Trustworthy Adversarial Robustness via Adaptive Learning-Driven Multi-Teacher Knowledge Distillation

Hayat Ullah[1,*], Talha Zaidi[2,*] and Arslan Munir[1,*]

[1]Department of Electrical Engineering and Computer Science, Florida Atlantic University, Boca Raton, Florida, 33431 USA
[2]Department of Computer Science, Kansas State University, Manhattan, Kansas, 66506 USA

Abstract

Convolutional neural networks (CNNs) excel in computer vision but are susceptible to adversarial attacks, crafted perturbations designed to mislead predictions. Despite advances in adversarial training, a gap persists between model accuracy and robustness. To mitigate this issue, in this paper, we present a multi-teacher adversarial robustness distillation using an adaptive learning strategy. Specifically, our proposed method first trained multiple clones of a baseline CNN model using an adversarial training strategy on a pool of perturbed data acquired through different adversarial attacks. Once trained, these adversarially trained models are used as teacher models to supervise the learning of a student model on clean data using multi-teacher knowledge distillation. To ensure an effective robustness distillation, we design an adaptive learning strategy that controls the knowledge contribution of each model by assigning weights as per their prediction precision. Distilling knowledge from adversarially pre-trained teacher models not only enhances the learning capabilities of the student model but also empowers it with the capacity to withstand different adversarial attacks, despite having no exposure to adversarial data. To verify our claims, we extensively evaluated our proposed method on MNIST-Digits, Fashion-MNIST, CIFAR-10 datasets across diverse experimental settings. The obtained results exhibit the efficacy of our multi-teacher adversarial distillation and adaptive learning strategy, enhancing CNNs' adversarial robustness against various adversarial attacks. https://github.com/iscaas/MTKD-AR

Keywords

Trustworthy AI, Adversarial robustness, Knowledge distillation, Multi-teacher learning, Adversarial training

1. Introduction

Despite their strong performance in real-world vision tasks [1, 2, 3], convolutional neural networks (CNNs) can be surprisingly fragile. A perturbation that is almost invisible to the human eye can be enough to change a correct prediction into a confident mistake [4, 5]. This weakness becomes even more concerning because adversarial examples do not always require direct access to the target model; attackers can exploit transferability from surrogate models and still mislead deployed systems. In safety-sensitive domains such as autonomous driving, surveillance, and medical imaging, such failures are not merely accuracy drops, but potential risks to trust, reliability, and decision safety [6]. Building CNNs that remain reliable under adversarial perturbations is therefore a central requirement for trustworthy AI.

A broad range of defenses has been developed to improve adversarial robustness, including input denoising [7, 8], adversarial-example detection [9, 10], certified robustness [11], and inference-time input transformations such as noise removal [12, 13], super-resolution [14], and JPEG compression [15]. Among these directions, adversarial training remains one of the most effective approaches because it directly exposes the model to perturbed examples during training [5, 16]. However, this robustness often comes with practical costs. Generating adversarial samples increases training complexity, robustness gains may reduce clean-image accuracy, and defenses based on input transformations can still be bypassed by stronger attacks [17]. Architecture-level modifications and ensemble-based defenses can

TRUST-AI: The Second European Workshop on Trustworthy AI. Organized as part of the International Joint Conference on Artificial Intelligence - IJCAI/ECAI 2026. August 2026, Bremen, Germany.
*Corresponding author.
hullah2024@fau.edu (H. Ullah); tzaidi@ksu.edu (T. Zaidi); arslanm@fau.edu (A. Munir)
0000-0001-9120-4882 (H. Ullah); 0000-0001-6716-9132 (T. Zaidi); 0000-0002-3126-8945 (A. Munir)

also improve robustness, but they may introduce extra computational overhead and limit deployment flexibility.

Knowledge distillation offers a promising alternative by transferring robustness from stronger teacher models to smaller student networks [18, 19, 20, 21]. This direction is particularly attractive for resource-constrained settings, where directly adversarially training every deployed model may be impractical. Yet existing adversarial robustness distillation methods still face an important limitation: a single teacher is often specialized to a particular attack, while multi-teacher approaches commonly assume that all teachers should contribute equally. In practice, this assumption is weak. A teacher that is reliable for one input or attack type may provide misleading guidance for another. To address this issue, we propose MTKD-AR, an adaptive multi-teacher knowledge distillation framework that transfers robustness from multiple adversarially trained teachers to a student model trained only on clean data. Instead of statically averaging teacher predictions, MTKD-AR dynamically weights each teacher according to its prediction reliability, allowing the student to benefit more from informative teachers while reducing the influence of unreliable ones. More precisely, the main contributions of this work include:

1. We propose a novel multi-teacher knowledge distillation framework, MTKD-AR, which enables the student model to achieve adversarial robustness without requiring adversarial data. The framework offers distillation from multiple adversarially trained teachers to a single student model, equipping the student to withstand diverse adversarial attacks.
2. We propose an adaptive teacher-weighting strategy for multi-teacher knowledge distillation. Since different teachers may be unreliable for different input samples, the proposed method dynamically assigns importance weights based on each teacher's prediction quality, allowing the student to rely more on the most informative teachers while reducing the influence of misleading predictions.
3. We conduct extensive quantitative experiments on MNIST-Digits, MNIST-Fashion, and CIFAR-10 datasets to demonstrate the effectiveness of our proposed method in terms of both knowledge distillation and adversarial robustness.

2. Related Works

Adversarial examples have been widely studied as a major vulnerability of deep neural networks. Early gradient-based attacks such as the Fast Gradient Sign Method (FGSM) [4, 5] showed that small perturbations aligned with the loss gradient can mislead a classifier while remaining visually subtle. Several stronger variants have since been proposed. The Refined Fast Gradient Sign Method (RFGSM) introduces a random perturbation before applying the gradient step [22, 23], while the Basic Iterative Method (BIM) [24] applies FGSM repeatedly with smaller updates. Projected Gradient Descent (PGD) [16] further strengthens this idea by using iterative optimization with projection onto a bounded perturbation region. Other attacks, including DeepFool [25], Carlini and Wagner (C&W) [26], and Simultaneous Perturbation Stochastic Approximation (SPSA) [27], have expanded the evaluation space through decision-boundary, optimization-based, and gradient-free attack formulations. Together, these methods show that robustness must be assessed across diverse attack types rather than a single perturbation strategy.

To defend against adversarial perturbations, several families of methods have been proposed. Adversarial training remains one of the most effective approaches, where adversarial samples are included during training to improve robustness [5, 16]. Later variants improved this direction by using larger models [28], additional unlabeled data [29], domain adaptation [30], robustness–accuracy trade-off objectives such as TRADES [31], misclassification-aware training [32], channel-wise activation suppression [19], and adversarial weight perturbation [33]. Although effective, these methods often require repeated adversarial-example generation, larger-capacity networks, or additional data, which makes them costly for resource-constrained settings such as mobile devices, autonomous systems, and embedded vision applications. Efficient adversarial training methods partially reduce this cost by updating model parameters and perturbations in a single backward pass [34], but the computational burden

remains a concern.

Knowledge distillation has therefore emerged as a promising alternative for transferring robustness from stronger models to smaller students [18, 20, 21]. Existing adversarial robustness distillation methods typically rely on a single robust teacher or combine multiple models without considering whether each teacher is reliable for a given input. This can be limiting because teachers trained under different attacks may specialize in different perturbation patterns and may not provide equally useful guidance across all samples. Beyond adversarial data augmentation, other defenses also explore robust feature representations through network ensembles or architecture-level modifications [35, 14, 22, 36]. However, these approaches can add architectural complexity or inference overhead. In contrast, our proposed framework uses multiple adversarially trained teachers only during distillation and introduces adaptive teacher weighting to control each teacher's contribution based on its prediction reliability, enabling robust knowledge transfer without adversarially training the student model.

3. Problem Formulation

Adversarial attacks involve generating small but carefully crafted perturbations to input data in order to manipulate a model's predictions while preserving the visual similarity of the original image. In CNN-based computer vision tasks, these perturbations are typically derived from the gradients of the model's loss function with respect to the input image. The objective is to produce imperceptible modifications that can mislead the model into incorrect predictions.

Considering a classification model $f(x, \theta) : \mathbb{R}^{h \times w \times c} \rightarrow \{1, \ldots, k\}$, where θ denotes the learnable parameters and h, w, and c represent the spatial and channel dimensions of the input image, the adversary aims to generate a perturbation $\delta \in \mathbb{R}^{h \times w \times c}$ with perturbation budget ϵ that maximizes the classification loss:

$$\delta = argmax_{|\delta|_p \leq \epsilon} \mathcal{L}_{ce}(\theta, x + \delta, y) \tag{1}$$

Here, y represents the ground-truth label, while $|\delta|_p$ denotes the L_p norm controlling the perturbation magnitude. The perturbed image x' is then obtained as:

$$x' = x + \delta \tag{2}$$

To mitigate the impact of adversarial attacks, defense mechanisms aim to learn robust model parameters that minimize vulnerability to perturbed inputs. Generally, adversarial defense can be formulated as the following optimization problem:

$$min_{\theta'} \mathcal{L}(\theta', x, y) + \lambda \cdot R(\theta') \tag{3}$$

where $\mathcal{L}(\theta', x, y)$ represents the classification loss on clean data, λ controls the trade-off between robustness and standard performance, and $R(\theta')$ denotes a regularization term designed to improve robustness against adversarial perturbations.

4. Proposed Method

To mitigate adversarial attacks and enhance CNN robustness, we propose a multi-teacher knowledge distillation (defensive distillation) framework with an adaptive weighting mechanism that assigns higher importance to more reliable teacher predictions. Unlike prior work that treats adversarial training and distillation separately, our approach integrates them in a unified setting to enable context-aware knowledge transfer. Notably, it achieves robustness using only clean data, avoiding the need for adversarial samples, thereby reducing computational cost and improving scalability.

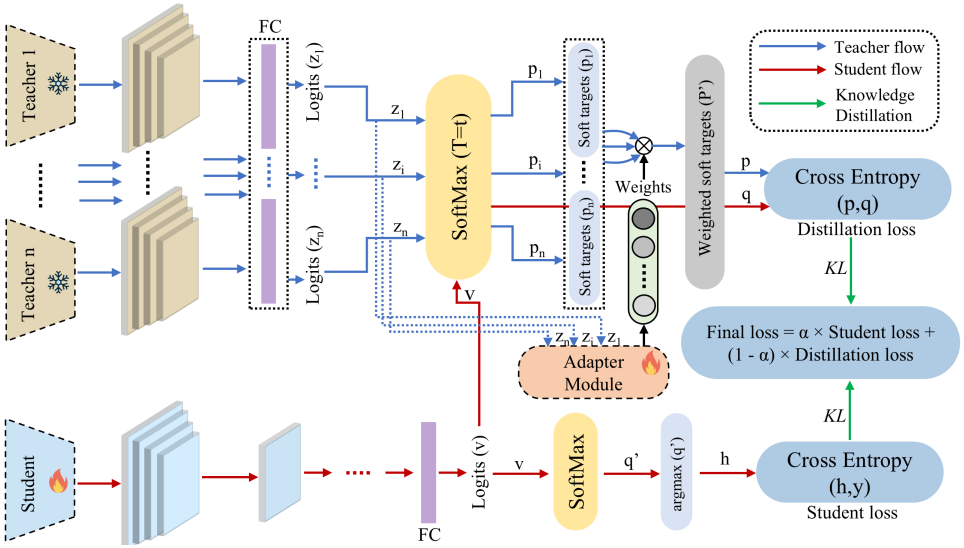

Figure 1: Detailed graphical overview of our proposed framework, depicting the overall workflow of adaptive learning-driven multi-teacher knowledge distillation for improving model robustness against adversarial attacks.

4.1. Adversarial Training Settings

The primary aim of this study is to enhance CNN robustness against adversarial attacks without relying on perturbed training data. To this end, we employ an adversarial training strategy that exposes the model to a range of attacks, including FGSM, FFGSM, RFGSM, and PGD. This formulation can be interpreted as the following min–max optimization problem in Eq. 4.

$$\min_{\theta} \frac{1}{|B|} \sum_{x,y \in B} \max_{\|\delta\| \leq \epsilon} \mathcal{L}(h_\theta(x+\delta), y) \tag{4}$$

Here, the inner maximization generates worst-case perturbations δ bounded by ϵ, which are added to input samples x to maximize the loss, while the outer minimization updates model parameters to reduce this loss over adversarially perturbed batches. A standard solution is to explicitly generate adversarial examples during training and incorporate them into optimization. To solve the outer minimization, Eq. 4 is reformulated using gradient-based optimization as shown in Eq. 5.

$$\theta := \theta - \alpha \frac{1}{|B|} \sum_{x,y \in B} \nabla_\theta \max_{\|\delta\| \leq \epsilon} \mathcal{L}(h_\theta(x+\delta), y) \tag{5}$$

Using the sub-differential property of the maximum, the inner maximization can be rewritten as Eq. 6, where the optimal perturbation is defined in Eq. 7.

$$\nabla_\theta \max_{\|\delta\| \leq \epsilon} \mathcal{L}(h_\theta(x+\delta), y) = \nabla_\theta \mathcal{L}(h_\theta(x+\delta^*(x)), y) \tag{6}$$

$$\delta^\star(x) = \text{argmax}_{\|\delta\| \leq \epsilon} \mathcal{L}(h_\theta(x+\delta), y) \tag{7}$$

Thus, model parameters are updated using gradients computed at the worst-case perturbation, as follows:

$$\theta := \theta - \alpha \frac{1}{|B|} \times \nabla_\theta \mathcal{L}(h_\theta(x+\delta^*(x)), y) \tag{8}$$

4.2. Adversarial Robustness Distillation

This section presents the MTKD-AR framework, which transfers knowledge from multiple adversarially trained teacher models to a single student model. By leveraging diverse robustness signals from these

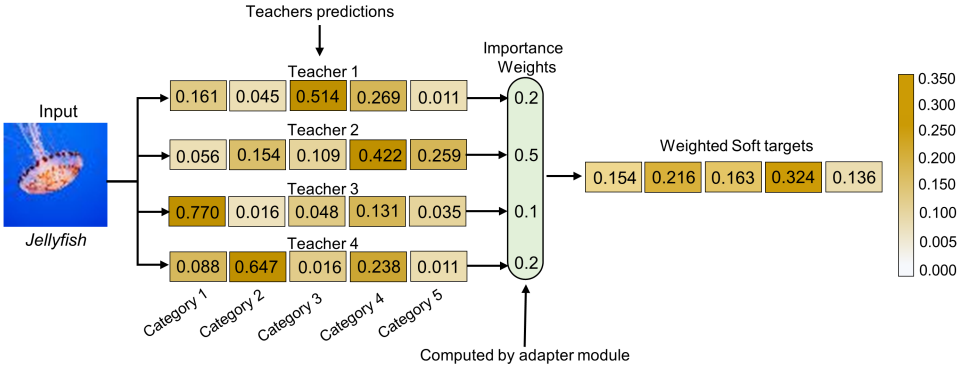

Figure 2: Overview of the proposed adaptive learning mechanism for multi-teacher knowledge distillation.

teachers, the proposed method enables the student to generalize against various adversarial attacks without requiring exposure to perturbed data. Figure 1 provides an overview of the framework and its key components.

4.2.1. Dynamic Weights Selection for Adaptive Learning

The *Adapter module* forms the basis for enabling the student model to learn from multiple adversarially trained teacher models. In conventional multi-teacher distillation, teacher outputs are typically averaged with equal contribution, regardless of individual performance. This naive averaging (average learning) can degrade performance when weaker or less reliable teachers introduce noisy or misleading predictions. To address this limitation, we propose an adaptive learning strategy that assigns dynamic importance weights to teachers based on their prediction quality, as illustrated in Figure 2. The weighting mechanism is guided by cosine similarity between teacher and student logits. Consider multiple adversarially trained teacher models $T_{M_{(1,n)}}$ and a student model S_M. During distillation, teachers and student produce logits $z_{(1,n)}$ and v, respectively, where logits represent pre-softmax outputs from the fully connected layer. These logits are transformed into soft targets $P_{(1,n)}$ and q using a temperature-controlled softmax function:

$$P_{(1,n)} = \sum_{i=1}^{N} SoftMax(\overbrace{T_{M_i}\theta(x)}^{z_{(1,n)}}) \tag{9}$$

$$q = SoftMax(\overbrace{S_M\theta(x)}^{v}) \tag{10}$$

Here, $T_{M_i}\theta$ and $S_M\theta$ denote the i^{th} teacher and student models, respectively, and x is the input image. To assign adaptive importance to each teacher, we compute cosine similarity between student logits v and teacher logits $z_{(1,n)}$, which is then used to derive teacher weights. These weights are subsequently normalized to ensure stability. The formulations are given below:

$$t_{similarity_{(1,n)}} = \sum_{i=1}^{N} CosineSimilarity(v, z_i) \tag{11}$$

$$t^{\theta}_{(1,n)} = \sum_{i=1}^{N} 1 + t_{similarity_i} \tag{12}$$

$$t^{\theta'}_{(1,n)} = \sum_{i=1}^{N} \frac{t^{\theta}_i}{t^{\theta}_1 + t^{\theta}_2 + ... + t^{\theta}_n} \tag{13}$$

Here, v and z_i represent student and i^{th} teacher logits, respectively. $t_{similarity_{1,n}}$ denotes the similarity vector between student and teacher outputs, $t^{\theta}_{1,n}$ represents the unnormalized teacher weights, and $t^{\theta'}_{1,n}$

denotes the normalized weight vector. Overall, the adaptive learning mechanism is the core contribution of our framework. Unlike conventional ensemble-based distillation that treats all teachers equally, our method dynamically adjusts each teacher's contribution based on its relevance to the current input. By leveraging cosine similarity between teacher and student logits, the framework prioritizes more reliable teachers and reduces the influence of noisy predictions, enabling more effective knowledge transfer.

4.2.2. Multi-Teacher Knowledge Distillation with Adaptive Learning

We propose an Adaptive Learning Driven multi-teacher knowledge distillation framework that supervises the student $\mathcal{S}$ model using multiple adversarially trained teacher models $\{\mathcal{T}_1, \mathcal{T}_2, ..., \mathcal{T}_n \mid n = 4\}$. The framework enables the student to learn adversarial robustness from teachers without access to adversarial data by introducing an adapter module that dynamically assigns weights $\mathcal{W}_i$ to each teacher $\mathcal{T}_i$ prediction. This adaptive weighting allows more informative teachers to contribute more strongly, improving robust feature learning and generalization. During training, the adapter produces weighted teacher supervision to guide the student. The distillation loss is defined as:

$$distillation\ loss = \mathcal{KLD}(\ \overbrace{\overbrace{p_{(1,n)} * W_{(1,n)}}^{p'}}^{\ }\ \overbrace{\sigma(z_{(1,n)}/\tau)}\ , \overbrace{\sigma(v/\tau)}^{q}) \cdot \tau^2, \tag{14}$$

where $z_{(1,n)}$ and v are the teacher and student logits, respectively. The function σ produces probability distributions $p_{(1,n)}$ and q, and τ is the temperature controlling output smoothness. The adapter weights $W_{(1,n)}$ generate the weighted teacher signal $p' = p_{(1,n)} * W_{(1,n)}$. The student is also optimized using cross-entropy:

$$student\ loss = CE(y, q), \quad where,\ q = \sigma(v) \tag{15}$$

Where y denotes ground-truth labels and v are student logits. The final objective combines both terms:

$$total\ loss = (1 - \alpha) * student\ loss + \alpha * distillation\ loss, \tag{16}$$

Where α balances clean-data accuracy and distillation-driven robustness. Overall, the proposed MTKD-AR framework leverages adaptive weighting of multiple adversarial teachers to improve robustness and generalization while avoiding adversarial training for the student.

5. Experimental Settings

5.1. Implementation Details

The proposed MTKD-AR framework was implemented in Python 3 using TensorFlow 2.0 and evaluated on a system equipped with an Intel Xeon E5-2640 CPU, 32 GB RAM, and two NVIDIA Tesla T4 GPUs with CUDA 12.0 support. A lightweight CNN architecture was adopted to ensure computational efficiency and suitability for resource-constrained environments. The implementation details and source code are publicly available to support reproducibility and verification of results.

5.2. Datasets

The experiments were conducted on the MNIST-Digits and MNIST-Fashion datasets, both of which are widely used benchmark datasets for image classification tasks. Each dataset consists of 60,000 training images and 10,000 test images with a spatial resolution of $28 \times 28 \times 1$. MNIST-Digits contains grayscale images of handwritten digits categorized into 10 classes (0–9), whereas MNIST-Fashion comprises grayscale images of fashion-related items such as shirts, shoes, and dresses across 10 categories. Following the standard evaluation protocol, we utilized 60,000 images for training (including adversarial and knowledge distillation training) and 10,000 images for model evaluation on both datasets.

5.3. Training Details

The proposed training scheme consists of two phases: adversarial training and multi-teacher knowledge distillation. During adversarial training, the training data was equally divided into clean and adversarial samples to enable the model to learn robust feature representations. The network was trained for 50 epochs with a batch size of 64 using the $cross-entropy$ loss and the $ADAM$ optimizer with a learning rate of $1e^{-3}$. In the multi-teacher knowledge distillation phase, the student model was trained solely on clean data by learning from the predictions of multiple adversarially trained teacher models. This strategy enables the student model to acquire both classification knowledge and adversarial robustness without direct exposure to adversarial samples. To evaluate robustness under different attack severities, the models were tested against FGSM, FFGSM, RFGSM, and PGD attacks using perturbation strengths of ($\epsilon \in$ 0.1, 0.2, 0.3). For both the single-teacher and multi-teacher knowledge distillation settings, the distillation hyperparameters were fixed to (α = 0.3) and a temperature of (T = 5). The remaining training hyperparameters were kept the same as those used during the adversarial training phase.

5.4. Threat Models

In this study, we evaluate the proposed framework against four white-box adversarial attacks: FGSM, FFGSM, RFGSM, and PGD, where the adversary has full access to model parameters and gradients. To analyze robustness under varying attack strengths, perturbation magnitudes of $\epsilon = 0.1, 0.2$, and 0.3 were employed. We further assessed attack transferability and robustness generalization by testing models under stronger perturbation settings than those used during training. Comparative analysis between single-teacher and multi-teacher distillation demonstrates that the proposed MTKD-AR framework achieves superior robustness and generalization against diverse adversarial attacks.

6. Main Results

6.1. Comparison with Adversarially Trained Models

We evaluated the proposed MTKD-AR framework against adversarially trained models using teacher models trained with a perturbation magnitude of $\epsilon = 0.1$. The results in Table 1(A), 1(B), and 1(C) demonstrate that MTKD-AR consistently outperforms comparative methods, despite training the student model solely on clean data. This validates the effectiveness of the proposed framework in achieving robust performance against both training-time and test-time adversarial attacks.

The results further reveal the limitations of single-teacher adversarially trained models, which perform well only on the attacks encountered during training but show significant performance degradation against unseen attacks. In contrast, MTKD-AR maintains consistently high accuracy across diverse adversarial settings, highlighting its superior generalization and robustness. By leveraging the complementary strengths of multiple teacher models through adaptive learning, MTKD-AR effectively transfers robust knowledge while preserving strong performance on clean data.

6.2. Adversarial Robustness Distillation Results

The results on the MNIST-Digits dataset in Table 2(A) show that the CNN_{WOKD} baseline achieves high clean accuracy but is highly susceptible to adversarial attacks, while the CNN_{WKD-1T} model improves robustness only against the attack used during training, exhibiting limited cross-attack generalization. In contrast, the proposed MTKD-AR consistently achieves superior robustness across all attack settings while maintaining high clean-data accuracy, demonstrating the effectiveness of adaptive multi-teacher knowledge distillation. Similar trends are observed on the MNIST-Fashion dataset in Table 2(B), where MTKD-AR outperforms both the baseline and single-teacher models across diverse adversarial attacks. By adaptively combining knowledge from multiple adversarially trained teachers, the proposed framework effectively balances clean accuracy and adversarial robustness. The results on the more challenging CIFAR-10 dataset in Table 2(C) further confirm the generalization capability of

Table 1
Comparative Analysis of the Proposed Multi-Teacher Knowledge Distillation with adversarially trained models: Evaluation on Clean and Perturbed Test Data of the MNIST-Digits, MNIST-Fashion, and CIFAR-10 Datasets with a Perturbation Budget of (ϵ = 0.1). The numbers highlighted as best and runner-up indicate the top and second-best performances, respectively.

Table 1(A): MNIST-Digits Dataset Results

Method	Clean data	FGSM	FFGSM	RFGSM	PGD
CNN_{WOKD}	99.05	21.92	40.40	11.66	12.25
$CNN_{ADV-FGSM}$	99.03	99.03	–	–	–
$CNN_{ADV-FFGSM}$	96.89	–	91.32	–	–
$CNN_{ADV-RFGSM}$	97.43	–	–	93.29	–
$CNN_{ADV-PGD}$	94.11	–	–	–	90.94
Proposed (MTKD-AR)	97.66	95.79	93.10	93.76	92.98

Table 1(B): MNIST-Fashion Dataset Results

Method	Clean data	FGSM	FFGSM	RFGSM	PGD
CNN_{WOKD}	91.41	17.91	28.53	13.47	15.30
$CNN_{ADV-FGSM}$	92.09	97.34	–	–	–
$CNN_{ADV-FFGSM}$	91.18	–	87.48	–	–
$CNN_{ADV-RFGSM}$	88.68	–	–	86.56	–
$CNN_{ADV-PGD}$	87.99	–	–	–	85.91
Proposed (MTKD-AR)	91.19	90.69	90.50	90.74	89.92

Table 1(C): CIFAR-10 Dataset Results

Method	Clean data	FGSM	FFGSM	RFGSM	PGD
CNN_{WOKD}	84.72	15.82	24.91	11.73	13.08
$CNN_{ADV-FGSM}$	84.06	83.89	–	–	–
$CNN_{ADV-FFGSM}$	83.84	–	83.56	–	–
$CNN_{ADV-RFGSM}$	81.76	–	–	81.55	–
$CNN_{ADV-PGD}$	81.21	–	–	–	81.37
Proposed (MTKD-AR)	83.91	83.80	84.47	82.12	82.76

Table 2
Comparative Analysis of Results: Evaluating the Impact of Multi-Teacher Knowledge Distillation (MTKD-AR) in Contrast to Single Teacher Knowledge Distillation (WKD-1T) under Diverse Adversarial Attacks on the MNIST-Digits, MNIST-Fashion, and CIFAR-10 Datasets with a Perturbation Budget of (ϵ = 0.1). The numbers highlighted as best and runner-up indicate the top and second-best performances, respectively.

Table 2(A): MNIST-Digits Dataset Results

Method	Clean data	FGSM	FFGSM	RFGSM	PGD
CNN_{WOKD}	99.05	21.92	40.40	11.66	12.25
$CNN_{WKD\_1T(FGSM)}$	97.22	95.55	51.16	48.36	52.79
$CNN_{WKD\_1T(FFGSM)}$	94.60	39.67	91.32	44.53	49.41
$CNN_{WKD\_1T(RFGSM)}$	95.11	44.60	49.00	93.29	47.53
$CNN_{WKD\_1T(PGD)}$	93.41	40.91	36.67	41.95	90.94
Proposed (MTKD-AR)	97.66	95.79	93.10	93.76	92.98

Table 2(B): MNIST-Fashion Dataset Results

Method	Clean data	FGSM	FFGSM	RFGSM	PGD
CNN_{WOKD}	91.41	17.91	28.53	13.47	15.30
$CNN_{WKD\_1T(FGSM)}$	88.45	87.16	41.36	51.90	47.13
$CNN_{WKD\_1T(FFGSM)}$	89.92	44.59	89.54	40.18	51.39
$CNN_{WKD\_1T(RFGSM)}$	90.59	34.11	48.93	88.47	49.10
$CNN_{WKD\_1T(PGD)}$	88.19	37.43	46.71	55.03	87.94
Proposed (MTKD-AR)	91.19	90.69	90.50	90.74	89.82

Table 2(C): CIFAR-10 Dataset Results

Method	Clean data	FGSM	FFGSM	RFGSM	PGD
CNN_{WOKD}	84.72	15.82	24.91	11.73	13.08
$CNN_{WKD\_1T(FGSM)}$	83.18	82.21	46.82	44.37	47.15
$CNN_{WKD\_1T(FFGSM)}$	81.42	36.74	81.93	40.86	43.98
$CNN_{WKD\_1T(RFGSM)}$	81.87	40.25	44.11	80.34	43.12
$CNN_{WKD\_1T(PGD)}$	80.95	37.06	33.44	38.57	80.78
Proposed (MTKD-AR)	83.91	83.80	84.47	82.12	82.76

Table 3
Comparative evaluation of quantitative results on the MNIST-Digits dataset under different adversarial attacks and perturbation magnitudes. The numbers highlighted as best and runner-up indicate the top and second-best performances, respectively.

		FGSM			FFGSM			RFGSM			PGD		
Method	Clean	ϵ=0.1	ϵ=0.2	ϵ=0.3	ϵ=0.1	ϵ=0.2	ϵ=0.3	ϵ=0.1	ϵ=0.2	ϵ=0.3	ϵ=0.1	ϵ=0.2	ϵ=0.3
CNN_{WOKD}	99.05	21.92	19.46	17.33	40.40	34.52	31.19	11.66	9.13	8.56	12.25	10.71	9.11
$CNN_{WKD\_1T(FGSM)}$	97.22	95.55	93.74	89.28	51.16	48.39	47.45	48.36	44.82	50.74	52.79	43.40	41.58
$CNN_{WKD\_1T(FFGSM)}$	94.60	39.67	36.57	35.14	91.32	89.63	87.94	44.53	41.53	39.89	49.41	46.09	41.63
$CNN_{WKD\_1T(RFGSM)}$	95.11	44.60	42.81	40.57	49.00	45.16	39.12	93.29	90.51	88.41	47.53	44.13	39.27
$CNN_{WKD\_1T(PGD)}$	93.41	40.91	37.36	35.00	36.67	33.15	29.67	41.95	37.39	35.56	90.94	88.69	87.85
Proposed (MTKD-AR)	97.66	95.79	94.13	90.77	93.10	92.12	90.53	93.76	91.51	89.87	92.98	90.06	89.33

MTKD-AR. Although CNN_{WOKD} achieves the highest clean accuracy, it remains highly vulnerable to adversarial perturbations, whereas MTKD-AR consistently delivers the best robustness across all evaluated attacks while maintaining competitive clean-data performance.

Table 4
Comparative evaluation of quantitative results for our proposed approach on the MNIST-Fashion dataset under varied adversarial attacks with different perturbation magnitudes. The numbers highlighted as best and runner-up indicate the top and second-best performances, respectively.

Method	Clean	FGSM			FFGSM			RFGSM			PGD		
		ϵ=0.1	ϵ=0.2	ϵ=0.3	ϵ=0.1	ϵ=0.2	ϵ=0.3	ϵ=0.1	ϵ=0.2	ϵ=0.3	ϵ=0.1	ϵ=0.2	ϵ=0.3
CNN_{WOKD}	91.41	17.91	14.62	11.39	28.53	25.49	23.98	13.47	10.55	8.13	15.30	12.45	10.56
$\text{CNN}_{WKD\_1T(\text{FGSM})}$	88.45	87.16	86.81	83.51	41.36	38.97	34.65	51.90	49.63	45.91	47.13	41.00	37.25
$\text{CNN}_{WKD\_1T(\text{FFGSM})}$	89.92	44.59	42.08	39.24	89.54	87.10	85.45	40.18	36.48	32.52	51.39	47.21	40.31
$\text{CNN}_{WKD\_1T(\text{RFGSM})}$	90.59	34.11	30.85	26.13	48.93	42.91	38.23	88.47	87.15	85.14	49.10	45.62	39.28
$\text{CNN}_{WKD\_1T(\text{PGD})}$	88.19	37.43	33.44	29.84	46.71	42.64	39.68	55.03	49.28	44.30	88.94	87.69	86.85
Proposed (MTKD-AR)	91.91	90.69	88.13	86.15	90.50	89.12	87.64	90.74	89.51	86.11	90.82	89.06	88.47

Table 5
Comparative analysis of adversarially trained best student mode, single best teacher model, MTKD-AR with uniform weight averaging, and MTKD-AR with proposed adaptive weighting on CIFAR-10 dataset.

Method	Clean data	FGSM	FFGSM	RFGSM	PGD
Adversarially Trained Best Student Model ($\text{CNN}_{ADV-FGSM}$)	84.06	83.89	–	–	–
Single Best Teacher	83.18	82.21	46.82	44.37	47.15
MTKD-AR (Uniform Weight Averaging)	82.70	81.93	82.56	81.0	81.29
MTKD-AR (Adaptive Weighting)	83.91	83.80	84.47	82.12	82.76

6.3. Ablation Study

This section presents additional results on MNIST-Digits and MNIST-Fashion under adversarial perturbations with varying magnitudes (ϵ = 0.1, 0.2, 0.3). Tables 3 and 4 compare MTKD-AR with baseline and single-teacher models on clean and adversarial data. MTKD-AR consistently outperforms all competing methods across different attack strengths, demonstrating the effectiveness of its adaptive learning mechanism in transferring robustness from multiple adversarially trained teachers to the student model. It achieves competitive clean accuracy while providing significant improvements under adversarial conditions compared to single-teacher and ensemble baselines. Unlike conventional approaches that require adversarial training of the student, MTKD-AR enhances robustness without exposing the student to perturbed samples during training, improving scalability and generalization. While all methods experience accuracy degradation as perturbation strength increases, MTKD-AR maintains stable performance across attack levels, highlighting its adaptability for robust deployment. Furthermore, Table 5 evaluates the proposed adaptive weighting strategy against the adversarially trained best student, the single best teacher, and MTKD-AR with uniform weight averaging on CIFAR-10. MTKD-AR achieves superior performance on both clean and adversarial examples, confirming that adaptive weighting enables more effective knowledge transfer and improved robustness compared to uniform averaging.

7. Conclusion

In this paper, we proposed a novel framework, MTKD-AR, to enhance the adversarial robustness of convolutional neural networks through adaptive learning-driven multi-teacher knowledge distillation. Our approach leverages multiple adversarially trained teacher models to supervise a student model trained solely on clean data, eliminating the need for adversarial samples during student training. To ensure effective knowledge transfer, we introduced an adaptive learning strategy that dynamically assigns importance weights to teacher models based on their prediction performance. Extensive experiments across diverse adversarial settings demonstrate that the proposed framework significantly improves robustness against various adversarial attacks while maintaining strong performance on clean data. The obtained results validate the effectiveness of the proposed multi-teacher adversarial distillation framework and highlight its potential as an efficient and scalable solution for robust and trustworthy AI systems.

Declaration on Generative AI

During the preparation of this work, the author(s) used GPT-5 for grammar and language refinement. The content generated with the assistance of these tools was carefully reviewed and refined prior to its inclusion in the final manuscript. Subsequently, the authors thoroughly revised the manuscript as necessary and assume full responsibility for the accuracy, originality, and integrity of the published work.

References

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

[2] R. Girshick, Fast r-cnn, in: Proceedings of the IEEE international conference on computer vision, 2015, pp. 1440–1448.

[3] D. Amodei, S. Ananthanarayanan, R. Anubhai, J. Bai, E. Battenberg, C. Case, J. Casper, B. Catanzaro, Q. Cheng, G. Chen, et al., Deep speech 2: End-to-end speech recognition in english and mandarin, in: International conference on machine learning, PMLR, 2016, pp. 173–182.

[4] 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).

[5] I. Goodfellow, J. Shlens, C. Szegedy, Explaining and harnessing adversarial examples. proceedings of the 3rd international conference on learning representations, iclr 2015 (2015).

[6] S. A. M. Zaidi, L. Shamir, W. Hsu, S. Dietrich, T. Zaidi, Graze: Grounded refinement and motion-aware zero-shot event localization, arXiv preprint arXiv:2604.01383 (2026).

[7] C. Guo, M. Rana, M. Cisse, L. Van Der Maaten, Countering adversarial images using input transformations, arXiv preprint arXiv:1711.00117 (2017).

[8] F. Liao, M. Liang, Y. Dong, T. Pang, X. Hu, J. Zhu, Defense against adversarial attacks using high-level representation guided denoiser, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 1778–1787.

[9] X. Ma, B. Li, Y. Wang, S. M. Erfani, S. Wijewickrema, G. Schoenebeck, D. Song, M. E. Houle, J. Bailey, Characterizing adversarial subspaces using local intrinsic dimensionality, arXiv preprint arXiv:1801.02613 (2018).

[10] K. Lee, K. Lee, H. Lee, J. Shin, A simple unified framework for detecting out-of-distribution samples and adversarial attacks, Advances in neural information processing systems 31 (2018).

[11] J. Cohen, E. Rosenfeld, Z. Kolter, Certified adversarial robustness via randomized smoothing, in: international conference on machine learning, PMLR, 2019, pp. 1310–1320.

[12] D. Meng, H. Chen, Magnet: a two-pronged defense against adversarial examples, in: Proceedings of the 2017 ACM SIGSAC conference on computer and communications security, 2017, pp. 135–147.

[13] S. A. M. Zaidi, W. Hsu, S. Dietrich, Vits for action classification in videos: An approach to risky tackle detection in american football practice videos, arXiv preprint arXiv:2604.01318 (2026).

[14] A. Mustafa, S. H. Khan, M. Hayat, J. Shen, L. Shao, Image super-resolution as a defense against adversarial attacks, IEEE Transactions on Image Processing 29 (2019) 1711–1724.

[15] N. Das, M. Shanbhogue, S.-T. Chen, F. Hohman, L. Chen, M. E. Kounavis, D. H. Chau, Keeping the bad guys out: Protecting and vaccinating deep learning with jpeg compression, arXiv preprint arXiv:1705.02900 (2017).

[16] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, A. Vladu, Towards deep learning models resistant to adversarial attacks, arXiv preprint arXiv:1706.06083 (2017).

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

[18] M. Goldblum, L. Fowl, S. Feizi, T. Goldstein, Adversarially robust distillation, in: Proceedings of the AAAI Conference on Artificial Intelligence, volume 34, 2020, pp. 3996–4003.

[19] 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).
[20] T. Chen, Z. Zhang, S. Liu, S. Chang, Z. Wang, Robust overfitting may be mitigated by properly learned smoothening, in: International Conference on Learning Representations, 2020.
[21] 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).
[22] F. Tramèr, A. Kurakin, N. Papernot, I. Goodfellow, D. Boneh, P. McDaniel, Ensemble adversarial training: Attacks and defenses, arXiv preprint arXiv:1705.07204 (2017).
[23] Z. Xu, A. Shafahi, T. Goldstein, Exploring model robustness with adaptive networks and improved adversarial training, arXiv preprint arXiv:2006.00387 (2020).
[24] A. Kurakin, I. Goodfellow, S. Bengio, Y. Dong, F. Liao, M. Liang, T. Pang, J. Zhu, X. Hu, C. Xie, et al., Adversarial attacks and defences competition, in: The NIPS'17 Competition: Building Intelligent Systems, Springer, 2018, pp. 195–231.
[25] S.-M. Moosavi-Dezfooli, A. Fawzi, P. Frossard, Deepfool: a simple and accurate method to fool deep neural networks, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 2574–2582.
[26] N. Carlini, D. Wagner, Towards evaluating the robustness of neural networks, in: 2017 ieee symposium on security and privacy (sp), Ieee, 2017, pp. 39–57.
[27] J. Uesato, B. O'donoghue, P. Kohli, A. Oord, Adversarial risk and the dangers of evaluating against weak attacks, in: International Conference on Machine Learning, PMLR, 2018, pp. 5025–5034.
[28] B. Wu, J. Chen, D. Cai, X. He, Q. Gu, Do wider neural networks really help adversarial robustness?, Advances in Neural Information Processing Systems 34 (2021) 7054–7067.
[29] Y. Carmon, A. Raghunathan, L. Schmidt, J. C. Duchi, P. S. Liang, Unlabeled data improves adversarial robustness, Advances in neural information processing systems 32 (2019).
[30] C. Song, K. He, L. Wang, J. E. Hopcroft, Improving the generalization of adversarial training with domain adaptation, arXiv preprint arXiv:1810.00740 (2018).
[31] H. Zhang, Y. Yu, J. Jiao, E. Xing, L. El Ghaoui, M. Jordan, Theoretically principled trade-off between robustness and accuracy, in: International conference on machine learning, PMLR, 2019, pp. 7472–7482.
[32] 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.
[33] D. Wu, S.-T. Xia, Y. Wang, Adversarial weight perturbation helps robust generalization, Advances in Neural Information Processing Systems 33 (2020) 2958–2969.
[34] A. Shafahi, M. Najibi, M. A. Ghiasi, Z. Xu, J. Dickerson, C. Studer, L. S. Davis, G. Taylor, T. Goldstein, Adversarial training for free!, Advances in Neural Information Processing Systems 32 (2019).
[35] S. A. Taghanaki, K. Abhishek, S. Azizi, G. Hamarneh, A kernelized manifold mapping to diminish the effect of adversarial perturbations, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 11340–11349.
[36] T. Pang, K. Xu, C. Du, N. Chen, J. Zhu, Improving adversarial robustness via promoting ensemble diversity, in: International Conference on Machine Learning, PMLR, 2019, pp. 4970–4979.