GCMRD: Global Consistency Multi-teacher Robustness Distillation

Yuhang Zhou[1], Zhongyun Hua[1], Rushi Lan[2], Qing Liao[1], and Wei Jiang[3]*

[1] Harbin Institute of Technology, Shenzhen, Shenzhen, China
{23B951015, huazhongyun}@hit.edu.cn
[2] Guilin University of Electronic Technology, Guilin, China
{rslan2016}@163.com
[3] Harbin Institute of Technology, Harbin, China
{jw}@hit.edu.cn

Abstract. Adversarial Robustness Distillation (ARD) aims to transfer both natural and robust knowledge from a strong teacher model to a lightweight student model, thereby achieving robust performance in resource-constrained settings. The dual-teacher robustness distillation paradigm further provides an inspiring solution for optimizing the 'accuracy-robustness' trade-off. However, existing dual-teacher distillation frameworks fail to fully align the supervision provided by teacher group, resulting in suboptimal performance. To fully utilize the potential of dual-teacher robustness distillation, we propose a novel multi-teacher framework called global consistency multi-teacher robustness distillation (GCMRD), which improves the existing dual-teacher paradigm from three perspectives. First, for internal maximization, we introduce an integrated soft-label attack objective to replace hard-label or single-teacher-guided objective, leading to consistent attack preferences across both natural and robust teachers. Second, for external minimization, we design a fully-repulsion regularization mechanism based on the erroneous adversarial outputs of the natural teacher, which enables more comprehensive utilization of teacher supervision. Finally, we employ the outputs of the natural teacher to guide the fine-tuning of the robust teacher, indirectly balancing the student model's accuracy-robustness trade-off by enforcing consistency with the teacher ensemble. Experimental results demonstrate that our GCMRD achieves state-of-the-art performance in both natural accuracy and adversarial robustness.

Keywords: Adversarial attack & defense · Robustness distillation

1 Introduction

Deep learning models are highly vulnerable to adversarial attacks induced by carefully crafted adversarial samples (5; 7; 14). By introducing small, human-imperceptible perturbations to the input data, an attacker can easily mislead

* corresponding author

a model into producing completely incorrect predictions. This inherent vulnerability poses significant security concerns to the reliable deployment of deep learning models in safety-critical and high-stakes domains, such as healthcare, finance services, and autonomous driving.

To maintain the discriminative capability of deep learning models under adversarial attacks, i.e., adversarial robustness, a variety of countermeasures have been proposed, collectively referred to as adversarial defenses (2; 16; 20). Among these approaches, adversarial training (AT) (11; 22; 25) is widely regarded as the most reliable empirical defense strategy. In adversarial training, adversarial examples are generated online during the training process and incorporated into the training data, enabling models to learn robustness within a min-max optimization framework. Despite its effectiveness, adversarial training suffers from several notable limitations. First, compared to vanilla training, adversarial training introduces substantially higher computational overhead due to the iterative generation of adversarial examples. Second, it tends to favor large-capacity models, while smaller models often exhibit significantly degraded performance under adversarial training. Third, improving adversarial robustness often comes at the cost of reduced accuracy on clean data, leading to the well-known 'accuracy–robustness' trade-off.

To overcome the limitations of adversarial training, robustness distillation (RD) (6; 17; 26; 27) has been proposed, where a lightweight student model learns robust knowledge from a larger adversarially trained teacher model within a min-max distillation framework. Furthermore, dual-teacher robustness distillation has been introduced to better address the inherent accuracy-robustness trade-off. MTARD (23; 24) first proposed an expert strategy where a natural teacher provides natural knowledge and a robust teacher provides robust knowledge to the student model, together with an adaptive loss weighting mechanism, thereby achieving an improved accuracy-robustness trade-off. Building on this idea, CIARD (12) further enhances the framework by encouraging the student's adversarial outputs to deviate from the natural teacher's erroneous adversarial predictions and introducing a fine-tuning strategy for the robust teacher during training. Compared with the single-teacher settings, dual-teacher frameworks demonstrate greater potential in balancing natural accuracy and adversarial robustness, and gradually become the mainstream configuration for robustness distillation.

However, state-of-the-art dual-teacher robustness distillation methods still suffer from suboptimal performance due to insufficient utilization of the dual-teacher supervision. First, the inner maximization process is typically restricted to hard-label cross-entropy objectives, which overlooks the potential benefits of incorporating soft supervision from multiple teachers to guide adversarial example generation. Second, the application of repulsion regularization remains incomplete. For example, the push loss introduced in CIARD only encourages the student's adversarial output to deviate from the natural teacher's erroneous adversarial predictions, while neglecting the alignment between the student and teacher on clean samples, thereby limiting the effectiveness of teacher supervi-

sion. Finally, existing dual-teacher approaches fail to consider the collaborative consistency between teacher models. In practice, the natural and robust teachers not only encode complementary knowledge but also often differ in backbone architectures, resulting in substantial discrepancies in their output distributions and prediction preferences. Ignoring such inter-teacher inconsistency introduces conflicts during knowledge transfer, making it difficult for the student to simultaneously reconcile natural accuracy and adversarial robustness, and ultimately leading to a suboptimal accuracy–robustness trade-off.

To address the aforementioned limitations, we propose a composite dynamic dual-teacher distillation framework called global consistency multi-teacher robustness distillation (GCMRD). First, to fully exploit the potential of dual teachers in the internal maximization stage, we introduce an integrated attack objective based on the weighted divergence between the student model and the ensemble outputs of the teachers, rather than relying on hard-label cross-entropy or single-teacher KL divergence. This design enables adversarial example generation that reflects consistent attack preferences across both the natural and robust teachers. Second, to more effectively employ the supervisory signals from the natural teacher, we refine the repulsion regularization in CIARD (12). Specifically, our regularization not only encourages the student's adversarial outputs to deviate from the natural teacher's erroneous adversarial predictions, but also enforces repulsion between the student's clean outputs and those erroneous adversarial output. In this way, the natural teacher's incorrect adversarial predictions are explicitly utilized as negative supervision, leading to more comprehensive knowledge transfer. Finally, during the fine-tuning of the robust teacher, we impose a consistency constraint on the output preferences within the teacher ensemble, thereby indirectly balancing the student model's accuracy-robustness trade-off. Instead of using hard-label cross-entropy to update the robust teacher, we guide its optimization with the natural teacher's soft predictions on clean samples. This strategy promotes alignment between teachers while mitigating the conflict between the student's natural and adversarial objectives, as all optimization signals are ultimately anchored to the natural teacher's clean-data oracle. GCMRD's intuitive differences from and connections to existing robustness distillation methods can be referenced in Figure 1, and we provide a formal comparison between GCMRD and existing methods in the Appendix. We also provide a more detailed discussion on related work in the Appendix. The contributions of this paper can be summarized as follows:

- We analyze the suboptimal performance of existing multi-teacher robustness distillation methods and find that it is mainly caused by the insufficient use of ensemble teacher supervision in adversarial generation, knowledge transfer, and teacher collaboration.
- Based on this analysis, we propose a novel dual-teacher robust distillation framework, GCMRD, that better utilizes both teachers to improve the student model's robustness and its accuracy-robustness trade-off.

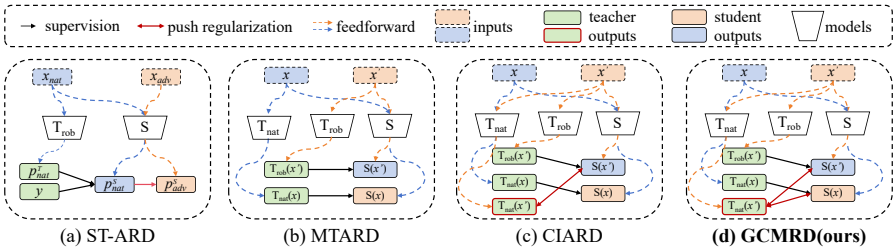


Fig. 1: A visual comparison of existing robustness distillation methods, where ST-ARD, MTARD, and CIARD represent single-teacher adversarial robustness distillation (6; 26; 27), multi-teacher adversarial robustness distillation (23; 24), and cyclic iterative adversarial robustness distillation (12), respectively.

– Extensive experiments demonstrate that GCMRD consistently improves both natural accuracy and adversarial robustness, validating the effectiveness of the proposed framework.

Note that our method does not introduce undistillable properties (13), despite incorporating push regularization. On the one hand, undistillable models are typically obtained by directly optimizing the teacher with a push regularization objective, thereby making its representations difficult for a student to mimic. In contrast, our approach applies repulsion only on the student model by encouraging it to deviate from the teacher's erroneous outputs, without modifying the teacher itself. On the other hand, our method does not enforce complete divergence between the student and teacher. Instead, the student is encouraged to align with the natural teacher's correct predictions on clean data while simultaneously distancing itself from the teacher's erroneous predictions on adversarial samples. In this way, the supervisory signals from the natural teacher are effectively exploited from both positive (clean) and negative (adversarial) perspectives.

2 Methods

2.1 Preliminary

Adversarial attack. Given a target model f_θ (simplified as f) parameterized by θ, the objective of the attacker can be formulated as a conditional optimization problem:

$$\arg\max_{x'} \mathcal{L}(f(x',\theta),y), \text{ s.t. } \|x'-x\|_p \leq \varepsilon, \tag{1}$$

where x' denotes the adversarial examples, y represents the corresponding labels, $\mathcal{L}(\cdot)$ represents the classification loss (e.g., cross entropy loss), and ε is the upper bound of the attack budget under the l_p-norm. The goal of the attacker is to fool the target model with visually imperceptible perturbations.

Traditional adversarial training achieves the inner maximization by maximizing the cross-entropy loss guided by hard labels, which can be formalized as:

$$x' = \underset{||\delta||_p \leq \epsilon}{\arg\max}\ \mathrm{CE}\left(S\left(x+\delta;\theta\right), y\right). \tag{2}$$

Later work, RSLAD, proposes achieving internal maximization by maximizing the output metric between the teacher and student models, which leads to improved robustness results. This can be formalized as:

$$x' = \underset{||\delta||_p \leq \epsilon}{\arg\max}\ \mathrm{KL}\left(S\left(x+\delta\right), T(x)\right). \tag{3}$$

Adversarial defense aims to preserve the discriminability capability of f_θ under adversarial attacks, i.e., achieving adversarial robustness. The objective of defense can be formalized as:

$$\underset{\theta}{\arg\min}\ \mathcal{L}(f(x',\theta), y). \tag{4}$$

Robustness distillation (RD) addresses both adversarial robustness and edge deployment efficiency challenges in a min-max distillation framework. Given a student model $S(\cdot)$ parameterized by θ_S and a robust teacher model $T_{rob}(\cdot)$ parameterized by $\theta_{T_{rob}}$, the main goal of robustness distillation is to transfer the robustness of the robust teacher model against adversarial examples to the smaller student model:

$$\underset{\theta_S}{\arg\min}\ \mathcal{L}(S(x), y) + \mathcal{D}(S(x'), T_{rob}(x')), \tag{5}$$

where $\mathcal{D}(\cdot)$ denotes a discrepancy metric in output distribution between the teacher and student models, typically measured using metrics like KL divergence.

Multi-teacher adversarial robustness distillation (MTARD) extends robustness distillation by no longer relying on labels or a single robust teacher for supervision during the distillation process, but instead employing an expert strategy: the natural teacher transfers natural knowledge, while the robust teacher transfers robust knowledge. The above process can be formalized as:

$$\arg\min_{\theta_S}[\underbrace{\alpha \cdot \mathcal{L}_{KL}\left(S(x), T_{nat}(x)\right)}_{\text{Natural Knowledge}} + \underbrace{\beta \cdot \mathcal{L}_{KL}\left(S\left(x'\right), T_{rob}\left(x'\right)\right)}_{\text{Robust Knowledge}}], \tag{6}$$

where T_{nat} and T_{rob} represents the natural teacher and robust teacher respectively, and α and β are the hyperparameters.

Cyclic iterative adversarial robustness distillation (CIARD) further introduces the push regularization to MTARD, utilizing the erroneous outputs of the clean teacher as negative supervision. This can be formalized as:

$$\begin{aligned}\mathcal{L}_{\text{student}} &= \alpha \cdot \underbrace{\mathrm{KL}(S(x), T_{nat}(x))}_{\text{Clean Knowledge}} + \beta \cdot \underbrace{\mathrm{KL}\left(S\left(x'\right), T_{rob}\left(x'\right)\right)}_{\text{Robust Knowledge}} \\ &\quad - \lambda \cdot \underbrace{\mathrm{Reg}\left(S\left(x'\right), T_{nat}\left(x'\right)\right)}_{\text{Repulsion Regularization}},\end{aligned} \tag{7}$$

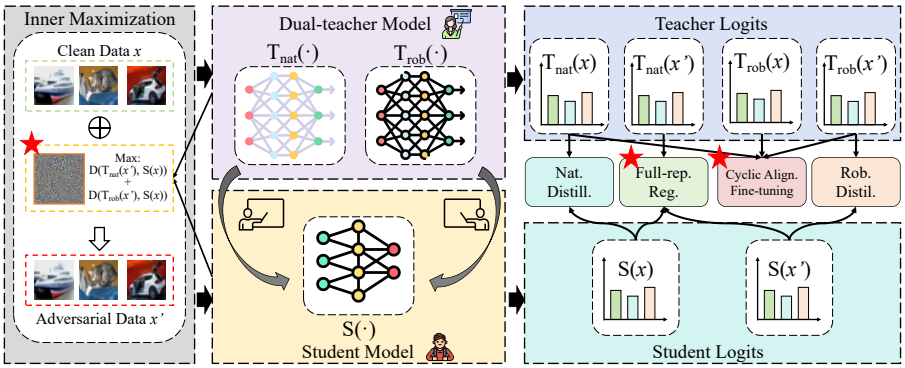


Fig. 2: The overall framework of the proposed GCMRD. It primarily comprises three innovative aspects: multi-teacher collaboration inner maximization, full-repulsion regularization, and natural-teacher guided fine-tuning of the robust teacher.

where λ is the hyperparameter of the regularization term. Still, as we have analyzed, CIARD does not fully exploit the potential of the teacher group for the inner maximization, repulsion regularization, and robust teacher fine-tuning.

2.2 Global Consistency Multi-teacher Robustness Distillation

Multi-teacher collaboration inner maximization (MCIM). Existing internal maximization schemes either overlook the role of soft labels or rely solely on single-teacher supervision, ignoring the adversarial potential of dual teachers. We propose simultaneously attacking both the natural and the robust teachers to construct the internal maximization objective, which can be formalized as:

$$x' = \underset{||x'-x||_p \leq \epsilon}{\arg\max} \ \mathrm{KL}\left(S\left(x'\right), T_{nat}(x)\right) + \mathrm{KL}\left(S\left(x'\right), T_{rob}(x)\right), \tag{8}$$

First, constructing internal maximization using soft labels has been demonstrated to be more effective than using hard labels (27). Moreover, the Kullback-Leibler divergence built with the robust teacher aligns with the attack objective of RSLAD, while its coupling with the natural model is equivalent to further exploring hard examples for the natural knowledge, thereby initially optimizing the trade-off between clean and robust knowledge distillation.

Full-repulsion regularization (FRR). CIARD proposes to push the student model's output on adversarial examples away from the natural teacher's erroneous outputs on adversarial examples. We argue that this approach overlooks the regularizing effect of the natural teacher's erroneous outputs on adversarial examples as a form of negative supervision for the student model's natural task. Further, we propose to also distance the student model's output on clean samples from the teacher model's erroneous outputs on adversarial examples. Such

a new regularization can be formalized as:

$$\begin{aligned}\mathcal{L}_{\text{student}} &= \alpha \cdot \underbrace{\mathrm{KL}(S(x), T_{nat}(x))}_{\text{Clean Knowledge}} + \beta \cdot \underbrace{\mathrm{KL}\left(S\left(x'\right), T_{rob}\left(x'\right)\right)}_{\text{Robust Knowledge}} \\ &- \lambda \cdot \underbrace{\left(\mathrm{Push}\left(S\left(x'\right), T_{nat}\left(x'\right)\right)\right) + \mathrm{Push}(S(x), T_{nat}(x'))}_{\text{Full-repulsion Regularization}},\end{aligned} \tag{9}$$

where Push$(\cdot)$ represents the repulsion regularization function formalized by KL divergence. Intuitively, the Full-repulsion Regularization simultaneously incorporates the repulsive optimization effect of the natural teacher's erroneous outputs on both the natural and adversarial tasks, providing the student model with a triplet-like optimization pattern. Here, besides $T_{nat}(x)$ and $T_{rob}(x')$ being treated as positive guiding labels, $T_{nat}(x)$ serves as a novel consistency negative supervision signal, leading to a consistency regularization between the natural and adversarial tasks at the logits level: both tasks are pushed away from this incorrect answer.

Consistency fine-tuning of the teachers (CFT). Current research has demonstrated that robust teacher models may experience a degradation in its robustness supervisory capability during the later stages of distillation. This is likely because, as the student model improves, the adversarial examples generated during the distillation process based on the student also become more challenging. CIARD proposes fine-tuning the robust teacher model using hard labels during the later epoch of the training process, which can be formalized as:

$$\mathcal{L}_{\text{rob\_teacher}} = \mathrm{CE}\left(T_{rob}\left(x'\right), y\right), \tag{10}$$

where x' represents the adversarial examples based on the student model.

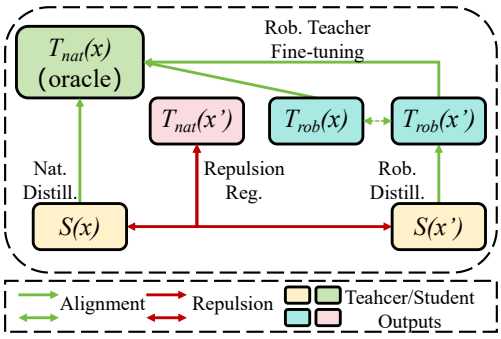


Fig. 3: Cyclic consistency. Fine-tuning the robust teacher with the natural teacher indirectly achieves consistency between natural and adversarial tasks, as all tasks directly or indirectly take the soft label outputs of the natural teacher on clean samples as the oracle.

However, while this fine-tuning approach mitigates the degradation of the robust teacher, it ignores the issue of collaboration between natural teacher and robust teacher, indirectly leading to a suboptimal trade-off between accuracy and robustness for the student model caused by inconsistent natural knowledge and robust knowledge preference. Specifically, the natural teacher and robust teacher often differ not only in model architecture but also in their training manner (vanilla training and adversarial training). This results in significant gap in both gradient patterns and output preferences. Directly

Table 1: Performance of the teacher models used in our experiments.

Dataset	Model	Clean	FGSM	PGD_{SAT}	$\text{PGD}_{\text{TRADES}}$	CW	Type
CIFAR10	RN-56	93.18%	19.23%	0%	0%	0%	Clean
	WRN-34-10	84.91%	61.14	55.30	56.61	53.84	Adv.
CIFAR100	WRN-22-6	76.65%	4.85%	0%	0%	0%	Clean
	WRN-70-16	60.96%	35.89%	33.58%	33.99%	31.05%	Adv.

applying an expert strategy to the student model using two such divergent models, i.e., the natural teacher performs natural knowledge distillation while the robust teacher performs robust knowledge distillation, would inevitably create conflicts between the natural and adversarial tasks. To alleviate this task conflict, we refine the fine-tuning objective of the robust model by incorporating supervision from the natural teacher, where the optimize objective can be rewrite as:

$$\mathcal{L}_{\text{rob\_teacher}} = \text{KL}\left(T_{rob}\left(x'\right), T_{nat}(x)\right) + \text{KL}\left(T_{rob}\left(x\right), T_{nat}(x)\right), \tag{11}$$

Intuitively, we supervise the robust teacher's output on adversarial samples with the natural teacher's output on clean samples. Such method is expected to reconcile the natural task and robust task of the student model, thereby optimizing the accuracy-robustness trade-off in the student model. Although such paradigm does not directly enforce consistency between the two tasks of the student model, it fosters consistency between the natural teacher and robust teacher, indirectly harmonizes the consistency of the two tasks in the student model under the expert strategy, and thus indirectly aligns the robustness objective with the natural accuracy objective of the student model. Intuitively, all learning objectives of the student model either directly or indirectly treat the natural teacher's output on clean samples as an oracle. Such a unified objective not only elegantly optimizes the trade-off relationship, but also avoids a suboptimal optimization target that would arise from directly aligning the natural and adversarial outputs of the student model (which would introduce new conflicts with the expert strategy). This cyclic consistency can be intuitively visualized in Figure 3.

2.3 Overall

Integrating the three modules above yields the complete model of our proposed GCMRD, whose framework can be intuitively illustrated in Figure 2.

3 Experiments

3.1 Experimental Settings

Datasets. We conduct our experiments on three datasets: CIFAR-10/100 and Tiny-ImageNet, which is the common setting for robustness distillation(12; 23).

Models. The selection of the teacher-student models remains consistent with existing work (12; 23; 24). For the student model, we consider two backbones including ResNet-18 (9) and MobileNet-V2 (18) for CIFAR10/100, and PreActResNet-18 (10) for Tiny-ImageNet. For the teacher model, we choose ResNet-56 (as the clean teacher) and WideResNet-34-10 (21) (as the robust teacher) for CIFAR-10, WideResNet-22-6 (as the clean teacher) and WideResNet-70-16 (as the robust teacher) for CIFAR-100, and PreActResNet-34 (10) for Tiny-ImageNet. For CIFAR-10, WideResNet-34-10 is trained using TRADES. For CIFAR-100, we use the WideResNet-70-16 model provided by Gowal et al (8); In Addition, for Tiny-ImageNet, we use PreActResNet-34 trained by TRADES (22) as the robust teacher. The performance of these teacher models is shown in Table 1.

Training settings. 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 128 and standard data augmentation (random cropping and horizontal flipping).

Evaluation settings. Following existing studies (12; 23; 24; 27), we evaluate the models against white box adversarial attacks: FGSM (7), PGD_{SAT} (14), $\text{PGD}_{\text{TRADES}}$ (22), CW_∞ (3) and AutoAttack (4), which are commonly used adversarial attacks in adversarial robustness evaluation. The step sizes of PGD_{sat} and $\text{PGD}_{\text{trades}}$ are $2/255$ and 0.003, respectively, and the step is 20. The total step of CW_∞ is 30. Maximum perturbation is bounded to the L_∞ norm $\epsilon = 8/255$ for all attacks. We also provide the black-box robustness evaluation, which includes the transfer-based attack and query-based attack. As for the transfer-based attack, we choose robust teachers (WideResNet-34-10 for CIFAR-10, WideResNet-70-16 for CIFAR-100, and PreActResNet-34 for Tiny-ImageNet) as the surrogate models to produce adversarial examples against the $\text{PGD}_{\text{trades}}$ and CW_∞ attack; As for the query-based attack, we choose a score-based attack: Square Attack (1) where the queries are 100.

3.2 Main Results

White-box robustness. The white-box results of CIFAR-10/100 on ResNet-18 are shown in Table 2. Results of CIFAR-10/100 on MobileNet-v2 and Tiny-ImageNet on PreActResNet-18 are provided in the Appendix. Further evaluation on Autoattack is also provided in the Appendix. GCMRD achieves the best performance in both natural accuracy and robustness. This demonstrates that GCMRD not only achieves better robustness but also optimizes the trade-off between robustness and accuracy in the student model.

Black-box robustness. Following the common setup (12; 23), we also evaluate the black-box robustness of GCMRD. The experimental results for CIFAR10/100 on ResNet-18 are shown in Table 3 and results of CIFAR-10/100 on

Table 2: White-box robustness of ResNet-18 on CIFAR-10 and CIFAR-100. The optimal results are highlighted in bold.

Attack	Defense	CIFAR10			CIFAR100		
		Clean	Robust	W-Robust	Clean	Robust	W-Robust
FGSM	Natural	94.57%	18.60%	56.59%	75.18%	7.96%	41.57%
	SAT	84.2%	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%	58.4%	71.26%	60.11%	33.61%	46.86%
	RSLAD	83.99%	60.41%	72.2%	58.25%	34.73%	46.49%
	MTARD	87.36%	61.2%	74.28%	64.3%	31.49%	47.90%
	B-MTARD	88.20%	61.42%	74.81%	65.08%	34.21%	49.65%
	CIARD	88.87%	61.88%	75.38%	65.73%	34.47%	50.10%
	GCMRD(Ours)	**89.26%**	**61.99%**	**75.32%**	**65.81%**	**35.42%**	**50.32%**
PGD_{SAT}	Natural	94.57%	0%	47.29%	75.18%	0%	37.59%
	SAT	84.2%	45.95%	65.08%	56.16%	21.18%	38.67%
	TRADES	83.00%	52.35%	67.68%	57.75%	28.05%	42.9%
	ARD	84.11%	50.93%	67.52%	60.11%	29.4%	44.76%
	RSLAD	83.99%	53.94%	68.91%	58.25%	31.19%	44.72%
	MTARD	87.36%	50.73%	69.05%	64.3%	24.95%	44.63%
	B-MTARD	88.20%	51.68%	69.94%	65.08%	28.50%	46.79%
	CIARD	88.87%	51.70%	70.29%	65.73%	28.05%	46.89%
	GCMRD(Ours)	**89.26%**	**52.84%**	**71.05%**	**65.81%**	**28.84%**	**47.33%**
PGD_{TRADES}	Natural	94.57%	0%	47.29%	75.18%	0%	37.59%
	SAT	84.2%	48.12%	66.16%	56.16%	22.02%	39.09%
	TRADES	83.00%	53.83%	68.42%	57.75%	28.88%	43.32%
	ARD	84.11%	52.96%	68.54%	60.11%	30.51%	45.31%
	RSLAD	83.99%	55.73%	69.86%	58.25%	32.05%	45.15%
	MTARD	87.36%	53.60%	70.48%	64.3%	26.75%	45.53%
	B-MTARD	88.20%	54.40%	71.30%	65.08%	29.94%	47.51%
	CIARD	88.87%	54.46%	71.67%	65.73%	29.45%	47.59%
	GCMRD(Ours)	**89.26%**	**55.01%**	**72.14%**	**65.81%**	**30.41%**	**48.11%**
CW_∞	Natural	94.57%	0%	47.29%	75.18%	0%	37.59%
	SAT	84.2%	45.97%	65.09%	56.16%	20.9%	38.53%
	TRADES	83.00%	50.23%	66.62%	57.75%	24.19%	40.97%
	ARD	84.11%	50.15%	67.13%	60.11%	27.56%	43.84%
	RSLAD	83.99%	52.67%	68.33%	58.25%	28.21%	43.23%
	MTARD	87.36%	48.57%	67.97%	64.3%	23.42%	43.86%
	B-MTARD	88.20%	49.88%	69.04%	65.08%	25.45%	45.27%
	CIARD	88.87%	50.61%	69.74%	65.73%	24.43%	45.08%
	GCMRD(ours)	**89.26%**	**51.42%**	**70.34%**	**65.81%**	**25.89%**	**45.85%**

MobileNet-v2 and Tiny-ImageNet on PreActResNet-18 are provided in the Appendix. In short, GCMRD achieves the best robustness under black-box attacks, demonstrating its transferable adversarial robustness.

3.3 Ablation Analysis

In this subsection, we analyze the effectiveness of each module we propose, which mainly includes three parts: MCIM, FRR, and CFT. Taking CIARD (12) as our baseline method, we discuss the effectiveness of each module when added individually. The experimental results are shown in Table 4. In short, the internal maximization involving dual teachers enables the student model to explore a richer set of adversarial behavior patterns, leading to a significant improvement

Table 3: Black-box robustness of ResNet-18 on CIFAR-10 and CIFAR-100. The optimal results are highlighted in bold.

Attack	Defense	CIFAR10			CIFAR100		
		Clean	Robust	W-Robust	Clean	Robust	W-Robust
PGD_{TRADES}	SAT	84.2%	64.74%	74.47%	56.16%	38.1%	47.13%
	TRADES	83.00%	63.56%	73.28%	57.75%	38.2%	47.98%
	ARD	84.11%	63.59%	73.85%	60.11%	39.53%	49.82%
	RSLAD	83.99%	63.9%	73.95%	58.25%	39.93%	49.09%
	MTARD	87.36%	65.17%	76.27%	64.3%	41.39%	52.85%
	B-MTARD	88.20%	65.29%	76.75%	65.08%	42.11%	53.60%
	CIARD	88.87%	66.28%	77.58%	65.73%	42.29%	54.01%
	GCMRD(Ours)	**89.26%**	**67.24%**	**78.25%**	**65.81%**	**43.12%**	**54.47%**
CW_∞	SAT	84.2%	64.88%	74.54%	56.16%	39.42%	47.79%
	TRADES	83.00%	62.85%	72.93%	57.75%	38.63%	48.19%
	ARD	84.11%	62.78%	73.45%	60.11%	38.85%	49.48%
	RSLAD	83.99%	63.02%	73.51%	58.25%	39.67%	48.96%
	MTARD	87.36%	64.65%	76.01%	64.3%	41.18%	52.74%
	B-MTARD	88.20%	64.64%	76.42%	65.08%	41.35%	53.22%
	CIARD	88.87%	64.79%	76.83%	65.73%	41.44%	53.59%
	GCMRD(Ours)	**89.26%**	**65.42%**	**77.34%**	**65.81%**	**42.27%**	**54.04%**
SA	SAT	84.2%	71.3%	77.75%	56.16%	41.27%	48.72%
	TRADES	83.00%	70.33%	76.67%	57.75%	41.96%	49.86%
	ARD	84.11%	73.3%	78.71%	60.11%	48.79%	54.45%
	RSLAD	83.99%	72.1%	78.05%	58.25%	45.34%	51.80%
	MTARD	87.36%	79.99%	83.68%	64.3%	48.13%	56.22%
	B-MTARD	88.20%	79.82%	84.01%	65.08%	49.40%	57.24%
	CIARD	88.87%	80.03%	84.45%	65.73%	49.76%	57.75%
	GCMRD(Ours)	**89.26%**	**81.16%**	**85.21%**	**65.81%**	**50.01%**	**57.91%**

in robustness. FRR provides appropriate negative supervision signals for both natural and adversarial tasks, thereby providing increments for both accuracy and robustness. Meanwhile, the natural teacher-guided fine-tuning of the robust teacher not only improves the performance of the robust teacher, resulting in further robustness increments, but also aligns the natural teacher and robust teacher, indirectly alleviates the optimization conflict between the clean task and adversarial task of the student model, and achieves a better trade-off between accuracy and adversarial robustness. Such results of the ablation study align with the intended design and expectations of the proposed modules.

3.4 Qualitative Analysis

Visualization of feature distribution. On the CIFAR10 dataset, we provide feature distribution maps based on UMAP (15) and compare them with those of the natural teacher and robust teacher. Intuitively, adversarial attacks cause the feature distribution of the natural teacher to become completely disordered, leading to incorrect classification. In contrast, our method maintains a discriminative feature distribution similar to that of the robust teacher under the threat of adversarial attacks. This provides interpretability for adversarial robustness from the perspective of feature distribution.

Table 4: Ablation Study on CIFAR-10/100 with ResNet-18. The optimal results are highlighted in bold.

Modules	CIFAR10					CIFAR100				
	Clean	FGSM	PGD_S	PGD_T	CW_∞	Clean	FGSM	PGD_S	PGD_T	CW_∞
Baseline	88.87%	61.88%	51.70%	54.46%	50.61%	65.73%	34.47%	28.05%	29.45%	24.43%
+MCIM	86.42%	61.90%	52.60%	54.84%	51.30%	63.27%	35.38%	28.74%	30.14%	25.76%
+FRR	87.06%	61.94%	52.75%	55.01%	51.06%	64.94%	35.22%	28.69%	30.22%	25.72%
+CFT	**89.37%**	61.87%	51.90%	54.90%	50.85%	**65.89%**	34.30%	28.10%	29.62%	24.62%
GCMRD	89.26%	**61.99%**	**52.84%**	**55.01%**	**51.42%**	65.81%	**35.42%**	**28.84%**	**30.41%**	**25.89%**

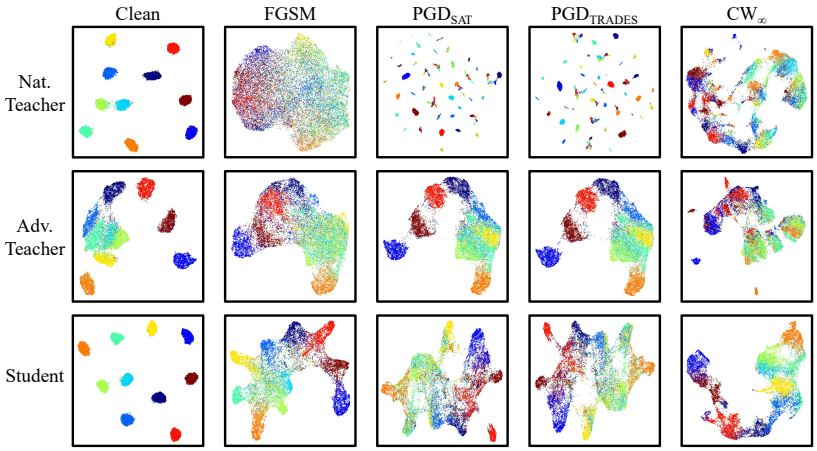
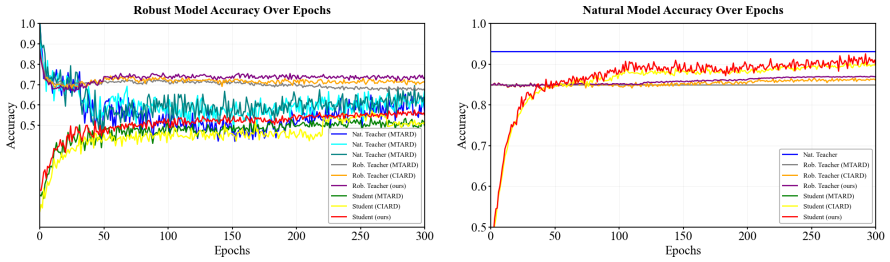


Fig. 4: Visualization of feature distribution based on UMAP. Here, data points of each color represent samples from the same category.

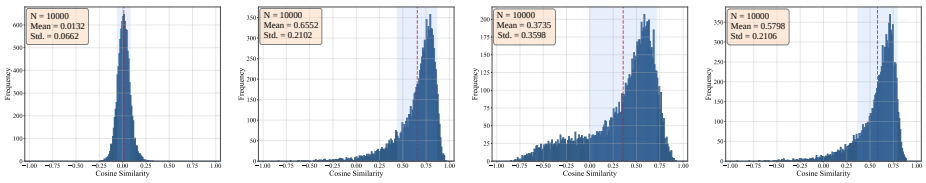
Task consistency analysis. As analyzed, the inherent structural differences between the two teachers can lead to optimization conflicts between the natural task and the adversarial task in the student model under the expert supervision framework. We use the gradient angle as a metric to describe task consistency, demonstrating that the GCMRD induces consistency between the two teachers, which in turn aligns the gradient directions of the natural and adversarial tasks, thereby explaining why GCMRD achieves a better trade-off between the accuracy and adversarial robustness.

Specifically, on CIFAR-10, we compute the cosine similarity between the natural task gradient and the adversarial task gradient on each sample for natural teacher, robust teacher, and student model in CIARD and our GCMRD. Intuitively, an acute angle (positive cosine value) between the natural and robust task gradients indicates better synergy between the two tasks, while an obtuse angle reflects a significant conflict in optimization direction. Results in Figure 6 shows that, the similarity between the natural teacher's natural task and robust task is the worst, with gradient angles for most samples being nearly orthogonal. The robust teacher exhibits the best gradient similarity. In comparison,



(a) Robustness curve **(b)** Accuracy curve

Fig. 5: Training curves of ResNet-18 on CIFAR-10.



(a) Nat. teacher **(b)** Rob. teacher **(c)** CIARD **(d)** GCMRD

Fig. 6: Gradient angle analysis between the natural task and the adversarial task.

GCMRD demonstrates better gradient similarity than the current state-of-the-art method, CIARD. This implies that a larger number of samples have coordinated natural and adversarial gradient directions, thereby achieving a better accuracy-robustness trade-off.

Analysis on the training curves. We further analyze the training curves of the teacher and student model in the proposed GCMRD, and provide an intuitive comparison with MTARD and CIARD. The training curves are shown in Figure 5. For the robust accuracy, the differences among the natural teachers in the three methods are subtle. However, the robust teacher of MTARD, being fixed, exhibits a decline in robustness after a certain stage. The fine-tuning strategy in CIARD alleviates this downward trend. GCMRD further enhances the robustness of the robust teacher in the later stages, and the student model in GCMRD learns better robustness. For the natural accuracy, the natural accuracy of the natural teachers across the three methods and the robust teacher of MTARD remains fixed. The natural accuracy of the robust teacher in CIARD shows a slight increase during training, while GCMRD achieves additional increments from soft labels. As a result, the natural accuracy of the student model is also enhanced. Such results once again confirm that our GCMRD reconciles the contradiction between natural and adversarial tasks in the student model through global consistency, achieving a better trade-off.

Attention map. To further verify the quantitative interpretability of GCMRD, we also provide attention heatmaps based on Grad-CAM (19), as shown in Figure 7. For each class, from left to right, the columns correspond to: clean samples,

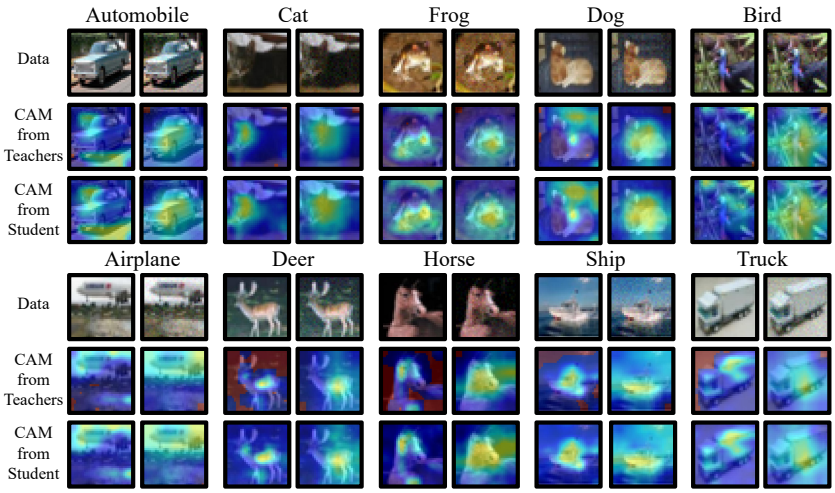


Fig. 7: Attention heatmap based on Grad-CAM of CIFAR-10. The warm-colored highlighted regions indicate the parts to which the model pays greater attention.

adversarial samples, attention maps of the natural teacher on clean samples, attention maps of the robust teacher on adversarial samples, attention maps of the student model on clean samples, and attention maps of the student model on adversarial samples. On the one hand, the attention maps of the student model on clean samples are largely similar to those of the natural teacher on clean samples, while the attention maps of the student model on adversarial samples closely resemble those of the robust teacher on adversarial samples, demonstrating the effectiveness of the expert strategy. On the other hand, we also find that the attention maps of the student model on natural samples and adversarial samples exhibit a certain degree of similarity, which again validates the preference for global consistency in GCMRD. In summary, Grad-CAM provides an interpretable perspective on the robustness of GCMRD from the viewpoint of model preferences.

4 Conclusion

In this paper, we propose a novel dual-teacher adversarial training framework named global consistency multi-teacher robustness distillation, designed to fully exploit the potential of dual-teacher frameworks and achieve global consistency for a better accuracy-robustness trade-off. Specifically, GCMRD consists of three core components: Multi-teacher collaboration inner maximization, full-repulsion regularization, and consistency fine-tuning of the robust teacher. Multi-teacher collaboration inner maximization explores a richer set of internal maximization patterns, full-repulsion regularization leverages negative supervision regularization to enhance the student model's generalization capability, and consistency

fine-tuning of the robust teacher indirectly reconciles the dual-task conflict of the student model by aligning with the teacher model. These three modules together achieve the global consistency of GCMRD. Experiments show that GCMRD achieves the best performance in both adversarial robustness and 'accuracy-robustness' trade-off, expanding the prospects for lightweight robust models.

5 Acknowledgments

This work was supported by the National Cyber Security-National Science and Technology Major Project under Grant No. 2026ZD1500500, the Key Program of the National Natural Science Foundation of China under Grant No. 62532016, the National Natural Science Foundation of China under Grant No. 62572150, the Guangdong Basic and Applied Basic Research Foundation under Grant No. 2026B1515020033, the National Social Science Fund of China under Grant No. 23VRC094, and the Major Program of the National Social Science Fund of China under Grant No. 22&ZD147.

Bibliography

[1] Andriushchenko, M., Croce, F., Flammarion, N., Hein, M.: Square attack: a query-efficient black-box adversarial attack via random search. In: European conference on computer vision. pp. 484–501. Springer (2020)

[2] Athalye, A., Carlini, N., Wagner, D.: Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples. In: International conference on machine learning. pp. 274–283. PMLR (2018)

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

[4] Croce, F., Hein, M.: Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks. In: International conference on machine learning. pp. 2206–2216. PMLR (2020)

[5] Dong, Y., Liao, F., Pang, T., Su, H., Zhu, J., Hu, X., Li, J.: Boosting adversarial attacks with momentum. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 9185–9193 (2018)

[6] Goldblum, M., Fowl, L., Feizi, S., Goldstein, T.: Adversarially robust distillation. In: Proceedings of the AAAI conference on artificial intelligence. vol. 34, pp. 3996–4003 (2020)

[7] Goodfellow, I.J., Shlens, J., Szegedy, C.: Explaining and harnessing adversarial examples. arXiv preprint arXiv:1412.6572 (2014)

[8] Gowal, S., Qin, C., Uesato, J., Mann, T., Kohli, P.: Uncovering the limits of adversarial training against norm-bounded adversarial examples. arXiv preprint arXiv:2010.03593 (2020)

[9] He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)

[10] He, K., Zhang, X., Ren, S., Sun, J.: Identity mappings in deep residual networks. In: European conference on computer vision. pp. 630–645. Springer (2016)

[11] Jia, X., Zhang, Y., Wu, B., Ma, K., Wang, J., Cao, X.: Las-at: adversarial training with learnable attack strategy. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 13398–13408 (2022)

[12] Lu, L., Pang, S., Zheng, X., Gu, X., Du, A., Liu, Y., Zhou, Y.: Ciard: Cyclic iterative adversarial robustness distillation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 350–359 (2025)

[13] Ma, H., Chen, T., Hu, T.K., You, C., Xie, X., Wang, Z.: Undistillable: Making a nasty teacher that cannot teach students. arXiv preprint arXiv:2105.07381 (2021)

[14] Madry, A., Makelov, A., Schmidt, L., Tsipras, D., Vladu, A.: Towards deep learning models resistant to adversarial attacks. arXiv preprint arXiv:1706.06083 (2017)
[15] McInnes, L., Healy, J., Melville, J.: Umap: Uniform manifold approximation and projection for dimension reduction. arXiv preprint arXiv:1802.03426 (2018)
[16] Papernot, N., McDaniel, P., Wu, X., Jha, S., Swami, A.: Distillation as a defense to adversarial perturbations against deep neural networks. In: 2016 IEEE symposium on security and privacy (SP). pp. 582–597. IEEE (2016)
[17] Rice, L., Wong, E., Kolter, Z.: Overfitting in adversarially robust deep learning. In: International conference on machine learning. pp. 8093–8104. PMLR (2020)
[18] Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., Chen, L.C.: Mobilenetv2: Inverted residuals and linear bottlenecks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 4510–4520 (2018)
[19] Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D.: Grad-cam: Visual explanations from deep networks via gradient-based localization. In: Proceedings of the IEEE international conference on computer vision. pp. 618–626 (2017)
[20] Xie, C., Wang, J., Zhang, Z., Ren, Z., Yuille, A.: Mitigating adversarial effects through randomization. arXiv preprint arXiv:1711.01991 (2017)
[21] Zagoruyko, S., Komodakis, N.: Wide residual networks. arXiv preprint arXiv:1605.07146 (2016)
[22] Zhang, H., Yu, Y., Jiao, J., Xing, E., El Ghaoui, L., Jordan, M.: Theoretically principled trade-off between robustness and accuracy. In: International conference on machine learning. pp. 7472–7482. PMLR (2019)
[23] Zhao, S., Wang, X., Wei, X.: Mitigating accuracy-robustness trade-off via balanced multi-teacher adversarial distillation. IEEE transactions on pattern analysis and machine intelligence **46**(12), 9338–9352 (2024)
[24] Zhao, S., Yu, J., Sun, Z., Zhang, B., Wei, X.: Enhanced accuracy and robustness via multi-teacher adversarial distillation. In: European Conference on Computer Vision. pp. 585–602. Springer (2022)
[25] Zhou, Y., Hua, Z.: Defense without forgetting: Continual adversarial defense with anisotropic & isotropic pseudo replay. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 24263–24272 (2024)
[26] Zhu, J., Yao, J., Han, B., Zhang, J., Liu, T., Niu, G., Zhou, J., Xu, J., Yang, H.: Reliable adversarial distillation with unreliable teachers. arXiv preprint arXiv:2106.04928 (2021)
[27] Zi, B., Zhao, S., Ma, X., Jiang, Y.G.: Revisiting adversarial robustness distillation: Robust soft labels make student better. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 16443–16452 (2021)