Eng. Sci., 2025, **37**, 1753

Engineered Science
DOI: https://dx.doi.org/10.30919/es1753

Adversarial Distillation via Attention Helps Enhance Accuracy and Robustness

Ruicheng Niu, Zhihong Liang,* Yuxiang Huang, Yuanyuan Ma and Linyi Zheng

Abstract

Lightweight neural networks are widely deployed in resource-constrained environments such as mobile devices and edge computing. However, they often struggle to achieve a reliable balance between accuracy and robustness, particularly under adversarial attacks. This limitation poses significant risks in safety-critical applications like autonomous driving and healthcare, where both high performance and reliability are essential. To address this challenge, we propose attention distillation enhancing robustness (ADER), a novel adversarial distillation framework that integrates self-attention mechanisms and a dual-teacher strategy. Unlike conventional single-teacher methods, ADER simultaneously distills knowledge from a clean teacher and an adversarially trained teacher. Furthermore, it incorporates cross-domain attention maps as auxiliary supervision to guide the student model's spatial focus during training. This design enables the student to capture both discriminative and robust features effectively. Extensive experiments on Canadian institute for advanced research (CIFAR)-10 and CIFAR-100 demonstrate that ADER consistently outperforms state-of-the-art adversarial training and distillation methods. The proposed method achieves substantial improvements in both clean accuracy and adversarial robustness, highlighting its potential for secure and efficient deployment of lightweight models.

Keywords: Adversarial robustness; Knowledge distillation; Attention mechanisms; Lightweight neural networks.
Received: 12 June 2025; Revised: 18 August 2025; Accepted: 21 August 2025
Article type: Research article.

1. Introduction

Deep Neural Networks (DNNs) have become indispensable tools for addressing complex tasks, including image recognition and biometric verification.[1,2] However, Szegedy *et al*.[3] revealed a critical vulnerability: DNNs are susceptible to adversarial manipulations—subtle perturbations imperceptible to humans—that can significantly degrade their performance. This vulnerability poses substantial risks in high stakes scenarios such as autonomous driving and medical diagnostics, where model reliability is paramount.

Since the pioneering work by Szegedy *et al*. unveiled the vulnerability of neural networks to adversarial samples,[3] numerous attack methodologies have been developed, including FGSM,[4] PGD,[5] and CW.[6] These attacks are generally categorized into white-box and black-box attacks.[7-10] In white-box attacks, the attacker has full access to the model parameters and utilizes gradient-based optimization to generate adversarial samples. In contrast, black-box attacks rely solely on the model's output, requiring numerous queries to approximate gradients indirectly. Despite extensive research, robust defense against adversarial attacks remains challenging, underscoring the urgent need for developing more effective and computationally efficient defensive strategies.

Current strategies to defend against adversarial attacks include gradient masking, auxiliary model deployment, adversarial training,[11-13] adversarial distillation,[14] fuzzy gradient techniques,[15] and robust network architectures.[16] Among these, adversarial training is widely regarded as the most reliable, as it improves model robustness by integrating adversarial examples during training. However, this approach typically demands considerable computational resources, especially for large neural networks,[17] and often compromises performance on clean inputs. Such trade-offs between robustness and accuracy become even more severe for lightweight models, which already face strict constraints on

College of Big Data and Intelligent Engineering, Southwest Forestry University, Kunming, 650224, China
**Email:* zhliang@swfu.edu.cn (Zhihong Liang)

computational resources and deployment feasibility. The principle of adversarial training can be formulated as a minimax optimization problem, as shown in Eq. (1):

$$\min_{\theta} \left[\!\left[E_{(x,y)\sim\mathbb{D}} \right]\!\right] \left[\max_{\delta\in\Omega} \mathcal{L}\left(f(x+\delta;\theta),y\right) \right] \quad (1)$$

where f represents the neural network with parameters θ, (x, y) are clean samples and labels from the data distribution D, and δ denotes the adversarial perturbation constrained by Ω to ensure imperceptibility. However, this optimization problem typically incurs substantial computational overhead, especially for large-scale models.

Adversarial training methods such as TRADES [11] and MART[18] represent important progress, with TRADES balancing robustness and clean accuracy, and MART further enhancing robustness by addressing misclassification risks. However, adversarial training in general still suffers from high computational cost and tends to reduce clean accuracy, especially for lightweight models.

To overcome these limitations, knowledge distillation has emerged as a promising alternative, particularly well-suited to resource-constrained scenarios. For example, RAD[14] transfers robustness from a large teacher to a lightweight student; IAD[19] recommends using identically structured teacher–student pairs for more effective transfer; and RSLAD[20] leverages robust soft labels to further improve student performance. While these methods achieve notable improvements, they predominantly rely on a single-teacher structure, which limits their ability to simultaneously preserve clean accuracy and adversarial robustness.

To address these specific challenges, we propose ADER (Attention Distillation Enhancing Robustness), a novel adversarial distillation method utilizing self-attention and multiple teachers to significantly enhance the robustness and accuracy of lightweight models. Unlike conventional single-teacher distillation methods, ADER leverages two high-capacity pre-trained models simultaneously: one model dedicated to robustness transfer and another focused on accuracy retention. In addition to traditional soft-label knowledge distillation, ADER explicitly incorporates attention maps as auxiliary guidance. Since attention mechanisms are widely recognized as powerful tools for improving neural network interpretability and feature localization, it is natural to integrate them into adversarial distillation. These mechanisms are generally categorized into gradient-based and response-based visual explanations. Gradient-based methods, such as Guided Back-propagation[21] and Grad-CAM,[22,23] use positive gradients to generate attention maps that highlight class-specific features across models. In contrast, response-based methods like CAM[24] and ABN[25] provide direct visualization of a network's attention during the forward pass, enabling immediate and intuitive interpretation of its focus areas. This design directs the student model's attention to critical input regions more effectively, a distinctive feature setting it apart from previous attention-based distillation approaches.

Extensive experimental evaluations demonstrate that ADER markedly improves adversarial accuracy for lightweight models, achieving increases of up to 11.99% on CIFAR-10 and 16.95% on CIFAR-100 under PGD-TRADES attacks when benchmarked against existing adversarial training and distillation methods. The main contributions of our method are summarized as follows:

(1) We introduce ADER, an attention map-based adversarial distillation approach that simultaneously boosts the robustness and clean accuracy of lightweight student models.

(2) We propose a unique auxiliary attention structure designed explicitly to help student models focus on essential input features, thus distinctly enhancing model performance.

(3) Comprehensive experiments validate that our proposed method outperforms existing techniques, clearly demonstrating its practical effectiveness in adversarial robustness scenarios.

Comprehensive experiments validate that our proposed method outperforms existing techniques, clearly demonstrating its practical effectiveness in adversarial robustness scenarios.

2. Method

This section introduces a novel attention-guided knowledge distillation framework, which aims to simultaneously improve both robustness and clean accuracy of lightweight neural networks (see Fig. 1). The proposed method integrates two key components: (1) an attention map transfer mechanism that enhances interpretability and robustness, and (2) a multi-teacher distillation strategy that allows the student model to learn from both clean and adversarial knowledge sources. The following subsections elaborate on each component.

2.1 Attention map transfer for robustness and interpretability

To address the challenge of poor attention quality in lightweight models, we propose to transfer attention maps from high-capacity teacher networks to guide the student's learning process. Attention maps act as interpretable indicators of the model's focus, highlighting the input regions most relevant for classification. However, due to limited representational capacity, lightweight models often produce noisy or diffuse attention distributions, which degrade both performance and interpretability.

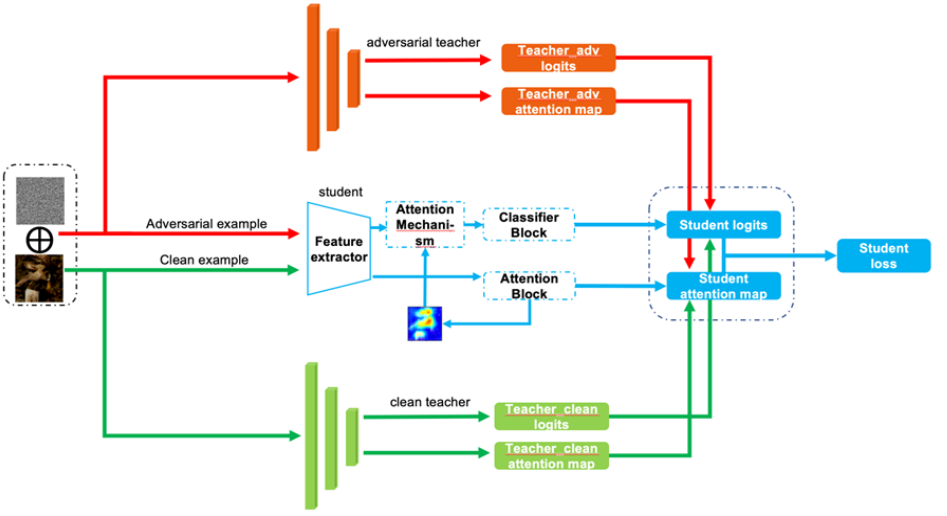

Fig. 1: Overview of the proposed multi-teacher adversarial distillation framework.

To mitigate this issue, we incorporate an attention transfer module into the distillation framework. The student model is explicitly encouraged to mimic the teacher's attention behavior, enabling it to concentrate more effectively on semantically salient regions. This design not only enhances interpretability but also improves the model's robustness against adversarial perturbations.

The design of the Attention Block in our study draws inspiration from.[20] It comprises a series of N × 3 × 3 convolutional layers, where N denotes the number of channels from the preceding layer. The input to the Attention Block undergoes transformation through K × 1 × 1 and 1 ×3×3 convolutional kernels, resulting in a 1 × h × w attention map. For effective attention map generation during training, the block is followed by K × 1 × 1 convolutions and Global Average Pooling (GAP), enabling the generation of class probabilities, where K represents the number of classes.

The attention map M(**x**i) modulates the feature map gc (**x**i) via a residual enhancement mechanism, as shown in Eq. (2):

$$g\,c^{\perp\prime}(xi) \;=\; (1 \;+\; M(xi))gc\,(xi) \qquad (2)$$

where g c′ (·) is the refined feature representation. This modulation amplifies

salient features while suppressing irrelevant ones, thereby improving both the discriminative power and interpretability of the student model.

From a theoretical perspective, adversarial examples typically disrupt model predictions by shifting attention toward misleading or irrelevant regions. Our attention distillation mechanism explicitly aligns the student's spatial focus with that of robust teacher models. This alignment is enforced during training by minimizing the divergence between the student and teacher attention distributions, as shown in Eq. (3):

$$Latt \;=\; KL(Ms\,(x)IIMt\,(x)) \qquad (3)$$

where Ms (x) and Mt (x) denote the attention maps of the student and teacher, respectively. By encouraging spatial consistency under both clean and adversarial inputs, the student learns to preserve stable attention patterns even when exposed to perturbations.

In addition, the residual attention mechanism functions as an implicit denoising operation. It selectively strengthens the impact of semantically meaningful features while reducing the influence of perturbed or noisy regions. Together, these mechanisms—spatial alignment and feature enhancement—form a robust foundation for improving both accuracy and adversarial resilience in lightweight neural networks.

2.2 Multi-Teacher distillation with clean and adversarial knowledge

To further boost the robustness of the student network, we propose a multi-teacher knowledge distillation strategy that leverages both clean and adversarial teacher models. While adversarial training has demonstrated strong robustness for large models, its benefits diminish when applied directly to compact architectures. To address this, recent work has explored knowledge distillation as a more efficient alternative.

The classic knowledge distillation objective can be defined as Eq. (4):

$$(arg\,min)_{\top}\theta S\;(1"\;-"\;\alpha)L(S(x),y)\;+\;\alpha\cdot KL(S(x),T(x)) \qquad (4)$$

where L is the standard cross-entropy loss, and KL denotes the KullbackLeibler divergence between the student S(·) and teacher T(·) predictions. The parameter α balances the contributions of hard labels and soft teacher outputs.

However, adversarially trained teachers often suffer from degraded cleansample accuracy. To overcome this trade-off, we employ a dual-teacher paradigm: a clean teacher ensures high standard accuracy, while an adversarial teacher imparts robustness. Furthermore, unlike traditional distillation methods that rely solely on soft logits, we also transfer attention maps from both teachers, enabling the student to learn feature-level and spatial cues associated with robustness and interpretability.

The distillation loss from the clean and adversarial teachers is shown in Eq. (5) and (6):

$$Lclean = KL(S(xclean), Tclean(xclean)) + KL(MS(xclean), MTclean(xclean)) \quad (5)$$

$$Ladv = KL(S(xadv), Tadv(xadv)) + KL(MS(xadv), MTadv(xadv)) \quad (6)$$

where MS and MT denote the attention maps of the student and teacher, respectively, and xadv is generated using the standard PGD attack, as shown in Eq. (7):

$$x\_"adv" = \arg\max\nolimits_{\delta \in \Omega} CE(S(x\_"clean" + \delta; \theta\_S), y) \quad (7)$$

with δ representing the adversarial perturbation constrained within a norm-bound set **Ω**.

To balance the contributions from both teachers, we define the total loss as shown in Eq. (8):

$$L = \omega \cdot Lclean + (1" - "\omega) \cdot Ladv \quad (8)$$

where ω is a dynamically adjusted coefficient, which can be annealed or adaptively updated during training to reflect the current learning stage. In our implementation, we adopt a convergence-aware adaptive weighting strategy inspired by,[26] which dynamically adjusts ω based on the relative convergence progress of each teacher branch.

Specifically, let Lclean (t) and Ladv (t) denote the KL divergence losses from the clean and adversarial teachers at training step t. We first normalize each loss with respect to its initial value, as shown in Eq. (9):

$$r\_i(t) = ((L\_i(t))/(L\_i(0)))\hat{}\beta, i \in \{"clean","adv"\} \quad (9)$$

where β > 0 is a tunable sensitivity factor. The final weight assigned to the clean teacher is then computed as shown in Eq. (10):

$$\omega(t) = (r\_"clean"(t))/(r\_"clean"(t) + r\_"adv"(t)) \quad (10)$$

This formulation ensures that the teacher whose guidance is less absorbed by the student receives more attention during training, leading to a more balanced and effective optimization process.

In summary, our method introduces a synergistic framework that combines attention-based spatial supervision and dual-teacher knowledge transfer. This design enables the lightweight student model to benefit from both clean accuracy and adversarial robustness while maintaining efficient inference. Extensive experiments demonstrate that this integrated strategy outperforms traditional single-teacher or output-only distillation methods, particularly in resource-constrained scenarios.

3. Experiments

This section presents our experimental evaluation. We begin by detailing the experimental setup, including datasets, model architectures, and attack configurations. We then compare the proposed method with four baseline models in terms of clean accuracy and robustness under both white-box and black-box attack settings. The black-box evaluation includes transfer-based and query-based attacks, covering a broad range of adversarial scenarios. Finally, we perform ablation studies to analyze the contribution of each component within our framework.

3.1 Experimental Setup

Datasets: We conduct experiments on two standard benchmark datasets:

CIFAR-10 and CIFAR-100, which are widely used in evaluating adversarial robustness in computer vision tasks. Their diversity and complexity make them suitable for assessing the effectiveness of defense methods under both clean and adversarial conditions.

Baseline Methods: To evaluate the performance of our method, we compare it against four state-of-the-art adversarial defense approaches:

- SAT (Standard Adversarial Training),[5] which directly incorporates adversarial examples during training.
- TRADES [11] which introduces a trade-off between clean accuracy and adversarial robustness via a surrogate loss.
- ARD (Adversarial Robustness Distillation),[14] which distills robustness from large teacher models into lightweight student networks.
- RSLAD (Revisiting Self-Labeled Adversarial Distillation),[20] an extension of ARD that further improves robustness using self-labeled adversarial examples.

Student and Teacher Networks: For student networks, we adopt two widely-used lightweight architectures: ResNet-18 and MobileNet-V2, selected for their favorable trade-off between computational efficiency and accuracy. For the teacher models, different networks are used depending on the dataset:

- On CIFAR-10, we employ ResNet-56 as the clean teacher and WideResNet34-10 as the adversarial teacher.
- On CIFAR-100, we adopt WideResNet-22-6 as the clean

Table 1: Pre-trained model's clean accuracy and adversarial accuracy.

	Model	Clean Acc (%)	FGSM (%)	PGDsat (%)	PGDtrades (%)	Type
CIFAR-10	RN-56	92.46	20.11	0	0	Clean Adv
	WRN-34-10	91.02	80.06	65.60	73.91	
CIFAR-100	WRN-22-6	76.78	27.00	0	0	Clean Adv
	WRN-70-16	67.63	50.12	31.24	33.18	

teacher and WideResNet-70-16 as the adversarial teacher.

All adversarial teachers are trained using the TRADES framework, which offers a balance between robustness and clean accuracy, making it a reliable source for robust supervision. Details of all pre-trained teacher models are summarized in Table 1.

Training Configuration: The Stochastic Gradient Descent (SGD) optimizer is selected for training the student network, with an initial learning rate of 0.1 and a momentum of 0.9. The training process consists of 300 iterations with a batch size of 128. The learning rate is reduced by a factor of 10 at iterations 215, 260, and 285, respectively. Additionally, Projected Gradient Descent (PGD) with 10 iterations (PGD-10) is applied during training, with a perturbation size of 0.001 and a step size of 2/255.

White-box Attack: To verify the security of the student model, this paper employs FGSM (Fast Gradient Sign Method), PGDsat, and PGDtrade. All methods use a step size of 2/255. The iteration number for PGDsat and PGDtrade is set to 20.

Black-box Attack: To evaluate the generalization of our method under black-box scenarios, we conduct robustness experiments on ResNet-18 across CIFAR-10 and CIFAR-100. Two types of black-box attacks are considered: transfer-based and query-based. For the former, adversarial examples are generated using surrogate models (WideResNet-34-10 and WideResNet-7016) via PGD-20 and CW (Carlini & Wagner Attack) attacks. For the latter, we adopt the Square Attack (SA), a query-based method that refines perturbations through model outputs.

3.2 ResNet-18 experimental results on CIFAR10

Table 2 presents the clean and robust accuracy of ResNet-18 under four white-box attacks. The proposed method consistently outperforms all baselines across all threat models. Under the FGSM attack, it achieves 75.59% robustness, surpassing SAT and RSLAD by 20.00 and 15.18 percentage points, respectively. For PGDsat and PGDtrade, the proposed method attains 61.35% and 67.72% robustness, with improvements of 7.41 and 11.99 points over the best baseline (RSLAD). Against the more challenging CW∞ attack, our method achieves the highest robustness at 62.19%, outperforming RSLAD and SAT by 9.52 and 16.22 points, respectively, while maintaining the highest clean accuracy (85.93%) across all settings.

Table 2: Quantitative analysis of the adversarial robustness of ResNet-18 on the CIFAR10 datasets.

		CIFAR-10	
Attack	Defense	Clean (%)	Robust (%)
FGSM	Natural	94.57	18.60
	SAT	84.20	55.59
	TRADES	83.00	58.35
	ARD	84.11	58.40
	RSLAD	83.99	60.41
	OURS	85.93	75.59
PGDsat	Natural	94.57	0
	SAT	84.20	45.95
	TRADES	83.00	52.35
	ARD	84.11	50.93
	RSLAD	83.99	53.94
	OURS	85.93	61.35
PGDtrade	Natural	94.57	0
	SAT	84.20	48.12
	TRADES	83.00	53.83
	ARD	84.11	52.96
	RSLAD	83.99	55.73
	OURS	85.93	67.72
CW∞	Natural	94.57	0
	SAT	84.20	45.97
	TRADES	83.00	50.23
	ARD	84.11	50.15
	RSLAD	83.99	52.67
	OURS	85.93	62.19

These results confirm the effectiveness of our dual-teacher distillation framework in enhancing both clean and adversarial performance. By leveraging complementary supervision from clean and adversarial teachers at both logit and attention levels, the student model learns robust yet generalizable representations, leading to superior defense across diverse attack types.

3.3 ResNet-18 experimental results on CIFAR100

Table 3 reports the performance of ResNet-18 on CIFAR-100 under various white-box attacks. The proposed method consistently achieves the highest robustness across all attack types. Under the FGSM attack, our method reaches 46.28%

robustness, outperforming RSLAD and SAT by 11.55 and 20.40 percentage points, respectively. For PGDsat and PGDtrade, it achieves 34.80% and 49.00%, exceeding RSLAD by 3.61 and 16.95 points. Against the stronger CW∞ attack, our method also performs best, with a robustness of 41.37%, which is 16.36 and 13.39 points higher than SAT and RSLAD, respectively.

Table 3: Quantitative analysis of the adversarial robustness of ResNet-18 on the CIFAR100 datasets.

		CIFAR-100	
Attack	Defense	Clean (%)	Robust (%)
FGSM	Natural	75.18	7.96
	SAT	56.16	25.88
	TRADES	57.75	31.36
	ARD	60.11	33.61
	RSLAD	58.25	34.73
	OURS	57.83	46.28
PGDsat	Natural	75.18	0
	SAT	56.16	21.18
	TRADES	57.75	28.05
	ARD	60.11	29.40
	RSLAD	58.25	31.19
	OURS	57.83	34.80
PGDtrade	Natural	75.18	0
	SAT	56.16	22.02
	TRADES	57.75	28.88
	ARD	60.11	30.51
	RSLAD	58.25	32.05
	OURS	57.83	49.00
CW∞	Natural	74.86	0
	SAT	59.19	25.01
	TRADES	55.41	27.72
	ARD	60.45	26.55
	RSLAD	59.01	27.98
	OURS	57.83	41.37

Despite slightly lower clean accuracy compared to some baselines, the proposed method demonstrates a superior balance between robustness and standard performance. These results highlight the effectiveness of our dual teacher distillation framework in improving resilience to adversarial attacks on more complex datasets.

3.4 MobileNet-V2 experimental results

Table 4 and Table 5 report the performance of MobileNet-V2 under various white-box attacks on CIFAR-10 and CIFAR-100. The proposed method consistently achieves the highest robustness across all attack types on both datasets. On CIFAR-10, our method achieves 75.33% robustness under FGSM, outperforming RSLAD and SAT by 15.86 and 19.44 percentage points, respectively. For PGDsat and PGDtrade, it reaches 69.98% and 62.27%, exceeding RSLAD by 16.73 and 6.54 points. Under the CW∞ attack, the method still performs best with 60.54% robustness, improving upon SAT and RSLAD by 13.92 and 8.76 points.

On CIFAR-100, similar trends are observed. Our method achieves 44.30% robustness under FGSM, 47.14% under PGDsat, 49.45% under PGDtrade, and 42.39% under CW∞, outperforming RSLAD by 10.42, 16.95, 17.40, and 14.18 percentage points, respectively.

Despite the lower model capacity of MobileNet-V2, the proposed method maintains a strong balance between clean accuracy and adversarial robustness. These results further validate the scalability and effectiveness of our dual-teacher distillation strategy across lightweight architectures and complex datasets.

Table 4: Quantitative analysis of the adversarial robustness of MobileNet-V2 on the CIFAR10 datasets.

		CIFAR-10	
Attack	Defense	Clean (%)	Robust (%)
FGSM	Natural	93.35	12.22
	SAT	83.87	55.89
	TRADES	77.95	53.75
	ARD	83.43	57.03
	RSLAD	83.2	59.47
	OURS	84.71	75.33
PGDsat	Natural	93.35	0
	SAT	83.87	46.84
	TRADES	77.95	49.06
	ARD	83.43	49.5
	RSLAD	83.2	53.25
	OURS	84.71	69.98
PGDtrade	Natural	93.35	0
	SAT	83.87	48.12
	TRADES	77.95	53.83
	ARD	83.43	52.96
	RSLAD	83.2	55.73
	OURS	84.71	62.27
CW∞	Natural	93.35	0
	SAT	83.87	46.62
	TRADES	77.95	46.06
	ARD	83.43	48.96
	RSLAD	83.2	51.78
	OURS	84.71	60.54

3.5 Block box attack

The proposed method, ADER, consistently achieves the highest robustness across both black-box settings. Under

Table 5: Quantitative analysis of the adversarial robustness of MobileNetV2 on the CI-FAR100 datasets.

		CIFAR-100	
Attack	Defense	Clean (%)	Robust (%)
FGSM	Natural	74.86	5.94
	SAT	59.19	30.88
	TRADES	55.41	30.28
	ARD	59.01	32.77
	RSLAD	58.25	33.88
	OURS	64.72	44.30
PGDsat	Natural	74.86	0
	SAT	59.19	25.64
	TRADES	55.41	23.33
	ARD	59.01	28.69
	RSLAD	58.25	30.19
	OURS	64.72	47.14
PGDtrade	Natural	74.86	0
	SAT	59.19	22.02
	TRADES	55.41	28.88
	ARD	59.01	30.51
	RSLAD	58.25	32.05
	OURS	64.72	49.45
CW∞	Natural	74.86	0
	SAT	59.19	20.9
	TRADES	55.41	24.19
	ARD	59.01	27.56
	RSLAD	58.25	28.21
	OURS	64.72	42.39

Table 6: Black-box robustness of ResNet-18 on CIFAR-10 dataset.

		CIFAR-10	
Attack	Defense	Clean	Robust
PGD-20	SAT	84.2%	64.74%
	TRADES	83.00%	63.56%
	ARD	84.11%	63.59%
	RSLAD	83.99%	63.9%
	OURS	85.93%	67.27%
CW∞	SAT	84.2%	64.88%
	TRADES	83.00%	62.85%
	ARD	84.11%	62.78%
	RSLAD	83.99%	63.02%
	OURS	85.93%	65.11%
SA	SAT	84.2%	71.3%
	TRADES	83.00%	70.33%
	ARD	84.11%	73.3%
	RSLAD	83.99%	72.1%
	OURS	85.93%	76.13%

transfer-based attacks on CIFAR-10, ADER achieves 67.27% (PGD-20) and 65.11% (CW∞) adversarial accuracy, outperforming RSLAD by 3.37 and 2.09 percentage points, respectively (Table 6). On CIFAR-100, it achieves 42.26% under PGD-20 and 41.17% under CW∞, exceeding the best baselines by 2.33 and 1.50 points (Table 7). These results highlight the effectiveness of attention-based knowledge transfer in enabling the student model to learn transferable and robust features, even under unseen adversarial sources.

Table 7: Black-box robustness of ResNet-18 on CIFAR-100 dataset.

		CIFAR-100	
Attack	Defense	Clean	Robust
PGD-20	SAT	56.16%	38.1%
	TRADES	57.75%	38.2%
	ARD	60.11%	39.53%
	RSLAD	58.25%	39.93%
	OURS	57.83%	42.26%
CW∞	SAT	56.16%	39.42%
	TRADES	57.75%	38.63%
	ARD	60.11%	38.85%
	RSLAD	58.25%	39.67%
	OURS	57.83%	41.17%
SA	SAT	56.16%	41.27%
	TRADES	57.75%	41.96%
	ARD	60.11%	48.79%
	RSLAD	58.25%	45.34%
	OURS	57.83%	51.22%

In query-based scenarios, ADER also exhibits superior performance. Against the Square Attack, it achieves 76.13% on CIFAR-10 and 51.22% on CIFAR-100, surpassing ARD by 2.83 and 2.43 percentage points, respectively. This advantage stems from the dual-teacher distillation strategy, where the adversarial teacher promotes robustness, and the clean teacher ensures feature discrimination. Through joint attention and logit distillation, ADER guides the student to focus on both semantically meaningful and adversarially resilient regions.

Overall, these results validate the robustness and generalization benefits of attention-guided dual-teacher distillation. By aligning attention distributions alongside soft labels, ADER enables the student model to learn representations that are both discriminative and resistant to black-box attacks.

The experimental results show that ADER consistently outperforms baseline methods across different datasets and attacks. The improvements mainly come from the complementary guidance of the clean and adversarial teachers, which helps the student maintain both discriminative and

robust features. On CIFAR-10, the gains are larger due to the relatively simpler dataset, while on CIFAR-100 the improvements are smaller but still significant, reflecting the higher complexity of the task. ResNet-18 also benefits more than MobileNet-V2 because of its higher representational capacity. Overall, these findings confirm that the dual-teacher and attention-based design is the key factor behind the observed robustness gains.

3.6 Distillation experiments

To evaluate the contribution of each teacher, we conduct two ablation experiments on CIFAR-10 with ResNet-18: (1) distillation using only the clean teacher's attention maps, and (2) using only the adversarial teacher's attention maps. All other training settings remain unchanged.

Clean Teacher Only: As shown in Fig. 2, using only the clean teacher's attention maps improves clean accuracy but significantly reduces robustness. This indicates that clean attention helps the student focus on discriminative features beneficial for standard classification, but lacks the guidance needed to handle adversarial perturbations.

Adversarial Teacher Only: Fig. 3 shows that distillation from only the adversarial teacher results in degradation of both clean accuracy and robustness. The student overfits to defensive cues while neglecting essential semantic features, leading to poor generalization on clean data and insufficient robustness under attack.

Combined Attention: Fig. 4 compares all three settings. The combined use of both clean and adversarial attention maps yields the best performance, enhancing both clean accuracy and robustness. The clean teacher contributes high-quality discriminative features, while the adversarial teacher provides guidance on defense-critical patterns. Together, they enable the student to learn a more balanced and robust feature representation.

Quantitatively, combining both sources improves clean accuracy by approximately 10 percentage points and adversarial robustness by 15 percentage points compared to using either attention map alone. These results confirm that dual-attention distillation provides complementary benefits and is essential for achieving robust yet accurate models.

3.7 Summary of Experimental Findings

Extensive experiments across CIFAR-10 and CIFAR-100 validate the effectiveness of the proposed ADER framework in improving adversarial robustness for lightweight models. On both ResNet-18 and MobileNet-V2, ADER consistently outperforms state-of-the-art defenses, including SAT, TRADES, ARD, and RSLAD, under a wide range of white-box attacks such as FGSM, PGD, and CW∞ . Notably, ADER improves robustness by an average of 10.80 percentage points across all evaluated attacks and datasets, while maintaining competitive clean accuracy.

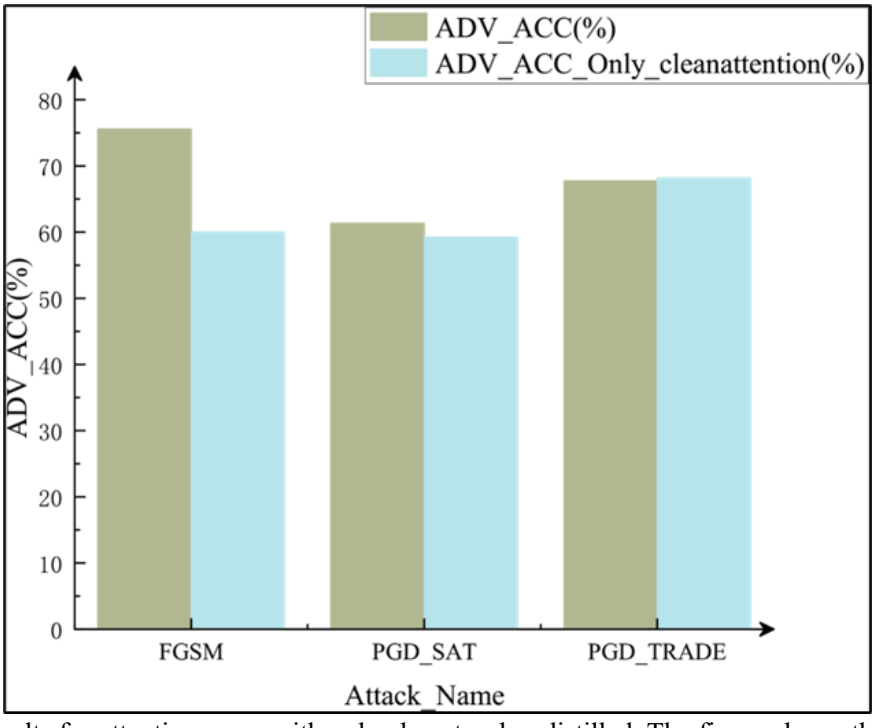

Fig. 2: Experimental results for attention maps with only clean teacher distilled. The figure shows that the model combining both types of attention maps performs better under various adversarial attacks compared to using only the clean teacher's attention map

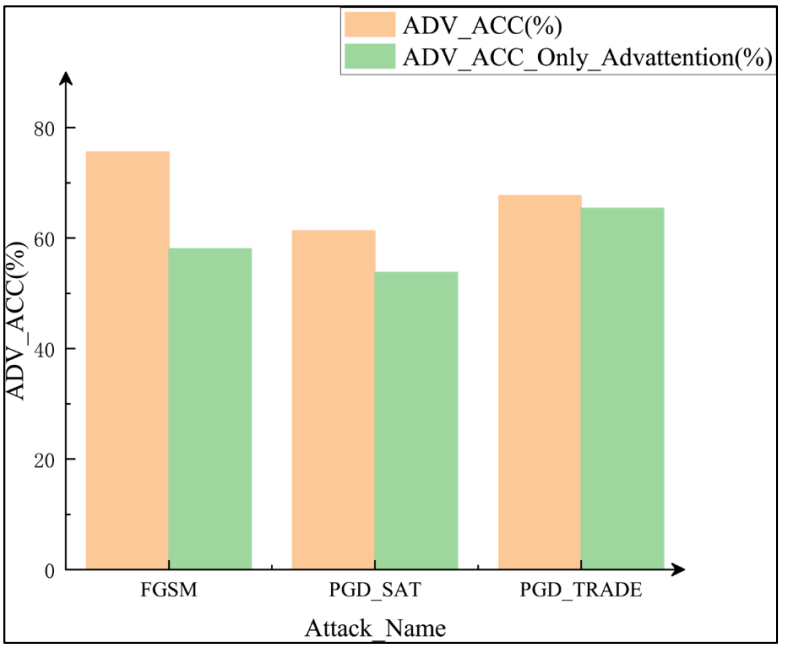

Fig. 3: Experimental results of the attention map for distillation-only adversarial teachers. The figure shows that the model combining both types of attention maps performs better under various adversarial attacks compared to using only the adversarial teacher's attention map.

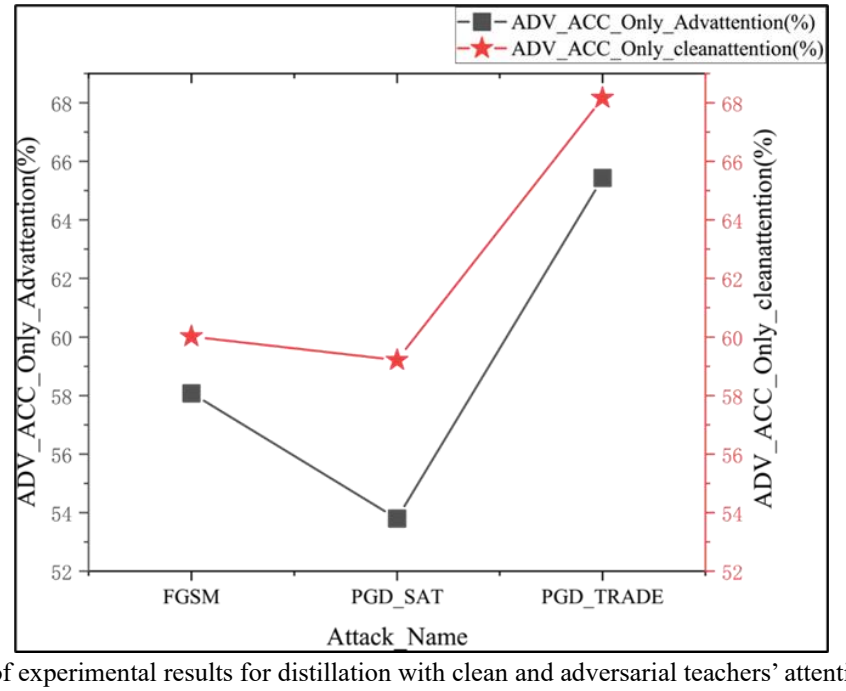

Fig. 4: Comparison of experimental results for distillation with clean and adversarial teachers' attention maps.

In black-box settings, including transfer-based and query-based attacks, ADER continues to demonstrate superior generalization. It achieves the highest adversarial accuracy across all evaluated attack methods, with consistent gains over the best-performing baselines. These results highlight the robustness and transferability of the features learned through attention-guided distillation.

Ablation studies further confirm that attention maps from both clean and adversarial teachers provide complementary supervision. Distilling from clean attention alone improves clean accuracy but weakens robustness, while adversarial attention alone harms generalization. Their combination enables the student model to capture both discriminative and defensive features, resulting in a balanced and robust performance.

Overall, the experiments demonstrate that ADER offers a principled and effective approach for training lightweight models with enhanced robustness and high accuracy. Its

generalizability across datasets, architectures, and attack scenarios underscores its potential for secure deployment in resource constrained environments.

4. Conclusion

In this paper, we address the challenge of enhancing both accuracy and robustness in lightweight models through adversarial distillation. We propose ADER, a novel multi-teacher distillation framework that incorporates self-attention mechanisms tailored for lightweight networks. By simultaneously distilling knowledge from clean and adversarial teachers—both at the logit and attention levels—our method enables the student model to learn features that are both discriminative and resilient. Experimental results demonstrate that integrating attention maps into the distillation process significantly improves the model's robustness without compromising clean accuracy. The proposed ADER framework is particularly suitable for deployment in resource-constrained, security-critical environments such as autonomous drones, mobile healthcare diagnostics, and smart surveillance systems. This work provides a balanced and effective solution for building secure and efficient lightweight models. Future research may further explore adaptive strategies to optimize the trade-off between robustness, accuracy, and computational cost.

Acknowledgments

This work was supported by the Key R&D Program Project of Yunnan Province (Grant No. 202503AA080013), the Yunnan Fundamental Research Project (Grant No. 202501AT070245), and the Doctoral Research Start-up Fund of Southwest Forestry University (Project: 25BS Research on the Robustness Mechanism and Enhancement Methods of Agroforestry Vision Models under Complex Environments, Grant No. 110225065).

Conflict of Interest

There is no conflict of interest.

Supporting Information

Not applicable.

CRediT Statement

Ruicheng Niu: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Data curation, Visualization, Writing – original draft, Writing – review and editing. **Zhihong Liang:** Resources, Funding acquisition. **Yuxiang Huang:** Funding acquisition. **Yuanyuan Ma:** Funding acquisition. **Linyi Zheng:** Funding acquisition.

References

[1] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 27-30, 2016, Las Vegas, NV, USA, 770-778, doi: 10.1109/CVPR.2016.90.

[2] H. Wang, Y. Wang, Z. Zhou, X. Ji, D. Gong, J. Zhou, Z. Li, W. Liu, CosFace: large margin cosine loss for deep face recognition, *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, June 18-23, 2018, Salt Lake City, UT, USA, 5265-5274, doi: 10.1109/CVPR.2018.00552.

[3] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, R. Fergus, *Intriguing properties of neural networks*, arXiv preprint arXiv:1312.6199 2013.

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

[5] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, A. Vladu, *Towards deep learning models resistant to adversarial attacks*, arXiv preprint arXiv:1706.06083 2017, doi:10.48550/arXiv.1706.06083.

[6] N. Carlini, D. Wagner, towards evaluating the robustness of neural networks, *IEEE Symposium on Security and Privacy (SP)*, May 22-26, 2017, San Jose, CA, USA, 39-57, doi: 10.1109/SP.2017.49.

[7] M. Andriushchenko, F. Croce, N. Flammarion, M. Hein, Square attack: a query-efficient black-box adversarial attack *via* random search, *Computer Vision – ECCV*, Cham: Springer International Publishing, 2020, 484-501, doi: 10.1007/978-3-030-58592-1_29.

[8] A. N. Bhagoji, W. He, B. Li, D. Song, Practical black-box attacks on deep neural networks using efficient query mechanisms, *Computer Vision – ECCV*, Cham: Springer, 2018, 158-174, doi: 10.1007/978-3-030-01258-8_10.

[9] X. Wei, H. Yan, B. Li, Sparse black-box video attack with reinforcement learning, *International Journal of Computer Vision*, 2022, **130**, 1459-1473, doi: 10.1007/s11263-022-01604-w.

[10] H. Yan, X. Wei, Efficient sparse attacks on videos using reinforcement learning, *Proceedings of the 29th ACM International Conference on Multimedia*, Virtual Event China, 2021, 2326-2334, doi: 10.1145/3474085.3475395.

[11] H. Zhang, Y. Yu, J. Jiao, E. Xing, L. El Ghaoui, M. Jordan, theoretically principled trade-off between robustness and accuracy, *International conference on machine learning*, PMLR, 2019, 7472-7482, doi:10.48550/arXiv.1901.08573.

[12] A. Ilyas, S. Santurkar, D. Tsipras, L. Engstrom, B. Tran, A. Madry, Adversarial examples are not bugs, they are features, *Advances in Neural Information Processing Systems*, 2019, **32**, doi:10.48550/arXiv.1905.02175.

[13] Y. Bai, Y. Zeng, Y. Jiang, S.-T. Xia, X. Ma, Y. Wang, *improving adversarial robustness via channel-wise activation suppressing*, arXiv preprint arXiv:2103.08307 2021, doi:10.1007/978-3-031-46677-9_10.

[14] M. Goldblum, L. Fowl, S. Feizi, T. Goldstein, Adversarially robust distillation, *Proceedings of the AAAI Conference on Artificial Intelligence*, 2020, **34**, 3996-4003, doi: 10.1609/aaai.v34i04.5816.

[15] J. A. Yu, L. Peng, Black-box attacks on DNN classifier based

on fuzzy adversarial examples, *IEEE 5th International Conference on Signal and Image Processing (ICSIP)*, October 23-25, 2020, Nanjing, China, 965-969, doi: 10.1109/ICSIP49896.2020.9339329.
[16] R. Niu, Z. Zhu, C. Li, D. Meng, Search robust and adaptable architecture, *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, April 14-19, 2024, Seoul, Republic of Korea, 5155-5159, doi: 10.1109/ICASSP48485.2024.10448197.
[17] B. Wu, J. Chen, D. Cai, X. He, Q. Gu, do wider neural networks really help adversarial robustness, *Advances in Neural Information Processing Systems*, 2021 **34** 7054-7067, doi:10.48550/arXiv.2010.01279.
[18] Y. Wang, D. Zou, J. Yi, J. Bailey, X. Ma, Q. Gu, improving adversarial robustness requires revisiting misclassified examples, in: *International conference on learning representations*, 2019, doi:10.48550/arXiv.2103.08307.
[19] J. Zhu, J. Yao, B. Han, J. Zhang, T. Liu, G. Niu, J. Zhou, J. Xu, H. Yang, *Reliable adversarial distillation with unreliable teachers*, arXiv preprint arXiv:2106.04928 2021, doi:10.48550/arXiv.2106.04928.
[20] B. Zi, S. Zhao, X. Ma, Y.-G. Jiang, revisiting adversarial robustness distillation: robust soft labels make student better, *IEEE/CVF International Conference on Computer Vision (ICCV)*, October 10-17, 2021, Montreal, QC, Canada, 16423-16432, doi: 10.1109/ICCV48922.2021.01613.
[21] J. T. Springenberg, A. Dosovitskiy, T. Brox, M. Riedmiller, Striving for simplicity, *The all convolutional net*, arXiv preprint arXiv:1412.6806 2014, doi:10.48550/arXiv.1412.6806.
[22] D. Smilkov, N. Thorat, B. Kim, F. Vi´egas, M. Wattenberg, Smoothgrad: *removing noise by adding noise*, arXiv preprint arXiv:1706.03825 2017, doi:10.48550/arXiv.1706.03825.
[23] A. Chattopadhay, A. Sarkar, P. Howlader, V. N. Balasubramanian, Grad-CAM: generalized gradient-based visual explanations for deep convolutional networks, *IEEE Winter Conference on Applications of Computer Vision (WACV)*, March 12-15, 2018, Lake Tahoe, NV, USA, 839-847, doi: 10.1109/WACV.2018.00097.
[24] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, A. Torralba, Learning deep features for discriminative localization, *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 27-30, 2016, Las Vegas, NV, USA, 2921-2929, doi: 10.1109/CVPR.2016.319.
[25] H. Fukui, T. Hirakawa, T. Yamashita, H. Fujiyoshi, Attention branch network: learning of attention mechanism for visual explanation, *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 15-20, 2019, Long Beach, CA, USA, 10697-10706, doi: 10.1109/CVPR.2019.01096.
[26] S. Zhao, J. Yu, Z. Sun, B. Zhang, X. Wei, Enhanced accuracy and robustness *via* multi-teacher adversarial distillation, *Computer Vision – ECCV*, Cham: Springer Nature Switzerland, 2022, 585-602, doi: 10.1007/978-3-031-19772-7_34.

Publisher's Note: Engineered Science Publisher remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access

This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International, which permits the use, sharing, adaptation, distribution and reproduction in any medium or format, as long as appropriate credit to the original author(s) and the source is given by providing a link to the Creative Commons license. This usage for commercial purposes is not allowed. If modifications, adaptations or any other transformation were made, it is not allowed for distribution. The images or other third-party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit https://creativecommons.org/licenses/by-nc-nd/4.0/.

© The Author(s) 2025.