Don't Teach Instability, Teach Robustness: Selective Sensitivity Gating for Adversarial Robust Distillation

Jingqi Ji, Quan Kong, Chaojie Gu, Yuanchao Shu, and Cong Wang*

Zhejiang University, Hangzhou, China
{jijingqi, qkv, gucj, ycshu, cwang85}@zju.edu.cn

Abstract. Adversarial Robustness Distillation (ARD) is the most viable way for transferring defensive capabilities from over-parameterized teachers to lightweight student models. However, existing ARD techniques rely on the assumption that the teacher's decision boundary is a perfect geometric oracle. In this paper, we identify a critical gap from the stability perspective that such exact matching forces blind inheritance of teacher's sensitivity noise, punishing the student's stability even if they are more robust. To address this challenge, we propose the AEGIS (Adversarial Error Gating and Instability Suppression) framework that replaces simple logit matching with gated supervision. It leverages a pair of logit and sensitivity gates to rectify potential errors/instability on the categorical and gradient-level, respectively. Extensive evaluations on CIFAR-10/100, and Tiny-ImageNet show that AEGIS significantly outperforms state-of-the-art methods, achieving a 3.12% gain under the strong AutoAttack and preserving nearly 50% accuracy under large perturbation intensities ($\epsilon = 16/255$) where traditional baselines collapse. Code is available at https://github.com/JingqiJi03/AEGIS.

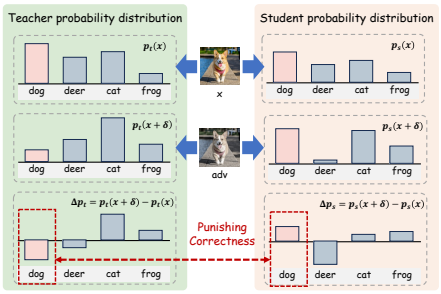
Keywords: Adversarial Robustness Distillation · Knowledge Distillation · Adversarial Attack

1 Introduction

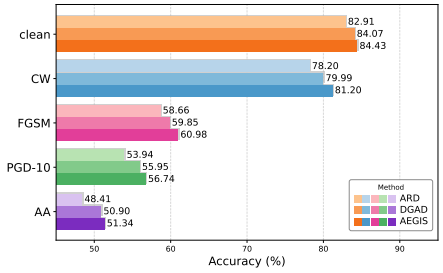
Adversarial Robustness Distillation (ARD) has emerged as one of the most effective techniques to enable mobile-friendly lightweight models with defensive capacity from adversarially-trained, over-parameterized models [8, 14, 18, 50, 51]. By forcing a student model to mimic the soft output distributions [8,14,50,51] or local gradient landscapes [18] of a robust teacher, ARD seeks to transfer the stable decision boundaries from the teacher [27]. Conventionally, the success of this process is measured by the degree to which the student matches the teacher's output probability.

Unfortunately, the existing ARD works rely on a risky assumption that the teacher's decision boundary is always optimal. This leaves a gap from a large

* Corresponding author.



(a) Illustration of Stability Punishment.



(b) Robustness comparison against various attacks.

Fig. 1: Analysis of stability punishment and potential performance. (a) Illustration of how the teacher's target-class probability variation under adversarial attacks could act as sensitivity noise to mislead the student model in distillation; (b) Comparison of distillation methods under various adversarial attacks.

body of parallel works in standard Knowledge Distillation (KD), which challenge the fundamental presumption whether the student should exactly match the teacher's probabilities [20, 25, 34]. E.g., Stanton et al. [34] show that students with tighter alignment may generalize worse than students that deviate from the teacher; Nagarajan et al. [25] demonstrate that students can achieve superior performance by amplifying the implicit biases of gradient descent rather than strictly following the teacher's probabilities. Meanwhile, another thread of early attempts also explored mutual [2,10,47] and self-distillation [44,45] that students learn from each other or even themselves without a teacher. These insights imply that the teacher is not an absolute oracle in standard KD, so how about ARD?

In ARD, beyond predicting the correct label, the student must match the teacher's stability under adversarial perturbations. Thus, it faces a more complex task of aligning the local geometry of adversarially-learned boundaries, which are often jagged with local variation and instability (possibly due to robust overfitting [32, 41]). Thus, blindly aligning the student's output/gradient to match this sensitivity from the teacher is deceptive, which leads to a phenomenon we call *Stability Punishment* as illustrated in Fig. 1a. Consider a scenario where the teacher correctly classifies an image but exhibits high instability, with its confidence dropping significantly under adversarial attack (e.g., $|\Delta p_t| \gg 0$). Nevertheless, the student might be more stable and resist this drop ($|\Delta p_s| \approx 0$), whereas standard gradient alignment [18] blindly punishes the student for this stability by forcing it to mimic the teacher's large confidence drop. This would inadvertently inject the teacher's instability noise into the student. Therefore, the ARD targets must be refined to not only correct static errors (wrong labels) [17] but also filter out more fine-grained *dynamic sensitivity* failures.

To bridge this gap missing in the current ARD works, we propose Adversarial Error Gating and Instability Suppression (AEGIS). Unlike previous methods that treat ARD as a single stream [8, 14, 18, 50, 51], we separate the distillation process into different paths, based on the prediction decisions and dynamic

sensitivity of the teacher. By leveraging *gating* as a dynamic conditioning operator, we introduce two gates to act as robust knowledge filters at different levels. First, on the logit level, the Semantic Logit Gate (SLG) utilizes a dynamic reliability mask to assess the teacher's correctness on each sample. If the teacher is wrong, the gate switches the target back to clean logits to prevent error propagation. Second, on the more fine-grained geometric level, the Sensitivity Calibration Gate (SCG) branches gradient-level supervision into: 1) regulating *target-class* alignment to filter out teacher instability to ensure the student only inherits valid robustness patterns; 2) performing *non-target-class* refinement by utilizing a masking strategy to filter out misleading classes where the teacher assigns higher confidence than the ground truth. These help selectively ignore the teacher when its local geometry is more volatile than the student, thereby mitigating potential instability. Our main contributions are summarized below:

- ✦ We identify a gap in current ARD techniques where forcing a student to align with the teacher's local manifold may inadvertently impair the student's geometric stability. This shifts the distillation targets to more fine-grained analysis of differential probability under adversarial attacks.
- ✦ We propose a selective gating framework to decouple knowledge transfer for both class-level and geometric alignments. Unlike previous ARD methods that are always on, we dynamically detach the student from unreliable teacher supervisions, such that only constructive dark knowledge is distilled.
- ✦ We conduct extensive experiments across CIFAR10/100 and Tiny-ImageNet and the results show that AEGIS consistently outperforms SOTA ARD techniques, achieving a notable 3.12% gain in robust accuracy under the strong AutoAttack (AA) [6]. AEGIS also demonstrates remarkable sustainability of preserving nearly 50% accuracy under large perturbation ($\epsilon = 16/255$) where traditional methods collapse below 40%.

2 Related Work

2.1 Standard Knowledge Distillation

KD is a process of transferring the dark knowledge from a large teacher to a small student model [13], which can be categorized into logit-based [13, 21, 35, 39, 48], feature-based [3, 5, 16, 37, 42], and relation-based approaches [22, 30, 31]. While these variants aim to maximize the fidelity of knowledge transfer, a growing body of research begins to challenge the fundamental assumption of perfect alignment [25, 34]. Specifically, Stanton et al. [34] demonstrate that even when a student has sufficient capacity to perfectly match a teacher, a significant fidelity gap persists due to student's small capacity and optimization difficulties. Nagarajan et al. [25] further delve into this skepticism by arguing students can outperform their teachers by exaggerating confidence levels and amplifying the implicit biases of gradient descent. ATS [20] discovers that complex teachers tend to be over-confident with less discriminative wrong class probabilities.

To tackle discrepancies of teacher errors, existing works have proposed techniques such as dynamic temperature adjustment [21,35], logits refinement through ground-truth integration [17, 36, 39], loss reformulation [46, 48] and teacher ensembles [7, 26]. However, these rectification strategies largely operate on the global logit level. They fail to account for fine-grained information such as local boundary sensitivity, which becomes critical in adversarial environments where the teacher's gradient noise can actively mislead the student.

2.2 Adversarial Robustness Distillation

Adversarial training is the most robust defense against adversarial attacks so far, yet its efficacy is often bottlenecked by the model's capacity [23, 28]. To bridge the performance gap in lightweight models, ARD [8] is introduced to leverage high-capacity teachers. Initial advancements focus on refining logit-level supervision: RSLAD [51] utilizes robust soft labels to guide the adversarial generation process, while DGAD [29] introduces error-corrective label swapping to rectify teacher misclassifications. Recognizing that teacher reliability is sample-dependent, IAD [50] and AdaAD [14] propose adaptive mechanisms where students only partially trust the teacher's signals based on confidence scores or inner-maximization difficulty. PeerAiD [15] further refines this by employing a specialized peer tutor to mitigate the distribution shift between student and teacher adversarial samples. MTARD [49] introduces a dual-teacher framework to improve the robustness-accuracy trade-off in the student model.

Recent works have further looked into the geometric and behavioral attributes of the teacher. SmaraAD [40] aligns attribution regions to help students inherit the teacher's decision logic, and IGDM [18] matches input gradients to transfer local smoothness directly. Nevertheless, the existing methods generally operate under the assumption that the local sensitivity of robust teacher is a high-fidelity oracle and focus on fixing what the teacher predicts (class labels). In this paper, we argue that this blind alignment fails to account for the local instabilities in adversarially-trained models [32, 41], which might lead to propagation of errors through the magnitude of sensitivity. We identify the resulting stability punishment and propose a framework to selectively filter erroneous supervision at both the logit and differential logit levels.

3 Preliminaries

Notations. We consider a standard classification task with a dataset $\mathcal{D} = \{(x, y)\}_{i=1}^{N}$, where $x \in \mathbb{R}^d$ is the input image and $y \in \{1, \ldots, K\}$ is the ground-truth label. Let $f(\cdot;\theta) : \mathbb{R}^d \to \mathbb{R}^K$ denote a neural network parameterized by θ that maps an input x to logit vector $z = f(x)$. The predicted probability distribution is obtained via the softmax function with a temperature τ: $p(x;\tau) = \frac{\exp z_k/\tau}{\sum_j \exp(z_j/\tau)}$. We denote the fixed pre-trained teacher as f_T and the student as f_S.

Adversarial Robustness Distillation (ARD) [8]. ARD aims to transfer the teacher's robust decision boundaries to the student. Similar to AT [23], it is also formulated as a min-max optimization problem,

$$\mathcal{L}_{\text{ARD}} = \min_{\theta} \mathbb{E}_{(x,y)\sim\mathcal{D}} \left[(1-\alpha) \cdot \mathcal{L}_{CE}(p_S(x), y) + \alpha\tau^2 \cdot \mathcal{L}_{KL}(p_S(x_{adv}, \tau) || \mathcal{T}) \right], \quad (1)$$

where $\mathcal{L}_{CE}$, $\mathcal{L}_{KL}$ are the cross-entropy and Kullback-Leibler losses. α balances the clean accuracy and robustness transfer, and $\mathcal{T}$ represents the distillation target. The adversarial example $x_{adv} = x + \delta$ is generated in the inner loop by maximizing a specific type of attack objective within an ϵ-ball constraint $||\delta||_p \leq \epsilon$.

Adaptive Adversarial Distillation (AdaAD) [14]. A key limitation of vanilla ARD is its reliance on a static teacher. AdaAD introduces a dynamic alignment strategy where the teacher is actively involved in the inner-maximization process. Instead of using a standard PGD attack, the inner loop identifies the hardest perturbation that maximizes the discrepancy between the student and the teacher,

$$\begin{aligned} \mathcal{L}_{\text{AdaAD}} = & \underset{\theta}{\text{argmin}} \, (1-\gamma) \cdot \mathcal{L}_{KL}(f_S(x) || f_T(x)) + \gamma \cdot \mathcal{L}_{KL}(f_S(x_{adv}) || f_T(x_{adv})), \\ & \text{s.t.} \qquad x_{adv} = \underset{||\delta||_p < \epsilon}{\text{argmax}} \, \mathcal{L}_{KL}(f_S(x+\delta), f_T(x+\delta)). \end{aligned} \quad (2)$$

By tracking the teacher's probability shifts under extreme perturbations, AdaAD encourages the student to replicate the teacher's manifold. However, as shown next, this objective implicitly assumes the teacher's local shift is always a valid target, while in fact, the teacher's manifold could be locally unstable.

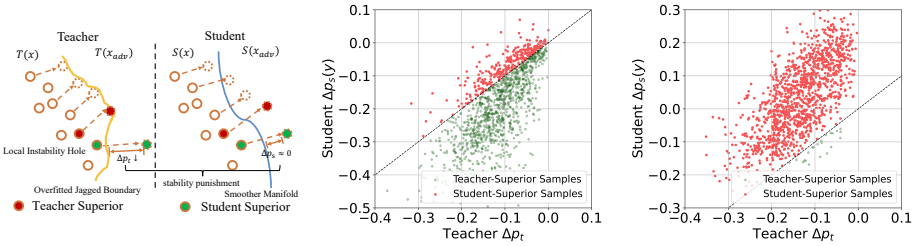
4 The Stability Punishment Phenomenon

In this section, we empirically demonstrate that ARD often ignores the geometric discrepancy between the teacher and student decision boundaries. As illustrated in Fig. 2a, high-capacity teachers frequently exhibit complex boundaries due to robust overfitting [32, 41], where the local boundaries are often more jagged. In contrast, the limited capacity of student models acts as an implicit regularizer, which leads to smoother manifolds in scattered local regions. Before the analysis, we illustrate two formal definitions below.

Definition 1 (Stability Punishment). It happens while forcing a student to align with the teacher's logits. If the target sample lies in a region with localized instability of the teacher, it injects sensitivity noise into the student who could have achieved a more robust local geometry than its teacher.

Definition 2 (Dynamic Probability Variation). It is defined as $\Delta p = p(x+\delta) - p(x)$. We use it as a proxy to characterize the local sensitivity of a model's manifold under perturbation δ.

Empirical Validation. We visualize the relationship between teacher stability (Δp_t) and student stability (Δp_s) across 2K test samples in Fig. 2b and



(a) Stability Punishment **(b)** Exact Logits Matching **(c)** Conditional Matching

Fig. 2: Empirical Analysis of Stability Punishment. We visualize the dynamic sensitivity (Δp) of teacher and student models across 2K test samples, where the diagonal $y = x$ represents the robustness parity. Red dots denote *Student-Superior Samples* ($\Delta p_s > \Delta p_t$), where the student exhibits higher adversarial stability than the teacher and green dots indicate teacher superiority. (a) Illustration of Stability Punishment due to boundary dynamics; (b) Exact Logit Matching forces the student to mimic the teacher's probabilities; (c) Conditional Logit Matching selectively filters the teacher's sensitivity noise resulting in a dramatic surge of Student-Superior samples.

Fig. 2c. In these scatter plots, the diagonal line $y = x$ serves as a threshold. That is, samples in the upper-left region ($\Delta p_s > \Delta p_t$) represent when the student is superior than the teacher and vice versa. Our goal is to achieve a more positive Δp since a more negative one stands for higher variance around $p(x + \delta)$.

When the student is forced to mirror the teacher's sensitivity (Fig. 2b), only 22.19% of samples achieves superior stability. In contrast, by filtering out unreliable logits, the proportion of red points surges to 97.75%. This demonstrates that with conditional logits matching, the student can maintain, or even enhance its local robustness without being penalized by the teacher. These empirical findings suggest that the standard KL-based alignment of gradients or logits is biased when the models' local geometries diverge. To understand the conditions under which this punishment occurs and how to mitigate it, we provide a formal theoretical analysis in the next section.

5 Theoretical Analysis

In this section, we analyze why aligning models in the differential probability space (Δp) compensates standard alignment of unnormalized logit space (Δz) and how standard alignment acts as a de-regularizer when a student outperforms its teacher.

Sensitivity Comparison. First, standard logits are unbounded and their scales vary significantly across heterogeneous models, whereas probability variations are naturally normalized and bounded within $\Delta p_y \in [-1, 1]$. Let $p_y(x) = \sigma(z(x))_y$ be the softmax output for class y. For an adversarial perturbation δ, the change in probability is,

$$\Delta p_y(x, \delta) = p_y(x + \delta) - p_y(x) \approx \nabla_x p_y(x)^\top \delta \tag{3}$$

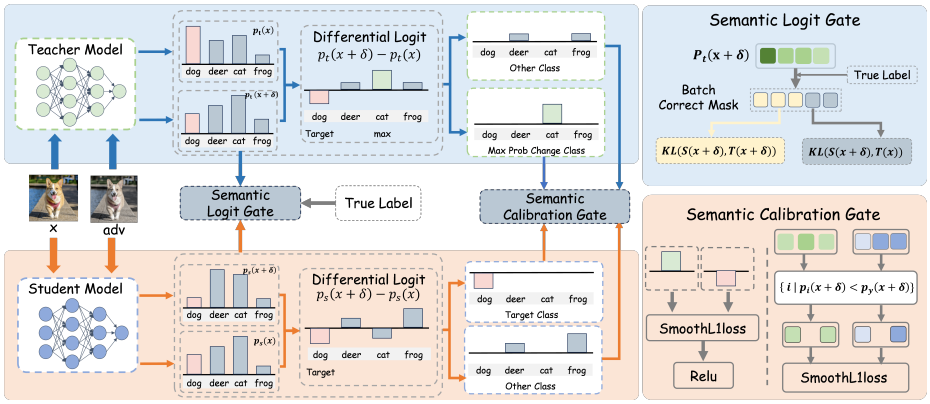


Fig. 3: Overview of AEGIS, which consists of two parallel gates to filter out inaccurate supervision. The Semantic Logit Gate (SLG) rectifies static prediction errors by switching distillation targets based on teacher correctness. The Sensitivity Calibration Gate (SCG) calibrates dynamic sensitivity by selectively aligning target-class gradients only when the student yields more stable probabilities under adversarial examples.

By the chain rule, $\nabla_x p_y = \left(\frac{\partial p_y}{\partial z}\right) \nabla_x z$. The term $\frac{\partial p_y}{\partial z}$ is the Jacobian of the softmax, which vanishes when the model is highly confident (approaching one-hot labels). **This means Δp naturally focuses alignment on uncertain samples near the decision boundary**, which is consistent with parallel works on improving KD efficiency by attending to samples with high entropy [38]. Nevertheless, standard logit alignment (Δz) continues to propagate gradient noise from the teacher even for easy samples where the student is already correct. Next, we formalize how it degrades student robustness.

Proposition 1 (Robustness De-regularization). *Consider the alignment loss $\mathcal{L}_{align} = \frac{1}{2}(\Delta p_S - \Delta p_T)^2$ [18]. If a student model is locally more robust than the teacher ($\Delta p_S > \Delta p_T$), the optimization of $\mathcal{L}_{align}$ forces the student to increase its local sensitivity, thereby degrading its own robustness.*

Proof. Considering the target-class probability under attack, we typically have decreased $\Delta p < 0$. A more robust model has a variation closer to zero. Let the student be more robust than the teacher, $\Delta p_T < \Delta p_S \leq 0$. The gradient of $\mathcal{L}_{\text{align}}$ with respect to the student's probability variation is,

$$\frac{\partial \mathcal{L}_{\text{align}}}{\partial \Delta p_S} = \Delta p_S - \Delta p_T > 0 \tag{4}$$

In a gradient descent step with learning rate η, the new variation becomes:

$$\Delta p'_S = \Delta p_S - \eta(\Delta p_S - \Delta p_T) \tag{5}$$

Since $(\Delta p_S - \Delta p_T) > 0$, we have $\Delta p'_S < \Delta p_S$. This drives the student's variation to be more negative that matches with the teacher's instability.

As shown by the directional derivative relation $|\Delta p_y| \leq \|\nabla p_y\|_2 \|\delta\|_2$, forcing Δp_S to become more negative is mathematically equivalent to forcing the student to increase its local gradient norm $\|\nabla p_S\|_2$. This actively injects the teacher's sensitivity noise into the student's geometry. Proposition 1 implies that a robust student should only follow the teacher when the teacher is more stable $|\Delta p_T| < |\Delta p_S|$, which lays the theoretical foundation for the AEGIS framework introduced next.

6 The AEGIS Framework

To mitigate the propagation of teacher-induced errors and stability noise, we propose the AEGIS ARD framework. As illustrated in Fig. 3, this framework partitions the distillation process into two levels: the Semantic Logit Gate (SLG) rectifies static prediction errors, and the Sensitivity Calibration Gate (SCG) calibrates the student's dynamic local sensitivity.

6.1 Semantic Logit Gate.

SLG is designed to prevent the student from learning incorrect predictions on adversarial samples. We define a *dynamic reliability mask*, G_{SLC} to evaluate whether the teacher's prediction on x_{adv} is semantically consistent with the ground truth y,

$$G_{\text{SLC}} = \mathbb{I}(\arg\max f_T(x_{adv}) = y), \tag{6}$$

where $\mathbb{I}(\cdot)$ is the indicator function. Using this mask, SLG dynamically redirects the distillation target. If the teacher is correct ($G_{\text{SLC}} = 1$), the student aligns with the teacher's adversarial logits to learn robust feature representations. However, if the teacher fails ($G_{\text{SLC}} = 0$), we adopt a fallback mechanism that pivots the student toward the teacher's stable, clean logits $f_T(x)$. This ensures the student is guided by reliable logits even when the teacher is fooled. The objective is formulated as,

$$\mathcal{L}_{\text{SLG}} = \mathbb{E}_x\big[G_{\text{SLC}} \cdot \mathcal{L}_{\text{KL}}\big(f_S(x_{adv}) \| f_T(x_{\text{adv}})\big) + (1 - G_{\text{SLC}}) \cdot \mathcal{L}_{\text{KL}}\big(f_S(x_{\text{adv}}) \| f_T(x)\big)\big]. \tag{7}$$

This reliability-weighted integration ensures that the student learns robust features from correct predictions while avoiding error propagation from the teacher.

6.2 Sensitivity Calibration Gate.

More importantly, SCG addresses the instability by calibrating the student's dynamic sensitivity. As analyzed in Sec. 5, aligning models in the probability variation space (Δp) allows for finer geometric control. We split this calibration into *target* and *non-target* components.

Target-Class Calibration. For the ground-truth class y, we aim to transfer the teacher's stability without inheriting its instability. In principle, adversarial perturbations push samples along the gradient toward the nearest decision

boundary in the feature space. Consequently, the non-target class with the maximum probability rise ($\max \Delta p_{\text{non-target}}$) serves as a precise indicator for this closest boundary. Hence, as opposed to the target class, non-target class with the maximum probability rise ($\max \Delta p_{\text{non-target}}$) represents the model's most critical vulnerability (the nearest decision boundary). By aligning the student's target drop (receding from the safety center) with the teacher's maximum non-target rise (approaching the error boundary), we impose a geometric constraint that the student's confidence should not degrade faster than the teacher's tolerance toward its most critical vulnerability:

$$\begin{aligned} G_{\text{TC}} &= \mathbb{I}\big((\Delta \mathbf{p}_S)_y < (\Delta \mathbf{p}_T)_{k^*}\big), \\ &\text{s.t. } k^* = \underset{k}{\text{argmax}} (\Delta \mathbf{p}_T)_k. \end{aligned} \tag{8}$$

Then the Target-class Robust Distillation (TRD) loss is formulated to suppress the student's sensitivity only when it exceeds the teacher's tolerance bound,

$$\mathcal{L}_{\text{TRD}} = G_{\text{TC}} \cdot \text{SmoothL1}\big((\Delta \mathbf{p}_S)_y - (\Delta \mathbf{p}_T)_{k^*}\big). \tag{9}$$

Gate Opening Ratios. To understand the status of the two gates function throughout training, we trace their opening ratios in Fig. 4. First, we see that the opening ratios of them decrease, but the SCG declines much more rapidly than the SLG. This trend reflects the learning process that the student initially relies on the teacher's guidance to establish basic robustness, but eventually achieves a level of stability that surpasses the teacher's local geometry. Second, to explain the interesting disparity of the decreasing rates between SLG and SCG, we reason through their different learning objectives. SLG handles high-level semantic correctness: matching the teacher's categorical decision boundaries is a more complex, long-term optimization task, corresponding to the gradual decline of the SLG curve. In contrast, SCG regulates local geometric sensitivity. Because the student has lower capacity, its architectural constraints act as an implicit regularizer, leading it to naturally converge toward a smoother local manifold much earlier in the training process. Once the student's local stability matches or exceeds the teacher, SCG closes to prevent gradient noise from the teacher.

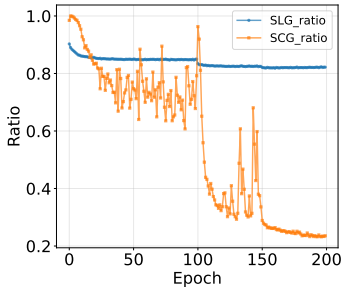


Fig. 4: Trace of gate activation of both SLG and SCG. Their decreasing opening ratios throughout training demonstrate selective adaptivity.

Non-Target Structural Knowledge. Next, to preserve the dark knowledge while filtering misleading information, we also identify beneficial non-target classes. A class k is considered valid only if the teacher remains more confident in

the ground truth than in class k. We define the multi-class gate $G_{NTC} \in \{0,1\}^K$,

$$G_{\mathrm{NTC}}^{(k)} = \mathbb{I}(p_k(x_{\mathrm{adv}}) < p_y(x_{\mathrm{adv}}) \wedge k \neq y). \tag{10}$$

The resulting Non-Target Robust Distillation (NTRD) loss aligns the structural variations between the two models,

$$\mathcal{L}_{\mathrm{NTRD}} = \frac{1}{\sum_k G_{\mathrm{NTC}}^{(k)} + \xi} \sum_{k=1}^{K} G_{\mathrm{NTC}}^{(k)} \cdot \mathrm{SmoothL1}\big((\Delta \mathbf{p}_S)_k, (\Delta \mathbf{p}_T)_k\big). \tag{11}$$

The total SCG objective is a weighted combination,

$$\mathcal{L}_{\mathrm{SCG}} = \beta \cdot \mathcal{L}_{\mathrm{TRD}} + (1-\beta) \cdot \mathcal{L}_{\mathrm{NTRD}}, \tag{12}$$

where β balances target robustness and dark knowledge preservation.

6.3 Overall Optimization Objective.

The overall training framework is formulated as a min-max optimization problem. For the inner maximization, we generate adversarial examples x_{adv} by maximizing the KL divergence between the student and teacher outputs following [14]. For the outer minimization, we integrate the proposed two gating mechanisms:

$$\min_{\theta_S} \mathbb{E}_{(x,y)\sim\mathcal{D}} \left[\mathcal{L}_{SLG}(x, x_{adv}) + \alpha \cdot \mathcal{L}_{SCG}(x, x_{adv})\right], \tag{13}$$

where α balances semantic correctness and geometric sensitivity calibration.

7 Experiments

7.1 Experimental Setup

Teacher and Student Models. Following [18], we select three teacher models including LTD [4], BDM-AT [19] for CIFAR-10 and CIFAR-100 datasets and PreActResNet-18 [12] for Tiny-ImageNet. For student models, we employ ResNet-18 [11] and MobileNetV2 [33] for the CIFAR-10/100 datasets.

Evaluation Metrics. We evaluate model performance by measuring accuracy on clean samples and robust accuracy against a suite of attacks: FGSM [9], PGD [23], CW_2 [1], and AutoAttack (AA) [6]. To ensure statistical stability and alleviate potential gradient masking concerns, all reported results are averaged across 4 independent random seeds. The perturbation budget (ϵ) for FGSM, PGD, and AA is set to 8/255. PGD is run for 10 steps and 20 steps with a step size (α) of 2/255. The CW_2 equilibrium constant is set to 0.1. Following the established evaluation protocols widely adopted in prior adversarial distillation works [8, 18, 29], all reported results are from the checkpoint that achieved the best PGD-10 accuracy.

Implementation Details. We employ an SGD optimizer with an initial learning rate of 0.1, a momentum of 0.9, and a weight decay of 5e-4 with standard data augmentation. All ARD methods (including ours) [8, 14, 29, 50, 51] are trained for 200 epochs with learning rate adjustments at epochs 100 and 150. For the inner optimization loop in adversarial training, we use 10 PGD iterations with a step size of 2/255 and an L_∞ perturbation bound of $\epsilon = 8/255$.

Table 1: Evaluation of clean and robust accuracy on CIFAR10.

Teacher: WideResNet-34-10 [4]								
Student	**Method**	Clean	FGSM	PGD_{10}	PGD_{20}	CW_2	AA	Avg.
ResNet18	Natural	95.20	37.12	0.00	0.00	0.00	0.00	7.42
	PGD-AT [23]	81.96	57.09	53.03	52.26	76.66	48.04	57.42
	TRADES [43]	79.59	56.82	53.99	53.52	75.52	48.65	57.70
	ARD [8]	85.10	61.40	55.29	53.91	79.74	50.10	60.09
	IAD [50]	83.26	61.83	57.73	56.76	79.38	52.20	61.58
	RSLAD [51]	82.92	59.42	55.34	54.54	79.29	50.25	59.77
	AKD [24]	83.76	59.63	54.89	54.14	78.83	50.26	59.55
	AdaAD [14]	84.16	59.87	56.26	55.43	80.23	51.41	60.64
	DGAD [29]	**85.46**	61.78	57.65	56.80	**81.54**	52.59	62.07
	IGDM [18]	82.21	60.51	57.15	56.57	78.91	51.80	60.99
	AEGIS	84.69	**62.66**	**58.68**	**57.89**	81.16	**53.22**	**62.72**
MobileNetV2	Natural	91.47	10.30	0.00	0.00	0.00	0.00	2.06
	PGD-AT [23]	80.63	55.52	50.65	49.82	74.80	45.61	55.28
	TRADES [43]	78.85	55.46	52.30	51.87	74.63	46.49	56.15
	ARD [8]	82.91	58.66	53.94	53.00	78.20	48.41	58.44
	IAD [50]	82.60	59.31	54.72	53.96	78.73	49.87	59.32
	RSLAD [51]	79.33	55.04	51.76	51.22	75.26	46.27	55.91
	AKD [24]	83.48	59.24	53.74	52.66	78.21	48.52	58.47
	AdaAD [14]	82.52	56.69	52.99	52.42	78.33	48.12	57.71
	DGAD [29]	84.07	59.85	55.95	55.11	79.99	50.90	60.36
	IGDM [18]	80.48	58.72	55.40	54.92	77.00	49.83	59.17
	AEGIS	**84.43**	**60.98**	**56.74**	**56.08**	**81.20**	**51.34**	**61.26**

7.2 Main Results

As shown in Tables 1 and 2, our method consistently improves robustness. On CIFAR-10, ResNet-18 achieves 53.22% AA and 84.69% clean accuracy, surpassing ARD [8] by **3.12%** and outperforming recent baselines like DGAD [29] and IGDM [18], which fall short under the strongest ensemble AA attack. This suggests that prior methods inherit sensitivity noise from the teacher's overfitted decision boundaries. For MobileNetV2, because of limited capacity and depthwise separable convolutions, lightweight models are susceptible to gradient noise during distillation. Traditional alignment often forces them to represent complex teacher geometries they cannot physically sustain. AEGIS's Semantic Logit Gate (SLG) effectively manages this capacity gap by switching back to the clean teacher when adversarial outputs degrade. This allows MobileNetV2 to achieve a remarkable 51.34% under AA. AEGIS also provides consistent performance on CIFAR-100. Under this complex label space, the SLG fallback mechanism prevents the student from learning the teacher's frequent misclassifications among hard labels. Furthermore, to verify that the performance gains of AEGIS are resilient to the specific form of attack used during training, we also conduct an ablation study on the inner maximization strategy (shown in the Appendix). The experiments demonstrate that AEGIS maintains high performance even when relying on standard PGD-10 to generate adversarial examples in the inner loop.

Table 2: Evaluation of clean and robust accuracy on CIFAR100.

Teacher: WideResNet-28-10 [19]								
Student	**Method**	Clean	FGSM	PGD_{10}	PGD_{20}	CW_2	AA	Avg.
ResNet18	Natural	77.23	7.99	0.01	0.00	0.00	0.00	1.60
	PGD-AT [23]	57.24	32.20	29.54	28.90	48.74	24.60	32.80
	TRADES [43]	54.39	31.95	29.67	29.33	48.37	24.57	32.78
	ARD [8]	58.46	34.26	31.93	31.55	51.42	26.18	35.07
	IAD [50]	61.13	36.13	33.12	32.57	53.17	26.79	36.36
	RSLAD [51]	59.19	34.76	31.79	31.34	51.43	25.68	34.20
	AKD [24]	58.62	34.45	32.05	31.62	51.28	26.28	35.14
	AdaAD [14]	63.88	37.56	34.47	33.96	56.09	28.16	38.05
	DGAD [29]	**64.35**	36.91	33.91	33.21	56.49	27.42	37.59
	IGDM [18]	60.24	37.22	34.42	33.91	53.71	**28.37**	37.53
	AEGIS	64.00	**37.80**	**34.76**	**34.13**	**56.50**	28.22	**38.28**
Teacher: WideResNet-34-10 [4]								
MobileNetV2	Natural	70.85	6.66	0.00	0.00	0.00	0.00	1.33
	PGD-AT [23]	55.46	31.30	28.39	27.82	48.39	23.28	31.84
	TRADES [43]	54.37	31.15	29.39	29.12	48.50	23.64	32.36
	ARD [8]	57.03	33.06	30.31	29.86	50.07	23.87	33.43
	IAD [50]	55.22	32.44	30.26	29.81	49.14	24.18	33.17
	RSLAD [51]	54.33	32.44	30.53	30.25	48.74	24.21	33.23
	AKD [24]	56.85	32.87	30.21	29.72	49.86	24.05	33.34
	AdaAD [14]	61.18	33.88	31.47	31.01	53.26	25.15	34.95
	DGAD [29]	**61.88**	34.63	31.96	31.62	**54.08**	25.90	35.64
	IGDM [18]	55.42	33.85	32.00	31.77	50.37	25.82	34.76
	AEGIS	61.23	**34.92**	**32.27**	**31.80**	53.83	**26.20**	**35.80**

Interestingly, we observe that DGAD [29] occasionally achieves higher clean accuracy (e.g., 85.46% vs. 84.69% on ResNet-18) and competitive results on CW attacks. It adopts Semantic Correction by employing Error-corrective Label Swapping, which fixes the teacher's misclassifications and ensures the student does not learn incorrect labels. This explains DGAD's high clean accuracy and competitive performance under the relatively weaker CW attack. Unfortunately, other than CW, it generally remains below AEGIS (by observing the last column of Average performance). This is because semantic correctness does not imply geometric stability. Even if DGAD swaps a label to make it correct, the underlying teacher model may still exhibit robust overfitting [32, 41]. By bringing another dimension of stability into consideration, AEGIS surpasses the previous methods under much stronger attacks.

Tiny-ImageNet. We also evaluate on Tiny-ImageNet by using a PreActResNet18 teacher and a ResNet18 student model. Table 3 demonstrates that our method significantly enhances model robustness, validating effectiveness even on larger-scale datasets, and outperforms other baselines.

7.3 Ablation Study

Contribution from Individual Components. As shown in Table 4, each proposed component contributes distinctly to the model's performance. First,

Table 3: Evaluating on Tiny-ImageNet. The teacher model is trained with TRADES (λ=6) for 110 epochs.

Model	Method	Clean	FGSM	PGD_{10}	CW_2	AA	Avg.
ResNet18	ARD* [8]	41.66	24.47	23.30	37.76	17.23	25.69
	RSLAD* [51]	40.83	23.45	22.58	37.12	17.05	25.05
	AdaAD* [14]	47.54	24.22	22.82	42.79	17.41	26.81
	DGAD* [29]	47.92	24.42	23.05	42.81	17.18	26.87
	AEGIS	**47.71**	**24.70**	**23.45**	**43.30**	**17.78**	**27.31**

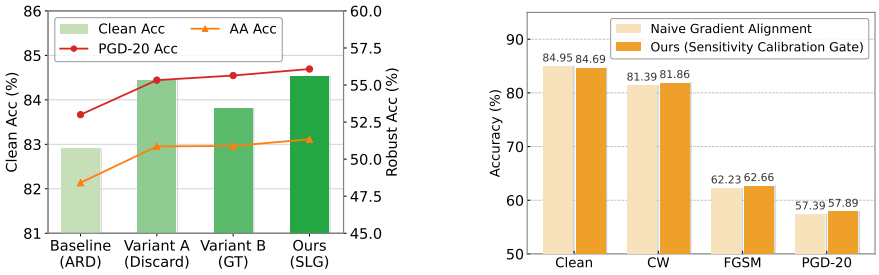
Table 4: Ablation study of component contribution on CIFAR10. We utilize ResNet18 (student) and WideResNet-34-10 (teacher) and additively introduce components including Target-class Gradient Robust Distillation (TRD), Non-Target-class Gradient Robust Distillation (NTRD) and Semantic Logit Gate (SLG).

Method	Clean	FGSM	PGD_{10}	PGD_{20}	CW_2	AA
Baseline [14]	84.16	59.87	56.26	55.43	80.23	51.41
+SLG	84.66	61.54	57.74	57.08	80.76	53.11
+TRD	84.82	62.39	58.63	57.66	81.14	53.09
+NTRD	84.76	61.54	57.84	57.13	80.88	52.96
+SLG+TRD	84.57	62.40	58.38	57.55	81.13	53.35
+SLG+NTRD	84.97	61.60	57.80	57.12	80.96	53.13
+SLG+TRD+NTRD	84.69	62.66	58.68	57.89	81.16	53.22

integrating SLG significantly enhances robustness stability that boosts accuracy under AutoAttack by 1.80% over the baseline. Second, Target-class Gradient Robust Distillation (TRD) calibrates the model's local geometry, yielding a 2.23% gain in PGD_{20} robustness by suppressing teacher-induced sensitivity noise. Finally, the complete integration of AEGIS achieves the best overall performance with the highest robustness on FGSM and PGD_{20} while maintaining competitive clean accuracy.

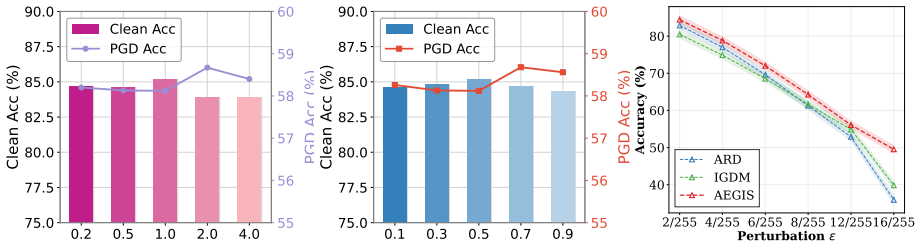
Impact of Semantic Logit Gate. We analyze the efficacy of the SLG by comparing it against two common strategies for handling teacher failure. Inspired by [50], we investigate two design variants: (Variant A) discarding erroneous samples and (Variant B) correcting them with ground truth labels. As illustrated in Fig. 5a, the baseline suffers significantly from indiscriminate error propagation. Although Variants A and B show some improvements by mitigating incorrect supervision, they fail to maximize the potential of knowledge transfer. In contrast, SLG achieves a superior trade-off, outperforming Variant B on AutoAttack. This demonstrates that even when a teacher's adversarial prediction fails, its clean-image logits encapsulate structural dark knowledge such as inter-class correlations that hard labels lack. By dynamically focusing on these clean logits, SLG provides a more informative and stable supervision.

Adaptive Selectivity in TRD. To verify the importance of the selectivity mechanism, we compare our Adaptive TRD against a Naive Gradient Alignment baseline [50] that indiscriminately enforces gradient matching across all samples regardless of the teacher's local stability. As shown in Fig. 5b, this greedy alignment leads to suboptimal robustness, particularly under FGSM and



(a) Effectiveness of SLG. **(b) Effectiveness of SCG.**

Fig. 5: Ablation studies on the design of SLG and SCG. (a) Analysis of SLG against ARD Baseline [8] and two variants [50]. **(b)** Validation of selectivity in Sensitivity Calibration Gate against Naive Gradient Alignment [18].



(a) Impact of α in SLG **(b)** Impact of β in SCG **(c)** Perturbation comparison

Fig. 6: Hyper-parameter sensitivity analysis and varying perturbation comparison. (a) α balances the SLG loss and the SCG loss. the model's robustness peaks at $\alpha = 2.0$. (b) Varying β for SCG indicates optimal trade-off at $\beta = 0.7$. Clean accuracy (bar) and PGD accuracy (line) are reported. (c) Robustness comparison under increasing attack strength ϵ.

PGD_{20} attacks. Additionally, this performance degradation re-confirms the existence of negative knowledge transfer, that hampers the student's optimization. Conversely, by filtering unreliable supervision through SCG, AEGIS ensures the student only aligns with the teacher when it is geometrically beneficial.

7.4 Parameter Analysis

Analysis of Scaling Parameters. We investigate the impact of the scaling parameters α and β in the optimization objectives. As shown in Fig. 6, optimal performance is achieved at $\alpha = 2.0$ and $\beta = 0.7$. Note that AEGIS favors a larger α compared to standard alignment methods, which weighs more on $\mathcal{L}_{SCG}$ that calibrates sensitivity. This confirms that our design of SCG plays an important role to improve the overall robustness. For β, the peak at 0.7 indicates that while target gradients are dominant, retaining non-target dark knowledge remains essential for maintaining clean accuracy.

Robustness under Varying Attack Intensities. To further evaluate the limits of the distilled robustness, we conduct a stress test by increasing the perturbation strength ϵ of PGD from 2/255 to 16/255. As shown in Fig. 6c, our method consistently outperforms baselines, and the performance of ARD and IGDM dips much faster under stronger attack strength 16/255, i.e., AEGIS preserves nearly 50% accuracy while competitors drop below 40%. Overall, connecting to the previous Tiny-ImageNet results in Table 3, it demonstrates that our method has better scalability under both stronger attack strength and more complex data distributions.

8 Conclusion

In this paper, we propose adversarial error gating to resolve stability punishment in ARD. By integrating logit rectification and gradient-level stability, we prevent students from inheriting teacher-induced sensitivity noise while preserving structural dark knowledge. Our extensive evaluations stand as a new SOTA with remarkable resilience against both extreme perturbation intensities and high-dimensional data distribution.

Acknowledgements

This work is supported by National Science Foundation of China under grants 62576310, Zhejiang Provincial National Science Foundation of China under Grant No. LZ25F020007 and ICT2026A20.

References

1. 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)
2. Chen, D., Mei, J.P., Wang, C., Feng, Y., Chen, C.: Online knowledge distillation with diverse peers. In: AAAI. vol. 34, pp. 3430–3437 (2020)
3. Chen, D., Mei, J.P., Zhang, H., Wang, C., Feng, Y., Chen, C.: Knowledge distillation with the reused teacher classifier. In: CVPR. pp. 11933–11942 (2022)
4. Chen, E.C., Lee, C.R.: Ltd: Low temperature distillation for robust adversarial training. arXiv preprint arXiv:2111.02331 (2021)
5. Chen, P., Liu, S., Zhao, H., Jia, J.: Distilling knowledge via knowledge review. In: CVPR. pp. 5008–5017 (2021)
6. Croce, F., Hein, M.: Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks. In: ICML. pp. 2206–2216. PMLR (2020)
7. Du, S., You, S., Li, X., Wu, J., Wang, F., Qian, C., Zhang, C.: Agree to disagree: Adaptive ensemble knowledge distillation in gradient space. NeurIPS **33**, 12345–12355 (2020)
8. Goldblum, M., Fowl, L., Feizi, S., Goldstein, T.: Adversarially Robust Distillation. In: AAAI (2020)
9. Goodfellow, I.J., Shlens, J., Szegedy, C.: Explaining and Harnessing Adversarial Examples. In: ICLR (2015)

10. Guo, Q., Wang, X., Wu, Y., Yu, Z., Liang, D., Hu, X., Luo, P.: Online knowledge distillation via collaborative learning. In: CVPR. pp. 11020–11029 (2020)
11. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR. pp. 770–778 (2016)
12. He, K., Zhang, X., Ren, S., Sun, J.: Identity mappings in deep residual networks. In: ECCV. pp. 630–645 (2016)
13. Hinton, G., Vinyals, O., Dean, J.: Distilling the knowledge in a neural network. arXiv preprint arXiv:1503.02531 (2015)
14. Huang, B., Chen, M., Wang, Y., Lu, J., Cheng, M., Wang, W.: Boosting Accuracy and Robustness of Student Models via Adaptive Adversarial Distillation. In: CVPR (2023)
15. Jung, J., Jang, H., Song, J., Lee, J.: Peeraid: Improving adversarial distillation from a specialized peer tutor. In: CVPR. pp. 24482–24491 (2024)
16. Kim, J., Park, S., Kwak, N.: Paraphrasing complex network: Network compression via factor transfer. NeurIPS **31** (2018)
17. Lan, W., Cheung, Y.m., Xu, Q., Liu, B., Hu, Z., Li, M., Chen, Z.: Improve knowledge distillation via label revision and data selection. IEEE Transactions on Cognitive and Developmental Systems (2025)
18. Lee, H., Cho, S., Kim, C.: Indirect gradient matching for adversarial robust distillation. In: ICLR. pp. 70007–70028 (2025)
19. Li, L., Wang, Y., Sitawarin, C., Spratling, M.: Oodrobustbench: a benchmark and large-scale analysis of adversarial robustness under distribution shift. In: ICML
20. Li, X.C., Fan, W.S., Song, S., Li, Y., Yunfeng, S., Zhan, D.C., et al.: Asymmetric temperature scaling makes larger networks teach well again. NeurIPS **35**, 3830–3842 (2022)
21. Li, Z., Li, X., Yang, L., Zhao, B., Song, R., Luo, L., Li, J., Yang, J.: Curriculum temperature for knowledge distillation. In: AAAI
22. Liu, Y., Cao, J., Li, B., Yuan, C., Hu, W., Li, Y., Duan, Y.: Knowledge distillation via instance relationship graph. In: CVPR. pp. 7096–7104 (2019)
23. Madry, A., Makelov, A., Schmidt, L., Tsipras, D., Vladu, A.: Towards Deep Learning Models Resistant to Adversarial Attacks. In: ICLR (2018)
24. Maroto, J., Ortiz-Jiménez, G., Frossard, P.: On the benefits of knowledge distillation for adversarial robustness. arXiv preprint arXiv:2203.07159 (2022)
25. Nagarajan, V., Menon, A.K., Bhojanapalli, S., Mobahi, H., Kumar, S.: On student-teacher deviations in distillation: does it pay to disobey? NeurIPS **36**, 5961–6000 (2023)
26. Nam, G., Yoon, J., Lee, Y., Lee, J.: Diversity matters when learning from ensembles. NeurIPS **34**, 8367–8377 (2021)
27. Ojha, U., Li, Y., Sundara Rajan, A., Liang, Y., Lee, Y.J.: What knowledge gets distilled in knowledge distillation? NeurIPS **36**, 11037–11048 (2023)
28. Pang, T., Yang, X., Dong, Y., Xu, K., Zhu, J., Su, H.: Boosting adversarial training with hypersphere embedding. In: NeurIPS
29. Park, H., Min, D.: Dynamic guidance adversarial distillation with enhanced teacher knowledge. In: ECCV. pp. 204–219 (2024)
30. Park, W., Kim, D., Lu, Y., Cho, M.: Relational knowledge distillation. In: CVPR. pp. 3967–3976 (2019)
31. Peng, B., Jin, X., Liu, J., Li, D., Wu, Y., Liu, Y., Zhou, S., Zhang, Z.: Correlation congruence for knowledge distillation. In: ICCV. pp. 5007–5016 (2019)
32. Rice, L., Wong, E., Kolter, Z.: Overfitting in adversarially robust deep learning. In: ICML. pp. 8093–8104. PMLR (2020)

33. Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., Chen, L.C.: Mobilenetv2: Inverted residuals and linear bottlenecks. In: CVPR. pp. 4510–4520 (2018)
34. Stanton, S., Izmailov, P., Kirichenko, P., Alemi, A.A., Wilson, A.G.: Does knowledge distillation really work? NeurIPS **34**, 6906–6919 (2021)
35. Sun, S., Ren, W., Li, J., Wang, R., Cao, X.: Logit standardization in knowledge distillation. In: CVPR. pp. 15731–15740 (2024)
36. Sun, W., Chen, D., Lyu, S., Chen, G., Chen, C., Wang, C.: Knowledge distillation with refined logits. In: ICCV. pp. 1110–1119 (2025)
37. Wang, C., Chen, D., Mei, J.P., Zhang, Y., Feng, Y., Chen, C.: Semckd: Semantic calibration for cross-layer knowledge distillation. IEEE Transactions on Knowledge and Data Engineering (2022)
38. Wang, Z., Yan, F., Wang, T., Wang, C., Shu, Y., Cheng, P., Chen, J.: Fed-dfa: Federated distillation for heterogeneous model fusion through the adversarial lens. In: AAAI. vol. 39, pp. 21429–21437 (2025)
39. Wen, T., Lai, S., Qian, X.: Preparing lessons: Improve knowledge distillation with better supervision. Neurocomputing
40. Yin, S., Xiao, Z., Song, M., Long, J.: Adversarial distillation based on slack matching and attribution region alignment. In: CVPR. pp. 24605–24614 (2024)
41. Yu, C., Han, B., Shen, L., Yu, J., Gong, C., Gong, M., Liu, T.: Understanding robust overfitting of adversarial training and beyond. In: ICML. pp. 25595–25610. PMLR (2022)
42. Zagoruyko, S., Komodakis, N.: Paying more attention to attention: improving the performance of convolutional neural networks via attention transfer. In: ICLR (2017)
43. Zhang, H., Yu, Y., Jiao, J., Xing, E., Ghaoui, L.E., Jordan, M.: Theoretically Principled Trade-off between Robustness and Accuracy. In: ICML (2019)
44. Zhang, L., Bao, C., Ma, K.: Self-distillation: Towards efficient and compact neural networks. TPAMI **44**(8), 4388–4403 (2021)
45. Zhang, L., Song, J., Gao, A., Chen, J., Bao, C., Ma, K.: Be your own teacher: Improve the performance of convolutional neural networks via self distillation. In: ICCV. pp. 3713–3722 (2019)
46. Zhang, R., Shen, J., Liu, T., Liu, J., Bendersky, M., Najork, M., Zhang, C.: Do not blindly imitate the teacher: Using perturbed loss for knowledge distillation. arXiv preprint arXiv:2305.05010 (2023)
47. Zhang, Y., Xiang, T., Hospedales, T.M., Lu, H.: Deep mutual learning. In: CVPR. pp. 4320–4328 (2018)
48. Zhao, B., Cui, Q., Song, R., Qiu, Y., Liang, J.: Decoupled knowledge distillation. In: CVPR. pp. 11953–11962 (2022)
49. Zhao, S., Yu, J., Sun, Z., Zhang, B., Wei, X.: Enhanced accuracy and robustness via multi-teacher adversarial distillation. In: ECCV. pp. 585–602 (2022)
50. Zhu, J., Yao, J., Han, B., Zhang, J., Liu, T., Niu, G., Zhou, J., Xu, J., Yang, H.: Reliable Adversarial Distillation with Unreliable Teachers. In: ICLR (2022)
51. Zi, B., Zhao, S., Ma, X., Jiang, Y.G.: Revisiting Adversarial Robustness Distillation: Robust Soft Labels Make Student Better. In: CVPR (2021)