Shuchao Pang are supported by the National Natural Science Foundation of China (Grant No.62206128), National Key Research and Development Program of China under (Grant No.2023YFB2703900) and the Postgraduate Research & Practice Innovation Program of Jiangsu Province (Grant No.KYCX24.0723). Anan Du is supported by the Start-up Fund for New Talented Researchers of Nanjing University of Industry Technology (Grant No.YK24-05-04) Yongbin Zhou is supported by the National Natural Science Foundation of China (Grant No.U2336205).

References

- [1] Maksym Andriushchenko, Francesco Croce, Nicolas Flammarion, and Matthias Hein. Square attack: a query-efficient black-box adversarial attack via random search. In *European conference on computer vision*, pages 484–501. Springer, 2020. 6, 7
- [2] Anish Athalye, Nicholas Carlini, and David Wagner. Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples. In *International conference on machine learning*, pages 274–283. PMLR, 2018. 1
- [3] Yang Bai, Yisen Wang, Yuyuan Zeng, Yong Jiang, and Shu-Tao Xia. Query efficient black-box adversarial attack on deep neural networks. *Pattern Recognition*, 133:109037, 2023. 1
- [4] Nicholas Carlini and David Wagner. Towards evaluating the robustness of neural networks. In *2017 IEEE Symposium on Security and Privacy (SP)*, pages 39–57. Ieee, 2017. 6
- [5] Alvin Chan, Yi Tay, and Yew-Soon Ong. What it thinks is important is important: Robustness transfers through input gradients. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 332–341, 2020. 3
- [6] Francesco Croce and Matthias Hein. Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks. In *International conference on machine learning*, pages 2206–2216. PMLR, 2020. 1
- [7] Kevin Eykholt, Ivan Evtimov, Earlene Fernandes, Bo Li, Amir Rahmati, Chaowei Xiao, Atul Prakash, Tadayoshi Kohno, and Dawn Song. Robust physical-world attacks on deep learning visual classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1625–1634, 2018. 1
- [8] Micah Goldblum, Liam Fowl, Soheil Feizi, and Tom Goldstein. Adversarially robust distillation. In *Proceedings of the AAAI conference on artificial intelligence*, pages 3996–4003, 2020. 2
- [9] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014. 2, 6
- [10] Sven Gowal, Chongli Qin, Jonathan Uesato, Timothy Mann, and Pushmeet Kohli. Uncovering the limits of adversarial training against norm-bounded adversarial examples. *arXiv preprint arXiv:2010.03593*, 2020. 6
- [11] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 1, 6
- [12] Geoffrey Hinton. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015. 1
- [13] Lei Hsiung, Yun-Yun Tsai, Pin-Yu Chen, and Tsung-Yi Ho. Towards compositional adversarial robustness: Generalizing adversarial training to composite semantic perturbations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 24658–24667, 2023. 1
- [14] Bo Huang, Mingyang Chen, Yi Wang, Junda Lu, Minhao Cheng, and Wei Wang. Boosting accuracy and robustness of student models via adaptive adversarial distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24668–24677, 2023. 2
- [15] Xiaojun Jia, Yong Zhang, Baoyuan Wu, Ke Ma, Jue Wang, and Xiaochun Cao. Las-at: adversarial training with learnable attack strategy. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13398–13408, 2022. 1
- [16] Xiaojun Jia, Jianshu Li, Jindong Gu, Yang Bai, and Xiaochun Cao. Fast propagation is better: Accelerating single-step adversarial training via sampling subnetworks. *IEEE Transactions on Information Forensics and Security*, 19: 4547–4559, 2024.
- [17] Xiaojun Jia, Yong Zhang, Xingxing Wei, Baoyuan Wu, Ke Ma, Jue Wang, and Xiaochun Cao. Improving fast adversarial training with prior-guided knowledge. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(9):6367–6383, 2024. 1
- [18] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009. 6
- [19] Yao Li, Minhao Cheng, Cho-Jui Hsieh, and Thomas CM Lee. A review of adversarial attack and defense for classification methods. *The American Statistician*, 76(4):329–345, 2022. 1
- [20] Liming Lu, Zhenghan Chen, Xiaoyu Lu, Yihang Rao, Lujun Li, and Shuchao Pang. Uniads: Universal architecture-distiller search for distillation gap. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 14167–14174, 2024. 1
- [21] Xingjun Ma, Bo Li, Yisen Wang, Sarah M Erfani, Sudanthi Wijewickrema, Grant Schoenebeck, Dawn Song, Michael E Houle, and James Bailey. Characterizing adversarial subspaces using local intrinsic dimensionality. *arXiv preprint arXiv:1801.02613*, 2018. 1
- [22] Xingjun Ma, Yuhao Niu, Lin Gu, Yisen Wang, Yitian Zhao, James Bailey, and Feng Lu. Understanding adversarial attacks on deep learning based medical image analysis systems. *Pattern Recognition*, 110:107332, 2021. 1
- [23] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017. 1, 4, 6

- [24] Javier Maroto, Guillermo Ortiz-Jiménez, and Pascal Frossard. On the benefits of knowledge distillation for adversarial robustness. *arXiv preprint arXiv:2203.07159*, 2022. 3
- [25] Nicolas Papernot, Patrick McDaniel, Somesh Jha, Matt Fredrikson, Z Berkay Celik, and Ananthram Swami. The limitations of deep learning in adversarial settings. In *2016 IEEE European symposium on security and privacy (EuroS&P)*, pages 372–387. IEEE, 2016. 1
- [26] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4510–4520, 2018. 6
- [27] Ali Shafahi, Parsa Saadatpanah, Chen Zhu, Amin Ghiasi, Christoph Studer, David Jacobs, and Tom Goldstein. Adversarially robust transfer learning. *arXiv preprint arXiv:1905.08232*, 2019. 3
- [28] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013. 1
- [29] Pratik Vaishnavi, Kevin Eykholt, and Amir Rahmati. Transferring adversarial robustness through robust representation matching. In *31st USENIX Security Symposium (USENIX Security 22)*, pages 2083–2098, 2022. 3
- [30] Haichen Wang, Shuchao Pang, Zhigang Lu, Yihang Rao, Yongbin Zhou, and Minhui Xue. dp-promise: Differentially private diffusion probabilistic models for image synthesis. In *33rd USENIX Security Symposium (USENIX Security 24)*, pages 1063–1080, 2024. 1
- [31] Yisen Wang, Xuejiao Deng, Songbai Pu, and Zhiheng Huang. Residual convolutional ctc networks for automatic speech recognition. *arXiv preprint arXiv:1702.07793*, 2017. 1
- [32] Yisen Wang, Difan Zou, Jinfeng Yi, James Bailey, Xingjun Ma, and Quanquan Gu. Improving adversarial robustness requires revisiting misclassified examples. In *International conference on learning representations*, 2019. 3
- [33] Boxi Wu, Jinghui Chen, Deng Cai, Xiaofei He, and Quanquan Gu. Do wider neural networks really help adversarial robustness? *Advances in Neural Information Processing Systems*, 34:7054–7067, 2021. 3
- [34] Wentao Xu, Qianqian Xie, Shuo Yang, Jiangxia Cao, and Shuchao Pang. Enhancing content-based recommendation via large language model. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, pages 4153–4157, 2024. 1
- [35] Zhendong Yang, Ailing Zeng, Zhe Li, Tianke Zhang, Chun Yuan, and Yu Li. From knowledge distillation to self-knowledge distillation: A unified approach with normalized loss and customized soft labels. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 17185–17194, 2023. 1
- [36] Xinli Yue, Mou Ningping, Qian Wang, and Lingchen Zhao. Revisiting adversarial robustness distillation from the perspective of robust fairness. *Advances in Neural Information Processing Systems*, 36, 2024. 2
- [37] Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. *arXiv preprint arXiv:1605.07146*, 2016. 6
- [38] Hongyang Zhang, Yaodong Yu, Jiantao Jiao, Eric Xing, Laurent El Ghaoui, and Michael Jordan. Theoretically principled trade-off between robustness and accuracy. In *International conference on machine learning*, pages 7472–7482. PMLR, 2019. 2, 3, 6
- [39] Shiji Zhao, Jie Yu, Zhenlong Sun, Bo Zhang, and Xingxing Wei. Enhanced accuracy and robustness via multi-teacher adversarial distillation. In *European Conference on Computer Vision*, pages 585–602. Springer, 2022. 2, 3, 4, 6
- [40] Shiji Zhao, Xizhe Wang, and Xingxing Wei. Improving adversarial robust fairness via anti-bias soft label distillation. *arXiv preprint arXiv:2312.05508*, 2023. 2
- [41] Shiji Zhao, Xizhe Wang, and Xingxing Wei. Mitigating accuracy-robustness trade-off via balanced multi-teacher adversarial distillation. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, (01):1–14, 2024. 2
- [42] Wangchunshu Zhou, Canwen Xu, and Julian McAuley. Bert learns to teach: Knowledge distillation with meta learning. *arXiv preprint arXiv:2106.04570*, 2021. 1
- [43] Jianing Zhu, Jiangchao Yao, Bo Han, Jingfeng Zhang, Tongliang Liu, Gang Niu, Jingren Zhou, Jianliang Xu, and Hongxia Yang. Reliable adversarial distillation with unreliable teachers. *arXiv preprint arXiv:2106.04928*, 2021. 3
- [44] Kaijie Zhu, Xixu Hu, Jindong Wang, Xing Xie, and Ge Yang. Improving generalization of adversarial training via robust critical fine-tuning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4424–4434, 2023. 2
- [45] Bojia Zi, Shihao Zhao, Xingjun Ma, and Yu-Gang Jiang. Revisiting adversarial robustness distillation: Robust soft labels make student better. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 16443–16452, 2021. 2, 3