



Dynamic Guidance Adversarial Distillation with Enhanced Teacher Knowledge

Hyejin Park¹  and Dongbo Min^{1*} 

Ewha Womans University, Seoul, South Korea
{clrara, dbmin}@ewha.ac.kr
<http://cv1.ewha.ac.kr>

Abstract. In the realm of Adversarial Distillation (AD), strategic and precise knowledge transfer from an adversarially robust teacher model to a less robust student model is paramount. Our Dynamic Guidance Adversarial Distillation (DGAD) framework directly tackles the challenge of differential sample importance, with a keen focus on rectifying the teacher model’s misclassifications. DGAD employs Misclassification-Aware Partitioning (MAP) to dynamically tailor the distillation focus, optimizing the learning process by steering towards the most reliable teacher predictions. Additionally, our Error-corrective Label Swapping (ELS) corrects misclassifications of the teacher on both clean and adversarially perturbed inputs, refining the quality of knowledge transfer. Further, Predictive Consistency Regularization (PCR) guarantees consistent performance of the student model across both clean and adversarial inputs, significantly enhancing its overall robustness. By integrating these methodologies, DGAD significantly improves upon the accuracy of clean data and fortifies the model’s defenses against sophisticated adversarial threats. Our experimental validation on CIFAR10, CIFAR100, and Tiny ImageNet datasets, employing various model architectures, demonstrates the efficacy of DGAD, establishing it as a promising approach for enhancing both the robustness and accuracy of student models in adversarial settings. The code is available at <https://github.com/kunsaram01/DGAD>.

Keywords: Adversarial Attack and Defense · Adversarial Training · Adversarial Distillation

1 Introduction

Deep Neural Networks (DNNs) have significantly advanced the frontiers of image classification [12, 19], speech recognition [11, 30], and natural language processing [6, 29], demonstrating remarkable success across a spectrum of complex tasks. Despite these advancements, their susceptibility to adversarial attacks [10, 28] poses a critical challenge, particularly in safety-sensitive domains such as autonomous vehicles [8, 27] and medical diagnostics [15, 21]. This vulnerability becomes even more pronounced in lightweight models designed for resource-constrained environments, where their limited capacity undermines robustness.

* Corresponding author.

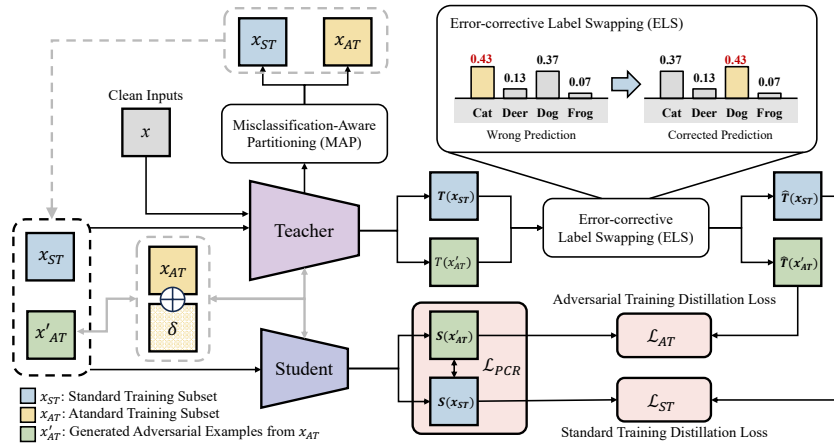


Fig. 1: The overview of Dynamic Guidance Adversarial Distillation (DGAD) framework. The DGAD framework refines adversarial distillation by employing a strategic approach: Misclassification-Aware Partitioning (MAP) categorizes inputs for tailored learning, Error-corrective Label Swapping (ELS) fixes teacher’s mispredictions, and Predictive Consistency Regularization (PCR) maintains learning uniformity. Together, these methods improve student model accuracy and robustness. $S(\cdot)$ and $T(\cdot)$ are the predictions of the student and teacher models, while $\hat{T}(\cdot)$ is the corrected teacher predictions after ELS.

Adversarial Training (AT) [22, 24, 32] has emerged as a crucial strategy to enhance the resilience of DNNs against adversarial attacks by training with adversarial examples. Although effective, the benefits of AT are more pronounced in larger models, leaving smaller models vulnerable due to their reduced capacity to handle adversarial perturbations. This limitation has led to the exploration of Adversarial Distillation (AD) [9, 16, 23, 33] as a method to transfer the robustness and accuracy from a larger, well-trained robust teacher model to a smaller, less robust student model, aiming to bridge the performance gap under adversarial conditions.

An often-overlooked issue in AD is the direct transfer of knowledge from the teacher to the student model without addressing potential inaccuracies in the predictions of the teacher. This oversight can significantly compromise the robustness and accuracy of the student model. In response to this challenge, recent advancements in AT have developed distinct treatment of samples according to their classification status. Methods such as Margin Maximization [4, 7] and Misclassification-Aware [1, 31] strategies have demonstrated that an indiscriminate approach—particularly using adversarial examples generated from misclassified clean inputs—can decrease model robustness. These findings underscore the necessity for AD to adopt a more thoughtful and strategic approach to knowledge transfer, specifically focusing on correcting teacher errors to effectively enhance the adversarial resilience of the student model.

In this study, we introduce the **Dynamic Guidance Adversarial Distillation (DGAD)** framework (see Fig. 1), embodying the principle of ‘dynamic guidance’. This concept transcends the traditional static approach to weighting distillation processes for clean and adversarial inputs by employing dynamic weighting to optimize the distillation focus. Dynamic guidance entails the real-time recognition and partitioning of training inputs within a batch, based on the teacher model’s misclassification status of clean inputs. It is followed by the immediate correction of any misclassified labels for both segregated clean and adversarial inputs during distillation. By pinpointing and separating misclassified samples, DGAD enables a custom distillation strategy that optimally addresses both standard and adversarial training needs. We employ three key interventions within this framework to ensure the precise and effective transfer of knowledge to the student model: 1) **Misclassification-Aware Partitioning (MAP)**: To realize dynamic weighting, this strategy separates the training dataset into two subsets based on the prediction of the teacher on clean inputs—The *Standard Training (ST)* subset comprises clean inputs incorrectly classified by the teacher, emphasizing correction of these misclassifications during standard training. Conversely, the *Adversarial Training (AT)* subset includes clean inputs correctly classified by the teacher, using adversarially perturbed versions of these inputs to increase the resistance of the student model to adversarial attacks. 2) **Error-corrective Label Swapping (ELS)**: Building upon the MAP, ELS is applied to inputs where the teacher’s predictions remain incorrect, specifically including the ST subset and adversarial examples generated using the AT subset. By replacing the incorrect labels predicted by the teacher with the correct ones, ELS ensures that the student model learns from accurate labels, directly addressing and amending the teacher’s prediction errors observed during distillation. 3) **Predictive Consistency Regularization (PCR)**: PCR addresses the imbalance between standard and adversarial training caused by the separate learning of ST and AT subsets in MAP. By regularizing the prediction consistency of the student model across the entire dataset, PCR ensures consistent predictions for both original inputs and their adversarial examples. This approach maintains balanced and effective learning, preventing biases toward any specific subset.

By integrating these innovative strategies, DGAD transcends traditional distillation enhancements, dynamically rectifying teacher model inaccuracies while fine-tuning the knowledge transfer. This dual-action approach not only elevates the student model’s defense against adversarial attacks but also significantly boosts its precision on clean data, setting a new standard for both robustness and accuracy in adversarial distillation.

2 Related Work

2.1 Adversarial Training

Adversarial Training (AT) [10, 22] is a defensive strategy against adversarial attacks, which aim to deceive machine learning models with subtly altered inputs.

The central goal of AT is to train models to accurately classify these manipulated inputs. However, treating all adversarially perturbed examples with the same target labels can lead to overfitting these adversarial examples. To address the trade-off between accuracy and robustness, approaches like Adversarial Logit Pairing (ALP) [17] focus on maintaining consistency between the logits of original and adversarial examples, while TRADES [32] and SCORE [24] introduce surrogate loss based on the Kullback-Leibler divergence and Squared Error loss, respectively, between the probability distributions of original and adversarial inputs.

Despite these advancements, previous research often overlooked whether adversarial examples were generated from correctly classified clean inputs. It has been highlighted that generating adversarial examples from misclassified images can exacerbate overfitting to adversarial examples. In response, methods such as MMA [7] and Misclassification-Aware Adversarial Training (MART) [31] suggest adjusting the weight of the adversarial perturbation or the loss function during adversarial training based on the misclassification of samples. These proposals underscore the importance of distinguishing between correctly classified and misclassified samples in generating adversarial examples, aiming to improve model robustness without compromising the model’s ability to generalize.

2.2 Adversarial Distillation

Adversarial Distillation (AD) emerged from the desire to convey the adversarial robustness of a well-trained teacher model to more compact student model. Adversarial Robust Distillation (ARD) [9] pioneered this realm by integrating Knowledge Distillation [14] with Adversarial Training. RSLAD [34] emphasized the significance of using robust soft labels in the inner optimization to generate adversarial examples. AdaAD [16] further refined this approach, optimizing the adversarial example generation to account for discrepancies between teacher and student predictions and leveraging these refined adversarial examples for more effective training. Introspective Adversarial Distillation (IAD) [33] addresses teacher’s unreliability in later training stages by incorporating a partial reliance on teacher’s predictions, increasingly favoring the student’s self-derived knowledge as training advances.

3 Preliminaries

There is active exploration into machine learning models based on the adversarial distillation (AD) that appropriately balance accuracy and resilience to adversarial attacks. Central to this challenge is an effective transfer of knowledge from an adversarially trained teacher model to a student model, aiming to instill both accuracy and robustness.

The foundation of our investigation is knowledge distillation (KD) [14], where a smaller student model is trained to mimic a more complex teacher model by

aligning its predictions with those of the teacher, following the objective function in Eq. (1):

$$\operatorname{argmin}_{\theta} (1 - \alpha) \cdot \mathcal{CE}(S_{\theta}(x), y) + \alpha \cdot \tau^2 \cdot \mathcal{KL}(S_{\theta}^{\tau}(x) || T^{\tau}(x)) \quad (1)$$

where \mathcal{CE} is the cross-entropy loss assessing the accuracy for an input x of the student with a ground truth y , \mathcal{KL} measures the disparity between the softened outputs of the student $S_{\theta}^{\tau}(x)$ with learnable parameters θ and a pretrained teacher model $T^{\tau}(x)$ modulated by a temperature parameter τ , and α weights the importance of classification accuracy versus prediction similarity.

Adversarial Robustness Distillation (ARD) [9] formulates the adversarial training in KD framework, harnessing the insights from a pretrained teacher model to guide a student model through adversarial scenarios. In contrast to AT, they use the predictions from the teacher model as reference signals. These signals aid the learning process of the student, encompassing both clean and adversarially perturbed inputs. AD adopts a min-max optimization framework that is similar to AT but is distinctively enhanced by the knowledge of the teacher model. The AD process is captured by the following optimization function:

$$\begin{aligned} & \operatorname{argmin}_{\theta} (1 - \alpha) \cdot \mathcal{CE}(S(x), y) + \alpha \cdot \mathcal{KL}(S(x'), T(x)) \\ & \text{where } x' = \operatorname{argmax}_{\|\delta\|_p < \epsilon} \mathcal{CE}(S_{\theta}(x + \delta), y). \end{aligned} \quad (2)$$

Robust Soft Label Adversarial Distillation (RSLAD) [34] showcases the use of robust soft labels, generated by a larger, robust teacher model, to guide the student model’s training on both clean and adversarial examples. This approach includes generating adversarial examples that leverage these robust soft labels for an enhanced training process. They apply robust soft labels in both of two processes:

$$\begin{aligned} & \operatorname{argmin}_{\theta} (1 - \alpha) \cdot \mathcal{KL}(S_{\theta}(x) || T(x)) + \alpha \cdot \mathcal{KL}(S_{\theta}(x') || T(x)), \\ & \text{where } x' = \operatorname{argmax}_{\|\delta\|_p < \epsilon} \mathcal{KL}(S_{\theta}(x + \delta), T(x)). \end{aligned} \quad (3)$$

In Eq. (3), the prediction deviation between the student and teacher models is assessed using both a clean input x and its corresponding adversarial example x' . The adversarial example is produced during an inner-maximization phase, where a deliberate perturbation δ is applied to the clean input within an ϵ -constrained sphere to maximize the divergence and hence challenge the model. The outer-minimization phase then involves the student model training, thereby bolstering the resilience of model to adversarial perturbations and mirroring robust predictive qualities of the teacher model.

Adaptive Adversarial Distillation (AdaAD) [16] further enhances adversarial distillation by integrating the teacher model in the generation of adversarial examples and guiding the student model with well-estimated probabilities for each

data point and its ϵ -neighborhood region. This approach mitigates model over-smoothness, thereby reducing the adversarial trade-offs for enhanced generalization. The AdaAD objective, shown in Eq. (4), emphasizes the teacher-directed adversarial learning:

$$\begin{aligned} & \underset{\theta}{\operatorname{argmin}} (1 - \alpha) \cdot \mathcal{KL}(S_{\theta}(x)||T(x)) + \alpha \cdot \mathcal{KL}(S_{\theta}(x')||T(x')), \\ & \text{where } x' = \underset{\|\delta\|_p < \epsilon}{\operatorname{argmax}} \mathcal{KL}(S_{\theta}(x + \delta), T(x + \delta)). \end{aligned} \tag{4}$$

A key advance lies in the inner-maximization process, which creates adversarial examples that maximize the discrepancy between the predictions of the student and teacher models on adversarially perturbed inputs. The teacher predictions on these adversarial examples are then used as supervisory signals to guide the training of the student model.

4 Dynamic Guidance Adversarial Distillation

4.1 Motivation of DGAD

In the realm of Adversarial Distillation (AD), reliance on static weighting for loss across all samples, notably in frameworks like AdaAD [16] and similar approaches [9, 23, 33, 34], often results in an imbalance between maintaining accuracy on original inputs and ensuring robustness against adversarial threats. This issue becomes more pronounced when adjusting the weighting parameter α , as depicted in Fig. 2. Static weights fail to account for the varying importance of individual samples, leading to a suboptimal balance between accuracy and robustness. This lack of consideration for sample importance means that some samples, particularly those misclassified by the teacher model, are not properly weighted during training. Consequently, the inaccuracies of the teacher model can disproportionately influence the training process, propagating these errors to the student model and undermining the overall effectiveness of the distillation process.

Our Dynamic Guidance Adversarial Distillation (DGAD) framework addresses this problem through dynamic adjustment of weighting, tailored to the precision of the teacher model’s predictions. DGAD hinges on a critical insight: the importance of each sample should be dynamically adjusted based on the accuracy of the teacher model’s predictions on clean inputs. When the teacher model’s predictions on clean inputs are incorrect, using these misclassified samples to generate adversarial examples can degrade the student’s learning experience during distillation. To mitigate this, DGAD dynamically adjusts the weights assigned to each sample based on whether the teacher model’s prediction for clean inputs is correct. By deliberately excluding misclassified samples from the adversarial generation process and focusing adversarial training on accurately classified samples, DGAD ensures that the student model’s training benefits from the most reliable information. For samples misclassified by the teacher model, the focus is on improving the student’s performance on clean data. This methodological

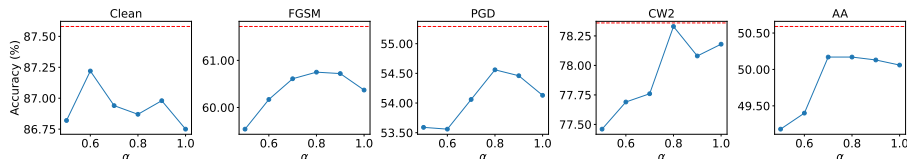


Fig. 2: Necessity of dynamically varying AD loss weights for individual samples. We compare the performance of AdaAD [16], which originally proposes to employ a static weight α in (Eq. (4)), against our Dynamic Guidance Adversarial Distillation (DGAD) that adapts the weight dynamically per sample as in Eq. (5). To validate the importance of dynamical weights, we adjust α for AdaAD and compare it across clean and adversarial scenarios. The blue solid line represents AdaAD’s performance with a fixed α across all samples, while the red dotted line indicates DGAD’s performance, showing improved accuracy due to its dynamic weighting approach.

pivot enhances the student model’s resilience to adversarial manipulations while either maintaining or improving accuracy on clean data, thereby strengthening the model’s overall performance.

The motivation behind DGAD is to navigate and mitigate the intrinsic trade-offs present in AD, employing a dynamic and discerning strategy for knowledge transfer that focuses exclusively on transmitting dependable insights from the teacher model. The subsequent sections will delve deeper into the strategies that embody this approach—Misclassification-Aware Partitioning (MAP), Error-corrective Label Swapping (ELS), and Predictive Consistency Regularization (PCR)—showcasing DGAD’s commitment to achieving both effectiveness and efficiency in enhancing model robustness and accuracy.

4.2 Misclassification-Aware Partitioning in AD

Dynamic sample weighting is essential for effective knowledge distillation. Our DGAD framework implements this through the Misclassification-Aware Partitioning (MAP) strategy, which categorizes the dataset into two subsets for specialized training: (1) **Standard Training Subset** (x_{ST}) consists of samples where the teacher model’s predictions for clean inputs x are incorrect, denoted as $x_{ST} = \{x \mid \arg \max(T(x)) \neq y\}$. These samples are used in the standard distillation path to improve the student’s accuracy on clean data; (2) **Adversarial Training Subset** (x_{AT}) consists of samples correctly classified by the teacher, represented as $x_{AT} = \{x \mid \arg \max(T(x)) = y\}$. These samples are used for adversarial distillation to enhance the student model’s robustness against adversarial perturbations.

This partitioning allows for targeted distillation, optimizing the contribution of each sample to the student’s learning process. The training objective combines standard and adversarial distillation to balance and streamline the learning pro-

cess:

$$\underset{\theta}{\operatorname{argmin}} \quad \underbrace{\mathcal{KL}(S_{\theta}(x_{ST})||T(x_{ST}))}_{\text{Standard Training Distillation } \mathcal{L}_{ST}} + \underbrace{\mathcal{KL}(S_{\theta}(x'_{AT})||T(x'_{AT}))}_{\text{Adversarial Training Distillation } \mathcal{L}_{AT}} \quad (5)$$

where $x'_{AT} = \underset{\|\delta\|_p < \epsilon}{\operatorname{argmax}} \mathcal{KL}(S_{\theta}(x_{AT} + \delta)||T(x_{AT} + \delta)).$

where x'_{AT} represents adversarially perturbed inputs from x_{AT} .

Standard Training Distillation (\mathcal{L}_{ST}): Aims to minimize the divergence between the student and teacher model predictions for the standard training subset x_{ST} , comprising samples inaccurately classified by the teacher. This focuses the distillation on enhancing the student’s accuracy on clean data, ensuring the student model learns more precisely from foundational data, contributing to overall performance improvements.

Adversarial Training Distillation (\mathcal{L}_{AT}): Targets adversarial resilience by distilling knowledge from adversarially perturbed examples x'_{AT} , derived from samples correctly identified by the teacher. This approach supports the student model in maintaining robustness in adversarial situations.

MAP’s approach of generating adversarial examples from accurately classified samples avoids propagating teacher model inaccuracies. This precision in knowledge transfer, as our ablation study (Tab. 2) demonstrates, substantially boosts the learning dynamics of the student model, marking a significant advance in adversarial distillation efficacy.

4.3 Error-corrective Label Swapping

Error-corrective Label Swapping (ELS) is a pivotal strategy designed to rectify inaccuracies in the predictions of the teacher model, especially focusing on samples misclassified after implementing MAP. ELS comes into play when a discrepancy is identified—specifically, when the teacher model wrongly places higher confidence in an incorrect label \hat{y} over the correct label y . This discrepancy is measured through a negative margin M , which triggers the corrective mechanism of ELS. By swapping the labels in such instances, ELS ensures that the student model receives and learns from correct labels, enhancing the precision and reliability of knowledge transfer. This corrective action is crucial for two scenarios.

All samples in **clean inputs** x_{ST} undergo label swapping to rectify the teacher’s initial misclassifications, ensuring x_{ST} contributes positively to the student’s learning. Here, y is a true label and \hat{y} is a predicted label, P_T is a softmax probability of the teacher model, and \hat{P}_T is the adjusted probability after swapping the incorrect prediction with the correct label:

$$\hat{P}_T \leftarrow \text{SWAP}(P_T(\hat{y}|x_{ST}), P_T(y|x_{ST})), \quad \forall x_{ST}. \quad (6)$$

Adversarial examples x'_{AT} are generated based on the student model using x_{AT} . According to IAD [33], the teacher’s predictions on x'_{AT} may become unreliable as student model training progresses. To prevent the propagation of these

unreliable predictions during later stages of training, ELS is applied only when the teacher’s predictions on x'_{AT} are incorrect. This ensures that the adversarial training of the student model is based on accurate teacher feedback:

$$\hat{P}_T \leftarrow \text{SWAP}(P_T(\hat{y}|x'_{AT}), P_T(y|x'_{AT})), \quad \text{if } M < 0, \quad (7)$$

where $M = P_T(y|x'_{AT}) - P_T(\hat{y}|x'_{AT})$, for generated x'_{AT} .

By systematically correcting these errors, ELS substantially enhances the quality of knowledge distilled to the student model and ensures a more effective and accurate learning process. This strategy is instrumental in overcoming the limitations posed by misclassifications, significantly contributing to the robustness and accuracy of the student model as demonstrated in our subsequent ablation studies. The training objectives, \mathcal{L}_{ST} for standard inputs and \mathcal{L}_{AT} for adversarial inputs, are refined through corrected teacher predictions $\hat{T}(\cdot)$ to ensure an optimal distillation path.

$$\underset{\theta}{\text{argmin}} \quad \underbrace{\mathcal{KL}(S_{\theta}(x_{ST})||\hat{T}(x_{ST}))}_{\text{Standard Training Distillation } \mathcal{L}_{ST}} + \underbrace{\mathcal{KL}(S_{\theta}(x'_{AT})||\hat{T}(x'_{AT}))}_{\text{Adversarial Training Distillation } \mathcal{L}_{AT}}. \quad (8)$$

4.4 Predictive Consistency Regularization

Predictive Consistency Regularization (PCR) directly addresses the challenge of maintaining consistency in the student model’s predictions across both Standard Training (ST) and Adversarial Training (AT) subsets. Given that ST focuses on correcting misclassifications of clean inputs and AT concentrates on enhancing resilience against adversarial perturbations, an inherent risk emerges: the student model might develop inconsistent responses to similar inputs under different contexts. PCR works to bridge this gap, ensuring that the student model applies a consistent learning approach to both subsets. By doing so, PCR mitigates the potential for divergent behaviors, fostering a unified model performance regardless of the input’s nature—clean or adversarially perturbed.

PCR introduced via \mathcal{L}_{PCR} , harmonizes the student model’s responses to clean (x) and their corresponding adversarial (x') inputs. This regularization approach is instrumental in fostering a balanced learning process, as evidenced by our ablation study results in Tab. 2. Here, $\mathcal{L}_{PCR} = \|S_{\theta}(x) - S_{\theta}(x')\|_2$. The comprehensive approach to adversarial distillation is encapsulated in the total loss \mathcal{L}_{DGAD} , defined as follows:

$$\mathcal{L}_{DGAD} = \mathcal{L}_{ST} + \mathcal{L}_{AT} + \beta \cdot \mathcal{L}_{PCR}, \quad (9)$$

This loss function strategically emphasizes predictive consistency through the parameter β , enhancing the student model’s accuracy and robustness in a comprehensive manner.

Implementing insights from AT research [24,32], PCR distinctly tailors these principles to our framework in AD, achieving a strategic balance between accuracy and adversarial resilience. Along with MAP and ELS, significantly elevates defense mechanisms against adversarial threats, a claim substantiated by our ablation study’s robust performance enhancements against a range of attacks.

Table 1: Performance of Teacher Models on CIFAR10/ CIFAR100 and Tiny-ImageNet.

Dataset	Teacher	Clean	FGSM	PGD	CW	AA
CIFAR10	ResNet18 [32]	82.94	59.02	53.71	77.04	49.34
CIFAR10	WideResNet-34-10 [25]	87.20	62.14	55.90	77.80	51.79
CIFAR10	WideResNet-34-20 [3]	86.03	66.01	63.33	82.60	57.71
CIFAR100	WideResNet-34-10 [3]	64.07	39.83	36.61	56.22	30.57
Tiny ImageNet	PreActResNet18 [13]	46.04	22.36	20.85	41.00	15.45

5 Experiments

Experimental Setup. The performance of DGAD was assessed on the CIFAR10, CIFAR100 [18], Tiny ImageNet [20] datasets, normalized between [0,1]. Benchmarks included PGD-AT [22], TRADES [32], and several AD methods (ARD [9], IAD [33], RSLAD [34], AKD [23], AdaAD [16]). We employed ResNet18 [12] and MobileNetV2 [26] as students, and WideResNet-34-10 (both datasets), WideResNet-34-20 (CIFAR10) [3,25], PreActResNet18 [13] (Tiny ImageNet) as teachers. Tab. 1 provides the performance of the teacher models used in our experiments. For fair comparison, models were trained following AdaAD’s basic settings. We used SGD with an initial learning rate of 0.1, momentum of 0.9, weight decay of $5e-4$, and standard data augmentation. Training duration varied: PGD-AT stopped at 110 epochs, TRADES and AD methods [9, 23, 33, 34], including DGAD, ran for 200 epochs with learning rate adjustments at epochs 100 and 150. Inner optimization parameters for adversarial training included 10 iterations, a step size of $2/255$, and a perturbation bound of $8/255$ under L_∞ constraint. Hyper-parameters α and distillation temperature τ were set as recommended. For the loss function, \mathcal{L}_{PCR} weight β was set to 5 for ResNet18, 10 for MobileNetV2, and 15 for PreActResNet18 models. Experiments were conducted in PyTorch with an adversarial training library.

Evaluation Metrics. Model performance is gauged through natural accuracy on clean samples and robust accuracy against adversarial samples, tested using FGSM [10], PGD [22], CW2 [2], and AutoAttack (AA) [5]. Perturbation size for FGSM, PGD, and AA is set at $8/255$, with PGD utilizing 10 steps of $2/255$ each. CW2’s equilibrium constant is 0.1. Results reflect the best PGD-10 checkpoint.

5.1 Ablation Study

Efficacy of Individual Components. The comprehensive ablation study presented in Tab. 2 meticulously dissects the distinct and combined influences of the proposed components, revealing a clear trajectory of performance enhancements and robustness against adversarial threats.

When Misclassification-Aware Partitioning (MAP) is applied independently, it yields a significant and vital enhancement in model robustness. This underscores the fundamental efficacy of MAP in directing the student model’s focus towards the most reliable predictions of the teacher.

Table 2: Efficacy of DGAD Components on CIFAR10. We utilize ResNet18 (student) and WideResNet-34-10 (teacher) to test components including Misclassification-Aware Partitioning (MAP), Error-corrective Label Swapping (ELS), and Predictive Consistency Regularization (PCR). Notations are as follows: x' - misclassified adversarial examples without consider misclassification on clean inputs, x_{ST} - misclassified clean inputs, x'_{AT} - misclassified adversarial examples.

Method	Clean	FGSM	PGD	CW	AA
Baseline [16]	86.75	60.37	54.13	78.18	50.06
+MAP	86.92	61.40	54.94	78.34	50.82
+MAP+PCR	87.19	61.14	54.92	78.72	50.52
+ELS(x')	87.27	60.66	54.47	78.42	50.13
+MAP+ELS(x'_{AT})	87.28	61.48	54.97	77.85	50.80
+MAP+ELS(x_{ST})	87.81	61.14	54.89	78.36	50.26
+MAP+ELS(x_{ST})+ELS(x'_{AT})	87.53	61.23	54.77	78.49	50.33
+MAP+ELS(x_{ST})+ELS(x'_{AT})+PCR	87.58	61.72	55.29	78.36	50.63

Table 3: Impact of different labeling methods in DGAD on CIFAR10. The experimental setup is identical to that described in Tab. 2.

Method	Clean	FGSM	PGD	CW	AA
Label Smoothing	87.24	61.97	56.09	77.51	49.86
Label Mixing	87.59	61.90	55.22	78.64	50.64
Label Swapping	87.58	61.72	55.29	78.36	50.63

Error-corrective Label Swapping (ELS) presents its own set of advantages. When ELS is applied to adversarial examples x' generated without MAP, we observe enhanced robustness and accuracy. Further improvements in robustness are noted when applying ELS on x'_{AT} after MAP, highlighting the benefits of excluding misclassified clean inputs and the crucial role of addressing misclassifications in bolstering adversarial resilience. Using ELS in x_{ST} with MAP amplifies performance. This demonstrates the critical role of rectifying teacher errors, as correcting misclassifications on clean data substantially boosts learning and robustness. Applying ELS to both x_{ST} and x'_{AT} with MAP enhances this effect, highlighting the synergy of these strategies in improving the performance of the student model.

Predictive Consistency Regularization (PCR) not only maintains robustness gains from MAP but also enhances accuracy on clean inputs, showcasing the synergistic effect of the two components. The integration of all components, including MAP and PCR, significantly outperforms configurations without PCR, highlighting the role of PCR in complementing and augmenting MAP and ELS.

The initial addition of MAP notably improved AA performance by 0.77%, with subsequent ELS and PCR enhancements showing smaller AA improvements. In total, these components resulted in an improvement of 0.57% over the baseline in AA. This non-linear improvement arises because our method aims to balance clean accuracy and adversarial robustness.

Effectiveness of Labeling Techniques in DGAD. Within the DGAD, we evaluate labeling techniques for correcting teacher misclassifications. Label

Table 4: Evaluating on CIFAR10. RN-18 and MN-V2 denote the student models ResNet-18 and MobileNetV2, respectively. Best results in **bold**; next-best underlined.

Teacher Model		WideResNet-34-10					WideResNet-34-20				
model	method	Clean	FGSM	PGD	CW2	AA	Clean	FGSM	PGD	CW2	AA
RN-18	PGD-AT [22]	82.95	57.16	52.87	77.56	47.69	82.95	57.16	52.87	77.56	47.69
	TRADES [32]	83.00	58.42	53.18	76.92	49.21	83.00	58.42	53.18	76.92	49.21
	ARD [9]	84.04	58.26	52.67	74.95	48.62	84.03	58.16	53.11	79.13	48.07
	IAD [33]	83.19	57.76	53.17	76.77	48.82	84.71	<u>61.28</u>	54.92	79.44	49.85
	RSLAD [34]	83.60	57.45	52.60	76.85	48.45	83.52	58.36	53.46	78.36	48.66
	AKD [23]	84.69	58.97	53.28	77.25	48.37	83.22	58.63	54.16	78.44	49.26
	AdaAD [16]	<u>86.75</u>	<u>60.37</u>	<u>54.13</u>	<u>78.18</u>	<u>50.06</u>	<u>85.58</u>	60.85	<u>56.40</u>	<u>80.83</u>	<u>51.37</u>
	DGAD	87.58	61.72	55.29	78.36	50.59	85.75	62.28	58.05	81.60	52.34
	+0.83	+1.35	+1.16	+0.17	+0.53	+0.17	+1.00	+1.65	+0.77	+0.97	
MN-V2	PGD-AT [22]	77.54	53.58	49.90	72.54	44.56	77.54	53.58	49.90	72.54	44.56
	TRADES [32]	79.80	54.84	50.51	75.30	45.67	79.80	54.84	50.51	75.30	45.67
	ARD [9]	84.63	58.00	50.82	72.93	46.48	79.56	53.17	49.06	74.51	44.04
	IAD [33]	82.11	55.27	50.20	75.41	45.66	83.31	58.29	52.98	78.03	47.11
	RSLAD [34]	83.24	56.69	51.57	76.52	47.18	81.11	56.39	51.66	76.20	46.75
	AKD [23]	82.64	56.17	50.49	75.31	45.67	83.41	<u>57.71</u>	52.35	77.97	46.82
	AdaAD [16]	<u>86.80</u>	<u>58.56</u>	<u>52.00</u>	<u>78.27</u>	<u>47.97</u>	<u>83.79</u>	57.29	<u>53.04</u>	<u>79.24</u>	<u>47.66</u>
	DGAD	87.19	60.11	53.56	79.40	49.19	85.30	61.20	56.77	80.98	51.10
	+0.39	+1.55	+1.56	+1.13	+1.22	+1.51	+3.49	+3.73	+1.74	+3.44	

Swapping is compared with Label Smoothing, represented as $(1 - \alpha) \cdot y + \alpha \cdot \frac{1}{C}$, where C is the number of classes, and Label Mixing, shown as $\alpha \cdot T(x) + (1 - \alpha) \cdot y$ (Tab. 3). While Label Smoothing slightly improves clean data accuracy, its impact on robustness varies. Label Mixing and Label Swapping show similar results in enhancing accuracy and robustness. However, Label Mixing’s reliance on the hyperparameter α can complicate corrections, especially with overly confident incorrect predictions. Label Swapping directly corrects misclassifications, simplifying the training process and ensuring precise knowledge transfer without complex parameters tuning. This highlights its advantage in adversarial training and provides a clear rationale for its use in DGAD.

5.2 Adversarial Robustness

CIFAR10/CIFAR100. We compared the performance of our DGAD with other existing methods on CIFAR10 and CIFAR100 datasets, focusing particularly on the best checkpoint results against PGD attacks, as established in [16].

For CIFAR10 evaluations in Tab. 4, our DGAD framework marks a distinct advancement in model performance. With the WideResNet-34-10 as the teacher model paired with ResNet18 as the student, DGAD has achieved a remarkable 0.83% uplift in clean data accuracy, surpassing the already impressive teacher model’s score with an overall accuracy of 87.20%. Moreover, DGAD’s fortification against attacks is evident with gains of 1.35% against FGSM and 1.16% against PGD attacks, showcasing a strengthened defense mechanism.

Switching to WideResNet-34-20 and MobileNetV2, DGAD achieves a notable 1.18% increase in clean accuracy and consistent robustness gains—3.49% against FGSM, 3.73% on PGD, 1.74% on CW2, and 3.44% on AA attacks—validating

Table 5: Evaluating on CIFAR100. RN-18 and MN-V2 denote the student models ResNet-18 and MobileNetV2, respectively. Best results in **bold**; next-best underlined.

Teacher Model		WideResNet-34-10				
Model	Method	Clean	FGSM	PGD	CW2	AA
RN-18	PGD-AT [22]	56.27	32.08	29.84	49.05	24.99
	TRADES [32]	57.82	32.52	30.38	51.30	25.02
	ARD [9]	60.94	35.31	32.72	53.67	26.04
	IAD [33]	60.43	35.75	32.80	52.71	26.84
	RSLAD [34]	59.55	35.68	33.35	52.89	27.77
	AKD [23]	57.84	34.32	31.98	51.06	26.06
	AdaAD [16]	<u>62.19</u>	<u>35.33</u>	<u>32.52</u>	<u>54.67</u>	26.74
	DGAD	63.24	36.09	33.68	55.47	27.66
	+1.05	+0.76	+1.16	+0.80	-0.11	
MN-V2	PGD-AT [22]	51.55	29.34	27.26	45.73	22.07
	TRADES [32]	53.05	29.07	27.44	47.62	21.82
	ARD [9]	57.18	33.13	30.91	51.50	24.20
	IAD [33]	56.33	32.88	30.18	49.00	24.07
	RSLAD [34]	56.04	32.76	30.29	50.14	24.56
	AKD [23]	56.75	33.11	30.50	49.53	24.65
	AdaAD [16]	<u>61.44</u>	<u>34.75</u>	<u>31.97</u>	<u>54.21</u>	<u>25.91</u>
	DGAD	62.25	34.90	32.64	54.54	26.56
	+0.81	+0.15	+0.67	+0.33	+0.65	

Table 6: Evaluating on Tiny ImageNet. The teacher model was trained with TRADES ($\lambda=6$) for 110 epochs.

Model	Method	Clean	FGSM	PGD	CW2	AA
RN-18	ARD [9]	41.66	24.47	23.30	37.76	17.23
	RSLAD [34]	40.83	23.45	22.58	37.12	17.05
	AdaAD [16]	47.54	24.22	22.82	42.79	17.41
	DGAD	47.92	24.42	23.05	42.81	17.18

its efficacy in improving both accuracy and defense in a cohesive manner. For WideResNet-34-10 with MobileNetV2, the results highlight its adaptability and strength in countering varied adversarial strategies while preserving or improving the accuracy of clean data.

For CIFAR100 in Tab. 5, DGAD has shown significant improvements over the AdaAD method. Specifically, for ResNet18 with WRN-34-10, there is an increase in clean accuracy by 1.05% and a boost in robustness against the PGD attack by 1.16%. For MobileNetV2, For the MobileNetV2 model, we observe a clean accuracy improvement of 0.81%, alongside a 0.67% uptick in PGD attack.

Tiny ImageNet. To evaluate performance on a more complex dataset, we tested DGAD on Tiny ImageNet using a PreActResNet18 teacher and a ResNet18 student model. As shown in Tab. 6, DGAD outperforms all other methods, including ARD, RSLAD, and AdaAD, achieving the highest accuracy on both clean and adversarial examples.

Transfer-based Attacks on CIFAR10. We tested DGAD against transfer-based attacks using surrogate models such as ResNet34 and VGG16 on CIFAR10. Our evaluation simulates real-world scenarios, where attackers lack specific details of the target models. In Tab. 7, DGAD outperforms all other methods,

Table 7: Evaluating Transfer-based Attacks Using Various Surrogate Models on CIFAR10 with a ResNet18 Target Model.

Surrogate Model Method	ResNet34			VGG16		
	FGSM	PGD	JSMA	FGSM	PGD	JSMA
PGD-AT [22]	63.05	60.58	84.90	64.06	62.78	85.77
TRADES [32]	65.57	63.93	84.71	66.88	66.00	85.36
ARD [9]	65.26	63.20	86.06	66.64	65.43	87.03
IAD [33]	65.48	63.23	84.71	66.62	66.06	86.04
RSLAD [34]	65.06	62.77	85.43	65.91	64.83	86.26
AKD [23]	64.34	62.23	85.22	65.24	64.30	86.24
AdaAD [16]	<u>66.81</u>	<u>64.57</u>	<u>88.00</u>	<u>68.74</u>	<u>67.89</u>	<u>88.39</u>
DGAD	67.77	65.20	90.56	70.92	70.29	90.34

Table 8: Evaluating Self-Adversarial Distillation on CIFAR10 using a TRADES ($\lambda = 6$) trained ResNet18 Teacher Model.

Method	Clean	FGSM	PGD	CW2	AA
PGD-AT [22]	82.95	57.16	52.87	77.56	47.69
ARD [9]	80.66	55.68	50.90	74.87	46.61
IAD [33]	81.32	57.54	52.91	75.69	48.20
RSLAD [34]	81.92	57.94	53.29	76.26	49.06
AKD [23]	83.74	<u>58.87</u>	<u>54.17</u>	77.97	48.84
AdaAD [16]	83.13	57.54	53.30	77.62	<u>49.61</u>
DGAD	<u>83.26</u>	58.91	54.37	<u>77.90</u>	50.54

including the state-of-the-art AdaAD, across all metrics, demonstrating superior transferability and robustness against diverse surrogate model-based attacks.

Adversarial Self-Distillation on CIFAR10. As shown in Tab. 8, DGAD outperforms both AdaAD and AKD [23] in adversarial resilience, achieving superior performance against FGSM, PGD, and AA attacks. While AKD is specifically designed for self-distillation within the same architecture, DGAD demonstrates superior performance in various setups, although it may occasionally lag behind AKD in scenarios tailored to AKD’s design.

6 Conclusion

In this study, we presented the Dynamic Guidance Adversarial Distillation (DGAD) framework, a novel strategy designed to enhance both the adversarial robustness and clean data accuracy of student models through a tailored approach in adversarial distillation. DGAD leverages Misclassification-Aware Partitioning (MAP), Error-corrective Label Swapping (ELS), and Predictive Consistency Regularization (PCR) to meticulously correct the inaccuracies in the teacher model’s predictions and fine-tune the student’s learning process.

Our findings affirm the effectiveness of DGAD, demonstrating substantial improvements in the model’s defense against adversarial threats and its accuracy on clean data. This advancement in adversarial and knowledge distillation sets new standards for developing resilient and accurate machine learning models, paving the way for future research in enhancing model robustness without compromising performance.

Acknowledgements

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. RS-2023-00222385) and partly by the Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (RS-2021-II212068, Artificial Intelligence Innovation Hub).

References

1. Altinisik, E., Messaoud, S., Sencar, H.T., Chawla, S.: A3t: accuracy aware adversarial training. *Machine Learning* **112**(9), 3191–3210 (2023)
2. 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)
3. Chen, E.C., Lee, C.R.: Ltd: Low temperature distillation for robust adversarial training. arXiv preprint arXiv:2111.02331 (2021)
4. Cheng, M., Lei, Q., Chen, P.Y., Dhillon, I., Hsieh, C.J.: Cat: Customized adversarial training for improved robustness. arXiv preprint arXiv:2002.06789 (2020)
5. Croce, F., Hein, M.: Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks. In: International conference on machine learning. pp. 2206–2216. PMLR (2020)
6. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018)
7. Ding, G.W., Sharma, Y., Lui, K.Y.C., Huang, R.: Mma training: Direct input space margin maximization through adversarial training. arXiv preprint arXiv:1812.02637 (2018)
8. Eykholt, K., Evtimov, I., Fernandes, E., Li, B., Rahmati, A., Xiao, C., Prakash, A., Kohno, T., Song, D.: Robust physical-world attacks on deep learning visual classification. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1625–1634 (2018)
9. Goldblum, M., Fowl, L., Feizi, S., Goldstein, T.: Adversarially robust distillation. In: Proceedings of the AAAI conference on artificial intelligence. vol. 34, pp. 3996–4003 (2020)
10. Goodfellow, I.J., Shlens, J., Szegedy, C.: Explaining and harnessing adversarial examples. arXiv preprint arXiv:1412.6572 (2014)
11. Graves, A., Mohamed, A.r., Hinton, G.: Speech recognition with deep recurrent neural networks. In: 2013 IEEE international conference on acoustics, speech and signal processing. pp. 6645–6649. IEEE (2013)
12. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
13. He, K., Zhang, X., Ren, S., Sun, J.: Identity mappings in deep residual networks. In: Computer Vision—ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part IV 14. pp. 630–645. Springer (2016)
14. Hinton, G., Vinyals, O., Dean, J., et al.: Distilling the knowledge in a neural network. arXiv preprint arXiv:1503.02531 **2**(7) (2015)

15. Hirano, H., Minagi, A., Takemoto, K.: Universal adversarial attacks on deep neural networks for medical image classification. *BMC medical imaging* **21**, 1–13 (2021)
16. Huang, B., Chen, M., Wang, Y., Lu, J., Cheng, M., Wang, W.: Boosting accuracy and robustness of student models via adaptive adversarial distillation. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 24668–24677 (2023)
17. Kannan, H., Kurakin, A., Goodfellow, I.: Adversarial logit pairing. *arXiv preprint arXiv:1803.06373* (2018)
18. Krizhevsky, A., Hinton, G., et al.: Learning multiple layers of features from tiny images (2009)
19. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems* **25** (2012)
20. Le, Y., Yang, X.: Tiny imagenet visual recognition challenge. *CS 231N* **7**(7), 3 (2015)
21. Ma, X., Niu, Y., Gu, L., Wang, Y., Zhao, Y., Bailey, J., Lu, F.: Understanding adversarial attacks on deep learning based medical image analysis systems. *Pattern Recognition* **110**, 107332 (2021)
22. Madry, A., Makelov, A., Schmidt, L., Tsipras, D., Vladu, A.: Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083* (2017)
23. Maroto, J., Ortiz-Jiménez, G., Frossard, P.: On the benefits of knowledge distillation for adversarial robustness. *arXiv preprint arXiv:2203.07159* (2022)
24. Pang, T., Lin, M., Yang, X., Zhu, J., Yan, S.: Robustness and accuracy could be reconcilable by (proper) definition. In: *International Conference on Machine Learning*. pp. 17258–17277. PMLR (2022)
25. Pang, T., Yang, X., Dong, Y., Su, H., Zhu, J.: Bag of tricks for adversarial training. *arXiv preprint arXiv:2010.00467* (2020)
26. Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., Chen, L.C.: Mobilenetv2: Inverted residuals and linear bottlenecks. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. pp. 4510–4520 (2018)
27. Sobh, I., Hamed, A., Kumar, V.R., Yogamani, S.: Adversarial attacks on multi-task visual perception for autonomous driving. *arXiv preprint arXiv:2107.07449* (2021)
28. Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., Fergus, R.: Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199* (2013)
29. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. *Advances in neural information processing systems* **30** (2017)
30. Wang, Y., Deng, X., Pu, S., Huang, Z.: Residual convolutional ctc networks for automatic speech recognition. *arXiv preprint arXiv:1702.07793* (2017)
31. Wang, Y., Zou, D., Yi, J., Bailey, J., Ma, X., Gu, Q.: Improving adversarial robustness requires revisiting misclassified examples. In: *International conference on learning representations* (2019)
32. Zhang, H., Yu, Y., Jiao, J., Xing, E., El Ghaoui, L., Jordan, M.: Theoretically principled trade-off between robustness and accuracy. In: *International conference on machine learning*. pp. 7472–7482. PMLR (2019)
33. Zhu, J., Yao, J., Han, B., Zhang, J., Liu, T., Niu, G., Zhou, J., Xu, J., Yang, H.: Reliable adversarial distillation with unreliable teachers. *arXiv preprint arXiv:2106.04928* (2021)

34. Zi, B., Zhao, S., Ma, X., Jiang, Y.G.: Revisiting adversarial robustness distillation: Robust soft labels make student better. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 16443–16452 (2021)