REVISITING SEMI-SUPERVISED ADVERSARIAL TRAINING VIA NOISE-AWARE ONLINE ROBUST DISTILLATION

Tsung-Han Wu[*1], **Hung-Ting Su**[2], **Shang-Tse Chen**[2], **Winston H. Hsu**[2]
[1]University of California, Berkeley, [2]National Taiwan University

ABSTRACT

Training adversarially robust models under a low-labeling regime is crucial for real-world deployment. Robust self-training (RST), with standard training for pseudo labels followed by adversarial robust training, has emerged as a key paradigm in this setting. Recent advancements in RST primarily focus on leveraging strong pre-trained models to improve robustness and performance. However, we find that these methods often overlook the critical role of pseudo labels in the training pipeline, leading to worse results on extremely low labeling regimes (<5%). In this work, we introduce SNORD, a simple yet effective approach that significantly improves robustness by enhancing pseudo-label quality in the first stage and effectively managing label noise in the second stage leveraging advanced standard semi-supervised learning techniques. Experiments on CIFAR-10, CIFAR-100, and TinyImageNet-200 demonstrate that SNORD outperforms prior methods by up to 22% in robust accuracy under low-labeling conditions. Furthermore, compared to fully supervised adversarial training, SNORD achieves 90% relative robust accuracy under $\ell_\infty = 8/255$ AutoAttack, requiring only 0.1%, 2%, and 10% labeled data for the three commonly used benchmarks, respectively. Additional analyses validate the contribution of each component and show that SNORD can be seamlessly integrated with existing adversarial pretraining strategies to further enhance robustness.

1 INTRODUCTION

The growing usage of machine learning models in safety-critical areas, like facial recognition systems, has highlighted their vulnerability to adversarial attacks (Rozsa et al., 2019). Despite advancements in adversarial defense strategies (Goodfellow et al., 2015; Madry et al., 2018; Zhang et al., 2019; 2021; Croce & Hein, 2020), reaching robustness under limited labeled data remains challenging (Schmidt et al., 2018). This issue has sparked a shift in research focus toward the semi-supervised learning (SSL) paradigm to enhance model robustness.

A dominant approach in this setting is robust self-training (RST), a two-stage pipeline that first generates pseudo labels via standard trained models and then applies adversarial training on the pseudo-labeled data (Carmon et al., 2019; Alayrac et al., 2019; Zhai et al., 2019). Recent research has primarily focused on improving RST by incorporating strong adversarial pretraining (Jiang et al., 2020; Luo et al., 2023), assuming that more powerful initialization enhances robustness. However, as

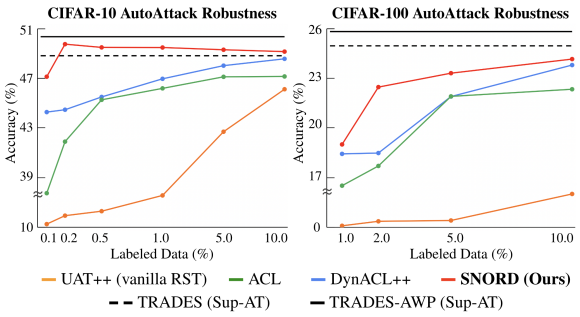

Figure 1: **Performance comparison of SSL adversarial training techniques.** While ACL (Jiang et al., 2020) and DynACL++ (Luo et al., 2023) enhance the robust self-training (RST) pipeline by leveraging strong adversarially pretrained models, their performance remains constrained by the inherent limitations of RST (Section 2). In contrast, our proposed SNORD framework surpasses these methods by a significant margin across various labeling budgets and datasets. Notably, on the CIFAR-10 dataset, SNORD attains comparable results to fully adversarial training methods like TRADES (Zhang et al., 2019) and TRADES-AWP (Wu et al., 2020) but requires merely **0.2% of the labeling effort (9 labeled data per class)**.

[*]Work done while at National Taiwan University (NTU).

illustrated in Figure 1, these advancements
fail under extremely low-labeling (few-shot) conditions, which are increasingly relevant in real-world applications.

Rather than solely improving adversarial pretraining, we argue that the fundamental bottleneck in semi-supervised adversarial training lies in pseudo-label quality and noise management, two aspects that remain underexplored. Specifically, our analysis in Figure 2 (b) reveals that pseudo labels in complex datasets like CIFAR-100 can have error rates exceeding 45% during the early training stages, leading to significant degradation in robustness. Moreover, as shown in Figure 2 (c), existing methods struggle to learn effectively from noisy pseudo-labeled data during adversarial training. For instance, UAT++ (Carmon et al., 2019), a baseline RST method, exhibits the lowest robustness due to its reliance on inaccurate, one-hot pseudo labels. While some works attempt to mitigate this by using adversarially pretrained models and soft-label distillation (Jiang et al., 2020; Luo et al., 2023), they suffer from poor standard accuracy, particularly when labeled data is extremely scarce (<5%).

To address these issues, we introduce a novel framework for SSL adversarial training, dubbed SNORD (**S**emi-supervised **N**oise-aware **O**nline **R**obust **D**istillation). SNORD is inspired by recent successes in standard SSL. Firstly, we utilize an off-the-shelf SSL algorithm (Sohn et al., 2020) for pseudo label generation, significantly improving label quality, as shown Figure 2 (b). Secondly, we present a novel noise-aware rectification strategy and an online robust distillation mechanism for the subsequent adversarial training process. The former further enhances pseudo label quality for downstream training by integrating entropy minimization techniques (Lee et al., 2013; Miyato et al., 2018) and accounting for noise in both labeled and unlabeled data. The latter, inspired by consistency regularization, allows the downstream adversarially trained model to learn from labels across different epochs. The whole pipeline of our SNORD is depicted in Figure 3.

Through extensive experiments on CIFAR-10, CIFAR-100, and TinyImageNet-200, we demonstrate that SNORD significantly outperforms prior SSL adversarial training methods across various labeling budgets. Notably, SNORD achieves 90% of the robust accuracy relative to full supervision under $\ell_\infty = 8/255$ AutoAttack, requiring only 0.1%, 2%, and 10% of labeled data for CIFAR-10, CIFAR-100, and TinyImageNet-200, respectively. Further ablation studies confirm that SNORD not only enhances robustness but also preserves standard accuracy, a common weakness in existing methods. Moreover, we show that SNORD is complementary to adversarial pretraining strategies, further improving robustness when combined with them. By shifting the focus from pretrained models to pseudo-label quality, SNORD redefines the SSL adversarial training paradigm and sets a new foundation for learning robust models under extreme data scarcity.

In summary, our contributions are:

1. We are the first to identify the two critical yet previously overlooked issues related to pseudo labels within the widely adopted RST pipeline, especially on extremely low labeling regimes.
2. We present SNORD, a simple, effective, and general SSL adversarial training framework to address the above challenges.
3. Our SNORD framework not only achieves SOTA performance in various robustness benchmarks but also demonstrates compatibility with existing adversarial pretraining methods.

2 Revisiting Semi-supervised Adversarial Training

In semi-supervised adversarial training, the objective is to develop a model that is robust against adversarial attacks while leveraging both a small labeled dataset and a larger pool of unlabeled data. Specifically, we focus on defending against ϵ-tolerant L_∞ attacks in image classification, an adversarial perturbation δ is added to the input x, subject to the constraint $||\delta||_\infty \leq \epsilon$. Formally, our goal is to minimize the following adversarial risk:

$$\min_{f_{rob}} \mathbb{E}_{x \in D} \left[\max_{||\delta||_\infty \leq \epsilon} \mathcal{L}(f_{rob}(x+\delta), y) \right] \tag{1}$$

where f_{rob} denotes the parameters of the robust model, D is the dataset, and $\mathcal{L}(f_{rob}(x+\delta), y)$ is a 0/1 error function measuring the model's performance under attack.

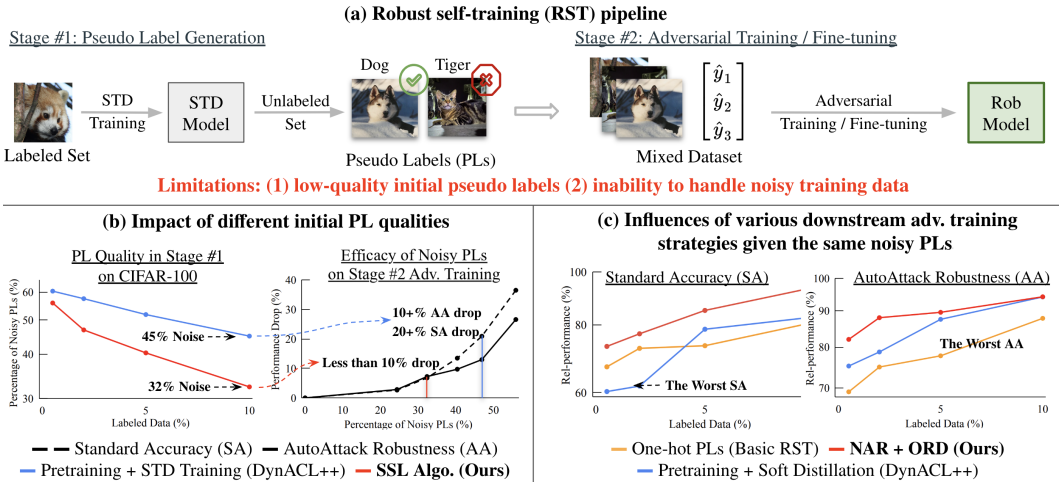

Figure 2: **Limitations of Existing Adversarial Training Methods.** (a) Overcoming the limitations of the two-stage RST framework requires improving pseudo-label (PL) quality in the initial stage and effectively handling noisy labels during downstream adversarial training. (b) On CIFAR-100, even SOTA RST-based methods like DynACL++ (Luo et al., 2023) produce low-quality pseudo labels, leading to suboptimal Standard Accuracy (SA) and Adversarial Accuracy (AA) compared to a fully supervised adversarially trained oracle model. By integrating a more advanced SSL-based pseudo-labeling mechanism, our approach significantly enhances performance under the same downstream adversarial training conditions. (c) We further analyze the impact of different adversarial training strategies when operating on equivalent noisy pseudo labels. The y-axis represents relative performance compared to the oracle model—a fully labeled adversarial training baseline. Results show that the basic RST method with hard labels performs the worst, suffering from degraded AA due to unreliable pseudo labels. While adversarial pretraining with soft-label distillation improves robustness (Jiang et al., 2020; Luo et al., 2023), it sacrifices SA in low-labeling regimes. In contrast, our proposed Noise-Aware Rectification (NAR) and Online Robust Distillation (ORD) strategies (Section 3) effectively mitigate these issues, achieving the best results on both metrics.

In this setting, robust self-training (RST) (Carmon et al., 2019; Alayrac et al., 2019; Zhai et al., 2019), a widely adopted two-stage training paradigm, has been instrumental in advancing semi-supervised adversarial training. As illustrated in Figure 2 (a), RST first generates pseudo labels using standard training and subsequently applies adversarial training on the pseudo-labeled data. While recent works (Jiang et al., 2020; Luo et al., 2023) have attempted to enhance RST by incorporating strong adversarially pretrained models, these methods inherently overlook a key limitation: the quality of pseudo labels. As shown in Figure 2 (b, c), we find that the primary bottleneck in semi-supervised adversarial training arises not from the choice of pretrained models but from the generation and management of pseudo labels during training.

Our analysis reveals that existing methods frequently suffer from high pseudo-label noise, leading to unreliable supervision and suboptimal robustness, particularly in extremely low-labeling regimes (such as those lower than 5%). Moreover, adversarial training on noisy pseudo labels further exacerbates performance degradation, limiting the effectiveness of the overall pipeline.

3 SNORD: A New Paradigm for Semi-Supervised Adversarial Training

To overcome the above-mentioned challenges, we introduce **SNORD** (**S**emi-supervised **N**oise-aware **O**nline **R**obust **D**istillation), a novel framework designed to directly tackle the pseudo-label quality bottleneck in RST. As shown in Figure 3, SNORD consists of three core components: (1) a reliable pseudo label generator, (2) the Noise-aware Rectification (NAR) strategy, and (3) the Online Robust Distillation (ORD) mechanism. We elaborate on each component in the following subsections.

3.1 Enhancing Pseudo Label Generation

The cornerstone of a robust self-training process heavily relies on the precision of pseudo labels in its initial phase (Gowal et al., 2021). Conventional approaches typically employ supervised training coupled with

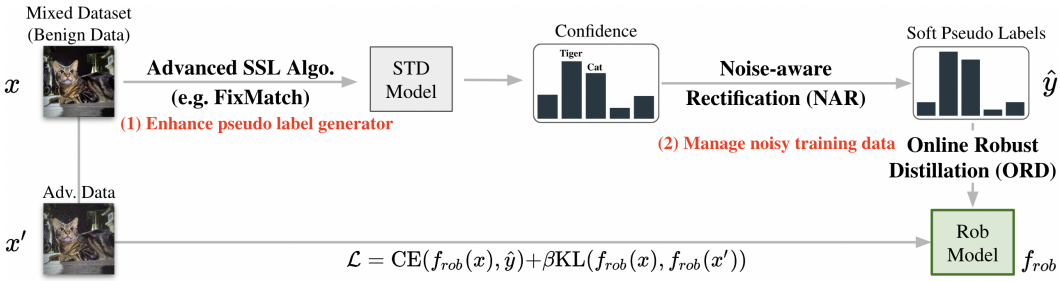

Figure 3: **Semi-supervised Noise-aware Online Robust Distillation (SNORD).** Our SNORD framework aims to address two important yet previously overlooked issues associated with noisy pseudo labels in the field of semi-supervised adversarial training. Firstly, it employs advanced SSL algorithms to improve the quality of pseudo labels (Section 3.1). Secondly, we introduce Noise-aware Rectification (Section 3.2) and Online Robust Distillation (Section 3.3) to enhance the learning capabilities of the downstream adversarial robust model in the context of handling noisy estimated pseudo labels.

a pretrained model to derive the pseudo label generator f_{std} (Jiang et al., 2020). While these methods are able to achieve 85% pseudo label accuracy with just 1% annotations on CIFAR-10, their potential to produce high-fidelity labels under extremely small amounts of labels remains largely unexplored. Evident from the CIFAR-100 analyses in Figure 2 (b), as the ratio of inaccurate predicted pseudo labels grows from 30% to around 50%, the final standard accuracy and adversarial robustness plummet by around 30% and 20%, respectively. Notably, given that the widely-adopted model (ResNet-18) consistently demonstrates a classification error surpassing 20% on CIFAR-100 (Zagoruyko & Komodakis, 2016), this issue can become quite serious during the subsequent adversarial training phase, not to mention when dealing with even scarcer labeling resources or more complex classification task, such as the case of TinyImageNet-200 (Le & Yang, 2015).

To acquire more accurate pseudo labels across varied scenarios, we embrace advanced semi-supervised learning (SSL) algorithms like FixMatch (Sohn et al., 2020) and ReMixMatch (Berthelot et al., 2020). After establishing the SSL-trained pseudo label generator f_{std}, we feed each benign image x through f_{std} to yield an estimated label distribution noted as $p(x) = \mathrm{softmax}(f_{std}(x))$. These estimations serve as the foundation for training the downstream adversarial robust model f_{rob}, which we elaborate on in the subsequent section.

3.2 Noise-aware Rectification

As shown in Figure 2 (c), on unlabeled data, prior methods either directly leveraged the soft predicted distribution as targets without applying entropy minimization technique or training the robust model with hard labels while overlooking inaccurate guidance. To enhance the use of the estimated label distribution in unlabeled data for subsequent adversarial training, we introduce a novel approach to improve pseudo labels, termed Noise-Aware Rectification (NAR). This method focuses on three key areas: mitigating the effects of mismatched label distribution noise in adversarial training, addressing inherent noise in incorrect predicted pseudo labels, and integrating the label sharpening techniques commonly used in the field of standard semi-supervised learning. The NAR strategy is mathematically represented as follows:

$$\hat{y} = \begin{cases} \lambda \cdot p(x) + (1-\lambda) \cdot y_{GT}, & \text{if } x \in D_L \\ \lambda \cdot p(x) + (1-\lambda) \cdot y_{PL}, & \text{if } x \in D_U. \end{cases} \tag{2}$$

In this formula, D_L and D_U indicate labeled and unlabeled datasets, respectively. y_{GT} represents the one-hot ground truth labels, λ is a label sharpening factor that balances the predicted distribution and one-hot labels, and y_{PL} is a one-hot label vector sampled from the predicted probability distribution $p(x)$. The sampling process for each dimension i of y_{PL} in a training batch is defined as:

$$y_{PL}^{i} = \begin{cases} 1, & \text{if } i \text{ is sampled from } p(x), \\ 0, & \text{Otherwise.} \end{cases} \tag{3}$$

Our design choices stem from three main factors. Firstly, recent studies point out a drawback of using one-hot label assignments for adversarial training, as adversarial perturbations can introduce label noise by

distorting data semantics (Dong et al., 2022). To counter this, we implement a label fusion technique that merges one-hot labels y_{GT} with the estimated distribution $p(x)$, resulting in a smoother label distribution.

Secondly, drawing inspiration from the success of entropy minimization in standard SSL tasks (Berthelot et al., 2019), we apply a similar technique to refine the label distribution $p(x)$ for the unlabeled data. This approach considers the potential inaccuracies of initial pseudo-label generators and opts for sampling a one-hot label y_{PL} from the distribution $p(x)$ instead of using a traditional $\mathrm{argmax}(\cdot)$ function. This strategy not only improves the label distribution for the unlabeled subset but also accounts for noise.

Finally, our approach applies the same label rectification process to both labeled and unlabeled data, using a single shared parameter. This contrasts with previous methods (Jiang et al., 2020; Carmon et al., 2019) and reduces the need for hyperparameter tuning while ensuring a balanced treatment of both data subsets. A detailed discussion and comparison of our NAR strategy are presented in Section 4.3.

3.3 Online Robust Distillation

While the combination of an improved pseudo label generator and a sophisticated noise-aware rectification strategy lays the foundation for the second-stage adversarial training, this adversarial training method cannot demonstrate the full potential of our approach. To further elevate its robustness, we introduce the concept of consistency regularization, a prevailing SSL strategy, into the realm of adversarial training. Drawing inspiration from the well-established research (Tarvainen & Valpola, 2017; Zi et al., 2021), we would like to leverage the efficacy of techniques such as label smoothing and ensemble methods across different epochs to enhance standard accuracy and adversarial robustness. In this work, we introduce an innovative online distillation mechanism that trains the robust model while concurrently updating the pseudo-label generator. In this process, we compute the estimated probability distribution at epoch t:

$$p^t(x) = f^t_{std}(\alpha(x)) \tag{4}$$

where $\alpha(\cdot)$ denotes a weak augmentation function, strategically introduced to diversify labels and further fortify robustness. To ensure training stability, the labels of the robust student undergo updates following the training of the pseudo label generator for T epochs. This dynamic knowledge distillation process in an online fashion empowers the teacher to offer more diverse and reliable guidance to its robust student, invariably leading to superior outcomes.

With an enhanced pseudo label generator, advanced label rectification strategy, and online distillation mechanism, the overall loss function for our holistic SNORD framework is as follows:

$$\mathcal{L} = \mathrm{CE}(f_{rob}(x), \hat{y}) + \beta \cdot \mathrm{KL}(f_{rob}(x), f_{rob}(x')) \tag{5}$$

where $\hat{y}$ is computed with equations 2, 3, and 4.

4 Experiments

4.1 Experimental Settings

We describe the datasets and experimental protocols in this section. Implementation details including hyper-parameter settings are reported in Appendix A.

Datasets. Our experiments were conducted with the widely-used CIFAR-10/100 datasets (Krizhevsky et al., 2009) and the TinyImageNet-200 dataset (Le & Yang, 2015). Following the established protocols from previous SSL adversarial training studies (Luo et al., 2023; Jiang et al., 2020), for CIFAR-10 and CIFAR-100, we evaluated our models using the official test set, while the official training set was divided into a 9:1 ratio for training and validation. For TinyImageNet-200, we directly used the official training, validation, and test set in the experiment. To create SSL settings, we randomly partitioned the training set into labeled and unlabeled portions to meet our specific experimental requirements following (Luo et al., 2023; Jiang et al., 2020). To mitigate potential class-imbalanced issues during training, we made sure to distribute the images of each class evenly across the splits. In cases where the number of images couldn't be divided exactly, a small difference of at most one image was allowed.

Training and Evaluation Protocols. For fair comparisons, we employed the widely-used ResNet-18 (He et al., 2016) model architecture for all experiments. During training, all methods were allowed to

Methods	0.1% labels			0.2% labels			0.5% labels			1% labels			5% labels			10% labels		
	SA	RA	AA	SA	RA	AA	SA	RA	AA	SA	RA	AA	SA	RA	AA	SA	RA	AA
UAT++	24.87	13.28	11.25	31.79	19.88	17.93	41.47	25.19	22.29	49.75	30.74	27.16	70.41	45.83	42.69	76.34	49.57	46.11
ACL	48.02	35.03	31.04	68.38	46.43	41.89	71.99	50.20	45.26	75.45	50.59	46.18	77.77	51.21	47.11	76.37	51.73	47.14
DynACL++	64.34	47.31	44.27	70.34	50.14	44.46	69.92	51.03	45.49	76.77	51.30	46.95	79.07	51.35	48.01	78.34	**53.00**	48.56
SNORD (Ours)	**71.71**	**50.07**	**47.12**	**76.83**	**53.28**	**49.74**	**80.99**	**53.46**	**49.48**	**80.60**	**53.22**	**49.47**	**81.96**	**52.90**	**49.29**	**82.73**	52.97	**49.14**

Table 1: **CIFAR-10 test accuracy (%) under SSL settings.** SA, RA, and AA denote standard accuracy, PGD-20 robust accuracy, and AutoAttack robust accuracy, respectively. For baselines, we report the maximum value of their official number and our reproduced results.

Methods	1% labels			2% labels			5% labels			10% labels		
	SA	RA	AA	SA	RA	AA	SA	RA	AA	SA	RA	AA
UAT++	10.43	4.14	3.59	17.48	6.52	5.66	25.12	12.13	10.41	38.63	18.86	16.01
ACL	27.87	19.39	16.51	32.10	19.83	17.70	42.57	25.64	21.90	44.05	26.78	22.33
DynACL++	35.34	21.55	18.43	32.92	21.33	18.48	42.81	25.93	21.89	45.64	27.98	23.79
SNORD (Ours)	**35.44**	**22.08**	**19.00**	**44.14**	**26.14**	**22.46**	**48.09**	**27.05**	**23.42**	**52.03**	**28.27**	**23.94**

Table 2: **CIFAR-100 test accuracy (%) under SSL settings.** Our method showed huge gains over all existing baseline methods across diverse labeling budgets even without the use of pre-trained models.

Methods	10% labels			20% labels		
	SA	RA	AA	SA	RA	AA
UAT++	31.96	12.98	9.26	33.36	13.30	9.88
SNORD (Ours)	**41.70**	**20.02**	**15.26**	**46.84**	**22.00**	**16.70**

Table 3: **TinyImageNet-200 test accuracy (%) under SSL settings.** We only compare our method to the basic RST method (UAT++), given that previous adversarial pretraining approaches had refrained from such a large-scale dataset.

optimize the model while considering a maximum perturbation of $L_\infty = 8/255$. The maximum number of perturbation steps was limited to 10 and the step size of each perturbation is set to $\alpha = 2/255$.

For our evaluation, we adhered to the protocols established in prior works (Luo et al., 2023; Jiang et al., 2020), utilizing three standard metrics for a thorough assessment: standard accuracy (SA), PGD-20 robust accuracy (RA) (Madry et al., 2018), and AutoAttack (AA) accuracy (Croce & Hein, 2020). PGD-20 refers to the robustness of the model when challenged with the Projected Gradient Descent method, a widely recognized adversarial attack approach, using 20 iterations. AutoAttack (AA), on the other hand, is an ensemble of diverse attack methods designed to provide a more comprehensive and stringent test of model robustness. In both metrics, the maximum perturbation was bounded by $L_\infty = 8/255$ across all three datasets: CIFAR-10, CIFAR-100, and TinyImageNet-200. To gain a comprehensive understanding of the problem, we evaluated all SSL adversarial training methods across a range of labeling budgets. Specifically, the labeling budgets for CIFAR-10, CIFAR-100, and TinyImageNet-200 varied from $0.1\% \sim 10\%$, $1\% \sim 10\%$, and $10\% \sim 20\%$, respectively.

4.2 Main Results

We compared our method with three semi-supervised adversarial training baselines. Among them, UAT++ (Alayrac et al., 2019) is the basic RST method, while ACL (Jiang et al., 2020) and DynACL++ (Luo et al., 2023) involve a self-supervised adversarial pretraining stage followed by a few RST-based finetuning steps. As UAT++ did not release the source code, we re-implemented the method and reported the number using our code. For ACL and DynACL++, we leveraged their provided pretrained model and source code to obtain the result. We reported the maximum value of their official number and our finetuned results.

CIFAR-10. The results in Table 1 underscore the remarkable efficacy of our SNORD framework over existing baselines across a spectrum of labeling budgets. In the context of scarce labeling resources ($<5\%$), a regime in which conventional methods falter in generating high-quality pseudo labels, SNORD exhibits an impressive advancement of around 3% in AA when compared to all established baselines. This highlights

the advantages inherent in our approach of leveraging sophisticated SSL algorithms to derive pseudo labels, as opposed to relying solely on pretrained models within standard training paradigms.

On the other hand, even when at least 1% labeled data is available—enough for baseline methods to generate satisfactory pseudo labels—SNORD not only slightly outperforms these baselines in terms of AA, but also provides a substantial improvement in SA. The huge gain actually comes from the benefits of our NAR and ORD modules, which collectively facilitate the effective management of noisy training data stemming from inaccurate pseudo label while also harnessing robust entropy minimization techniques to bolster standard accuracy.

Notably, even with a mere 0.2% labeled data (equivalent to only 9 labeled images per class), SNORD achieves an incredible 49% AA robustness. This impressive achievement underscores the superior capabilities of our method in harnessing minimal labeled data to rival the outcomes of extensively supervised methods such as TRADES (Zhang et al., 2019) and TRADES-AWP (Wu et al., 2020) shown in Figure 1.

CIFAR-100. As presented in Table 2, as the CIFAR-100 is much more difficult than CIFAR-10, the performance of the baseline vanilla RST method (UAT++) is notably bad, failing to exceed 20% AA with even with 10% available annotated data. In such a complicated task, while the integration of pretrained models indeed provides improvement compared with UAT++, our SNORD framework continues to outperform these approaches across a diverse range of labeling budgets. Echoing the results seen in CIFAR-10, SNORD consistently yields significantly improved standard accuracy (SA) over baseline methods, particularly in scenarios with ample labeling resources.

It is important to note that in scenarios where only 1% of labeled data is available (equivalent to only 4 to 5 labeled images per class), the prior state-of-the-art DynACL++ method manages to achieve results that closely approximate those of SNORD. This situation can be attributed to the relatively high proportion of inaccurate pseudo labels generated by SNORD, thus obtaining only compromised results compared with the performance under 2% labels.

TinyImageNet-200. As depicted in Table 3, our SNORD method consistently shows exceptional performance, even when faced with a more challenging dataset, as compared directly to the basic RST method (UAT++). It is worth highlighting that the utilization of the supervised TRADES method yields SA and AA scores of only 48.49 and 17.35, respectively. In stark contrast, SNORD achieves results that outperform 95% of it with merely 20% labeled data.

Comparison with Full Adversarial Training. We also compare the proposed SNORD pipeline with the state-of-the-art fully adversarial training method, ADR (Wu et al., 2024), which utilizes the entire set of labeled data. As depicted in Table 4, SNORD, even with 10 times fewer labels, achieves impressive results in both Standard Accuracy (SA) and AutoAttack (AA) across different datasets. Most notably, SNORD substantially narrows the AA performance gap to less than 5% when compared to the leading full adversarial training method. Furthermore, SNORD significantly surpasses existing semi-supervised adversarial training methods, particularly with the large-scale TinyImageNet-200 dataset and in the low-label regimes of CIFAR-10/100. These findings underscore SNORD's remarkable efficacy and contribution in the realm of SSL adversarial training.

Method (labels)	CIFAR-10		CIFAR-100		TinyImageNet	
	SA	AA	SA	AA	SA	AA
ADR (100%)	82.41	50.38	56.10	26.87	48.19	19.46
Prior SOTA (0.2%)	70.34	44.46	-	-	-	-
SNORD (0.2%)	**76.83**	**49.74**	-	-	-	-
Prior SOTA (5%)	79.07	48.01	42.81	21.89	-	-
SNORD (5%)	**81.96**	**49.29**	**48.09**	**23.42**	-	-
Prior SOTA (10%)	78.34	48.56	45.64	23.79	31.96	9.26
SNORD (10%)	**82.73**	**49.14**	**52.03**	**23.94**	**41.70**	**15.26**

Table 4: **More Comparisons.** We conduct a further comparison of SNORD against previous state-of-the-art SSL adversarial training methods and the fully adversarial supervision method, ADR (Wu et al., 2024), as well. Our findings demonstrate that SNORD not only outperforms all existing SSL adversarial training approaches across various settings but also significantly closes the gap in AutoAttack robustness (AA).

4.3 Discussions

Ablation Studies. We conducted comprehensive ablation studies to assess the efficacy of our proposed components on both the CIFAR-10 and CIFAR-100 datasets, as presented in Table 5. The results exhibit clear improvements when transitioning from standard training (row (a)) to advanced semi-supervised learning (SSL) algorithms (row (b)) under limited labeled data. The introduction of our NAR and ORD modules, individually illustrated in rows (c) and (d), yields evident enhancements in both performance and robustness over the baseline (row (b)). Notably, the combination of both NAR and ORD modules in row (e) showcases the most favorable outcomes in two prominent robustness evaluation benchmarks. However, a marginal performance drop in standard accuracy is observed in row (e) in comparison to applying NAR or ORD individually (rows (c) and (d)). This minor decline could be attributed to a cumulative effect of smoothing operations stemming from the combined deployment of NAR and ORD modules, aligning with insights from prior research (Zi et al., 2021). Furthermore, we evaluate the efficacy of our NAR and ORD modules in conjunction with established RST-based methods like ACL and DynACL++. By intentionally excluding advanced SSL algorithms in favor of improved pseudo labels, our approach consistently yields superior results in Table 6, firmly underscoring the value of NAR and ORD.

Additionally, we delve deeper into the mechanics of our proposed NAR method. Illustrated in Figure 4, the hyper-parameter λ, responsible for harmonizing one-hot labels with predicted distributions, manifests consistent performance across a substantial range of values ($0.25 \sim 0.5$) on both datasets, affirming the robustness of our method. This observation further confirms the limitations inherent in the adversarial finetuning strategy employed by ACL and DynACL++, as the classifier's proficiency leans more toward one-hot labels than sole reliance on soft predicted distributions. Moreover, the comparison between our sampling process in equation (3), and the use of $\mathrm{argmax}(\cdot)$ function to obtain y_{PL} (indicated by the X-mark in the figures) clearly shows the superiority of our sampling approach. This highlights the importance of employing sampling to enhance label diversity and mitigate noise inherent in noisy pseudo labels.

Applicability to Adversarial Pretraining Methods. We extend our evaluation to assess the broader applicability of our approach by integrating it with existing adversarial pretraining methods. In this context, we initialize the robust model with adversarially pretrained weights from DynACL++ and subsequently perform adversarial finetuning with the NAR and ORD modules from an SSL-trained pseudo label generator over 30 epochs. As demonstrated in Table 7, when pseudo label quality is high, such as achieving over 85% precision with 1% CIFAR-10 labeled data, the application of adversarial pretraining on top of our SNORD method yields the best results, consistent with findings in (Luo et al., 2023). However, when the initial pseudo label quality is lower, as indicated by over 40% error with 5% CIFAR-100 annotations,

	Components			CIFAR-10 1%			CIFAR-100 5%		
	SSL algo.	NAR	ORD	SA	RA	AA	SA	RA	AA
(a)				55.58	34.52	31.68	33.90	17.19	14.47
(b)	✓			79.91	52.06	48.72	48.57	26.44	22.78
(c)	✓	✓		**81.81**	52.50	49.09	**49.84**	26.92	23.13
(d)	✓		✓	81.55	52.07	48.34	48.52	26.85	23.34
(e)	✓	✓	✓	80.60	**53.22**	**49.47**	48.29	**27.05**	**23.42**

Table 5: **Ablation studies.** By harnessing the capabilities of our developed NAR and ORD modules, in conjunction with an SSL-trained pseudo label generator, we are able to achieve the optimal results for both RA and AA.

Methods	CIFAR-10 1%			CIFAR-100 5%		
	SA	RA	AA	SA	RA	AA
ACL	75.45	50.59	46.18	42.57	25.64	**21.90**
+NAR+ORD	**81.21**	**52.50**	**48.16**	**50.67**	**26.22**	21.83
DynACL++	76.77	51.30	46.95	42.81	25.93	21.89
+NAR+ORD	**80.54**	**52.30**	**48.45**	**48.92**	**26.58**	**22.20**

Table 6: **Effectiveness of NAR and ORD.** Without the use of SSL-trained pseudo label generator, we showcase the benefit of NAR and ORD on top of prior RST-based pipelines.

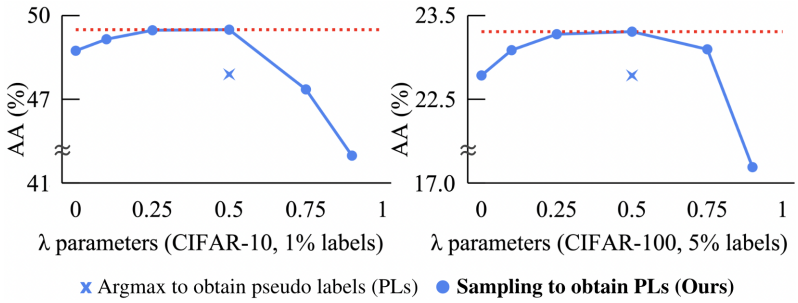

Figure 4: **Sensitivity analyses of our NAR methods.** The result underscores the importance of using sampling rather than argmax to obtain pseudo labels and the robustness of our approach over a wide range of hyper-parameters.

Methods	CIFAR-10 1%			CIFAR-100 5%		
	SA	RA	AA	SA	RA	AA
DynACL++	76.77	51.30	46.95	42.81	25.93	21.89
SNORD	80.60	53.22	49.47	**48.09**	27.05	**23.42**
DynACL++ & SNORD	**82.17**	**53.78**	**50.37**	48.00	**27.29**	23.19

Table 7: **Combination of SNORD with adversarial pretraining methods.** With high-quality pseudo labels on CIFAR-10, the combined utilization of a robust pretrained model and our SNORD mechanism yields enhanced outcomes in comparison to employing either one in isolation. Nonetheless, when confronted with severe noisy training data on CIFAR-100, the potential effectiveness of the pretrained model becomes constrained, underscoring the substantial significance of SNORD.

the effectiveness of initializing the network with robust pretrained weights is negligible compared to direct adversarial training from scratch using our SNORD framework. This highlights that in the realm of SSL adversarial training, the impact of robust pretrained models may be constrained when finetuning with noisy data. Instead, the absence of a robust pseudo label generator or an effective method to handle noisy training data can lead to a remarkable performance drop, as evident in the comparison between the first and second rows of the table.

5 Related Work

5.1 Semi-supervised Learning (SSL)

SSL has become increasingly popular due to its capacity for leveraging abundant unlabeled data to enhance model performance. Within the realm of SSL, existing strategies can be broadly categorized into two primary domains: entropy minimization and consistency regularization.

In the context of entropy minimization, methods often assume that a classifier's decision boundary should steer clear of high-density regions in the data distribution. To enact this principle, (Lee et al., 2013) introduced the "pseudo-labeling" technique, a straightforward yet remarkably effective approach utilizing one-hot encoding. Recent advancements (Berthelot et al., 2019; Miyato et al., 2018) have further refined this approach through label sharpening, enhancing label distributions via softer labels.

Conversely, consistency regularization leverages data augmentation to reinforce the SSL process, ensuring a classifier's output class distribution remains consistent for unlabeled instances even post-augmentation. Several methods and loss functions (Rasmus et al., 2015; Sajjadi et al., 2016; Berthelot et al., 2020; Xie et al., 2020; Tarvainen & Valpola, 2017) have been proposed to realize this idea in diverse ways.

While recent studies have achieved remarkable success by combining these strategies for standard image classification, such as FixMatch (Sohn et al., 2020), their application to the domain of SSL adversarial training remains limited. In this work, we introduce a pioneering unified framework that integrates both

entropy minimization and consistency regularization into the field of adversarial robustness. This framework not only significantly surpasses existing baselines but also offers a novel perspective to the field.

5.2 Adversarial Robustness

Research in adversarial robustness can be broadly categorized into two fronts: attacks and defenses. Adversarial attacks aim to craft adversarial samples that are misclassified by models through introducing minimal perturbations to benign data, while defensive approaches seek to enhance model robustness against such attacks. Over the past few years, numerous classical attacks, such as Fast Gradient Sign Method (FGSM) (Goodfellow et al., 2015) Projected Gradient Descent (PGD) (Madry et al., 2018), and AutoAttack (AA) (Croce & Hein, 2020), have generated adversarial examples by back-propagating loss functions. On the other hand, defensive methods have employed techniques like obfuscated gradients (Athalye et al., 2018) or various adversarial training strategies (Madry et al., 2018; Zhang et al., 2019; 2021; Wu et al., 2024).

Recent studies have delved into the impact of noisy training data on adversarial training. (Chen et al., 2021; Dong et al., 2022; Zhu et al., 2021) mitigated mismatched distribution noise between benign and adversarial samples by applying knowledge distillation loss to smoothed label distributions. In contrast, (Zhang et al., 2022b) introduced a noise injection mechanism to counter robust overfitting. Unlike these endeavors, which focus on noisy labels within supervised learning contexts, our novel noise-aware label rectification strategy addresses inaccurate pseudo label noise specific to the SSL adversarial training paradigm. This addresses a distinct challenge and contributes to enhanced robustness in scenarios with limited labeled data.

5.3 Semi-supervised Adversarial Training

Schmidt et al. (2018) demonstrated that increasing the amount of labeled data can bolster the adversarial robustness of models. This insight led to the emergence of research on achieving robustness with limited labeled data, termed semi-supervised adversarial robustness. Early studies (Carmon et al., 2019; Alayrac et al., 2019; Zhai et al., 2019; Li et al., 2022) introduced the two-stage Robust Self-Training (RST) pipeline, involving pseudo label generation from a standard-trained model in the initial stage and adversarial training on the entire dataset in the subsequent stage.

Despite the simplicity of RST, it often experiences significant performance degradation when labeled data is scarce (e.g.,$<$10% on CIFAR-10) (Gowal et al., 2021). To address this, several researchers have developed self-supervised adversarial training strategies to obtain pretrained models with robust feature representations (Naseer et al., 2020; Hendrycks et al., 2019; Chen et al., 2020a;b; Gowal et al., 2021; Luo et al., 2023; Zhang et al., 2022a; Fan et al., 2021; Kim et al., 2020; Ho & Nvasconcelos, 2020). These techniques have achieved over 85% robustness compared to fully supervised methods on relatively simple datasets like CIFAR-10, using only 1% to 10% annotations of the entire training set.

Building upon this line of research, we observe that existing RST-based methods struggle with complex tasks or extremely limited labeling scenarios due to the low-quality pseudo labels in the initial stage and the inability to handle noisy training data in the subsequent stages. To address these challenges, we propose a novel approach that seamlessly integrates contemporary SSL techniques into the realm of adversarial training. Our proposed framework not only achieves state-of-the-art across established benchmarks but is also applicable to existing adversarial pretrained models, signifying a significant advancement in the domain of semi-supervised adversarial robustness.

6 Conclusion

We present SNORD, a simple, effective, and general semi-supervised adversarial training framework. Instead of focusing on adversarial pertained models as a lot of prior work, we revised the widely-used RST-based methods and pointed out that the bottleneck to this problem is the quality of pseudo-labels and the management of noisy training data. With the aid of existing and our developed SSL techniques, SNORD demonstrates a substantial performance advantage over conventional RST-based approaches across multiple well-established benchmarks, regardless of the usage of adversarial pretrained models. This success paves the way for a novel approach to label-efficient and adversarial robust machine learning systems.

ACKNOWLEDGEMENT

This work was supported in part by the National Science and Technology Council under Grants NSTC 113-2634-F-002-007, 113-2222-E-002-004-MY3, 113-2923-E-002-010-MY2, 113-2634-F-002-001-MBK, and by the Center of Data Intelligence: Technologies, Applications, and Systems, National Taiwan University under Grant NTU-113L900903.

REFERENCES

Jean-Baptiste Alayrac, Jonathan Uesato, Po-Sen Huang, Alhussein Fawzi, Robert Stanforth, and Pushmeet Kohli. Are labels required for improving adversarial robustness? *Advances in Neural Information Processing Systems*, 32, 2019. 1, 3, 6, 10

Anish Athalye, Nicholas Carlini, and David Wagner. Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples. In *International conference on machine learning*, pp. 274–283. PMLR, 2018. 10

David Berthelot, Nicholas Carlini, Ian Goodfellow, Nicolas Papernot, Avital Oliver, and Colin A Raffel. Mixmatch: A holistic approach to semi-supervised learning. *Advances in neural information processing systems*, 32, 2019. 5, 9

David Berthelot, Nicholas Carlini, Ekin D. Cubuk, Alex Kurakin, Kihyuk Sohn, Han Zhang, and Colin Raffel. Remixmatch: Semi-supervised learning with distribution matching and augmentation anchoring. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=HklkeR4KPB. 4, 9

Yair Carmon, Aditi Raghunathan, Ludwig Schmidt, John C Duchi, and Percy S Liang. Unlabeled data improves adversarial robustness. *Advances in neural information processing systems*, 32, 2019. 1, 2, 3, 5, 10

Kejiang Chen, Yuefeng Chen, Hang Zhou, Xiaofeng Mao, Yuhong Li, Yuan He, Hui Xue, Weiming Zhang, and Nenghai Yu. Self-supervised adversarial training. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 2218–2222. IEEE, 2020a. 10

Tianlong Chen, Sijia Liu, Shiyu Chang, Yu Cheng, Lisa Amini, and Zhangyang Wang. Adversarial robustness: From self-supervised pre-training to fine-tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 699–708, 2020b. 10

Tianlong Chen, Zhenyu Zhang, Sijia Liu, Shiyu Chang, and Zhangyang Wang. Robust overfitting may be mitigated by properly learned smoothening. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=qZzy5urZw9. 10

Francesco Croce and Matthias Hein. Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks. In *International conference on machine learning*, pp. 2206–2216. PMLR, 2020. 1, 6, 10

Chengyu Dong, Liyuan Liu, and Jingbo Shang. Label noise in adversarial training: A novel perspective to study robust overfitting. *Advances in Neural Information Processing Systems*, 35:17556–17567, 2022. 5, 10

Lijie Fan, Sijia Liu, Pin-Yu Chen, Gaoyuan Zhang, and Chuang Gan. When does contrastive learning preserve adversarial robustness from pretraining to finetuning? In A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, 2021. URL https://openreview.net/forum?id=70kOIgjKhbA. 10

Ian J. Goodfellow, Jonathon Shlens, Christian Szegedy, et al. Explaining and harnessing adversarial examples. In Yoshua Bengio and Yann LeCun (eds.), *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015. URL http://arxiv.org/abs/1412.6572. 1, 10

Sven Gowal, Po-Sen Huang, Aaron van den Oord, Timothy Mann, and Pushmeet Kohli. Self-supervised adversarial robustness for the low-label, high-data regime. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=bgQek2O63w. 3, 10

Kaiming He, Xiangyu Zhang, Shaoqing Ren, Jian Sun, et al. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016. 5, 14

Dan Hendrycks, Mantas Mazeika, Saurav Kadavath, and Dawn Song. Using self-supervised learning can improve model robustness and uncertainty. *Advances in neural information processing systems*, 32, 2019. 10

Chih-Hui Ho and Nuno Nvasconcelos. Contrastive learning with adversarial examples. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 17081–17093. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/c68c9c8258ea7d85472dd6fd0015f047-Paper.pdf. 10

Ziyu Jiang, Tianlong Chen, Ting Chen, and Zhangyang Wang. Robust pre-training by adversarial contrastive learning. 2020. 1, 2, 3, 4, 5, 6, 14, 15

Minseon Kim, Jihoon Tack, and Sung Ju Hwang. Adversarial self-supervised contrastive learning. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 2983–2994. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/1f1baa5b8edac74eb4eaa329f14a0361-Paper.pdf. 10

Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009. 5

Ya Le and Xuan Yang. Tiny imagenet visual recognition challenge. *CS 231N*, 7(7):3, 2015. 4, 5

Dong-Hyun Lee et al. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In *Workshop on challenges in representation learning, ICML*, volume 3, pp. 896. Atlanta, 2013. 2, 9

Yiming Li, Baoyuan Wu, Yan Feng, Yanbo Fan, Yong Jiang, Zhifeng Li, and Shu-Tao Xia. Semi-supervised robust training with generalized perturbed neighborhood. *Pattern Recognition*, 124:108472, 2022. 10

Rundong Luo, Yifei Wang, and Yisen Wang. Rethinking the effect of data augmentation in adversarial contrastive learning. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=0qmwFNJyxCL. 1, 2, 3, 5, 6, 8, 10, 14, 15

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=rJzIBfZAb. 1, 6, 10

Takeru Miyato, Shin-ichi Maeda, Masanori Koyama, and Shin Ishii. Virtual adversarial training: a regularization method for supervised and semi-supervised learning. *IEEE transactions on pattern analysis and machine intelligence*, 41(8):1979–1993, 2018. 2, 9

Muzammal Naseer, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, and Fatih Porikli. A self-supervised approach for adversarial robustness. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 10

Tianyu Pang, Xiao Yang, Yinpeng Dong, Hang Su, and Jun Zhu. Bag of tricks for adversarial training. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=Xb8xvrtB8Ce. 14

Antti Rasmus, Mathias Berglund, Mikko Honkala, Harri Valpola, and Tapani Raiko. Semi-supervised learning with ladder networks. *Advances in neural information processing systems*, 28, 2015. 9

Leslie Rice, Eric Wong, and Zico Kolter. Overfitting in adversarially robust deep learning. In *International Conference on Machine Learning*, pp. 8093–8104. PMLR, 2020. 14, 15

Andras Rozsa, Manuel Günther, Ethan M Rudd, and Terrance E Boult. Facial attributes: Accuracy and adversarial robustness. *Pattern Recognition Letters*, 124:100–108, 2019. 1

Mehdi Sajjadi, Mehran Javanmardi, and Tolga Tasdizen. Regularization with stochastic transformations and perturbations for deep semi-supervised learning. *Advances in neural information processing systems*, 29, 2016. 9

Ludwig Schmidt, Shibani Santurkar, Dimitris Tsipras, Kunal Talwar, and Aleksander Madry. Adversarially robust generalization requires more data. *Advances in neural information processing systems*, 31, 2018. 1, 10

Kihyuk Sohn, David Berthelot, Nicholas Carlini, Zizhao Zhang, Han Zhang, Colin A Raffel, Ekin Dogus Cubuk, Alexey Kurakin, and Chun-Liang Li. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. *Advances in neural information processing systems*, 33:596–608, 2020. 2, 4, 9, 14

Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. *Advances in neural information processing systems*, 30, 2017. 5, 9

Dongxian Wu, Shu-Tao Xia, and Yisen Wang. Adversarial weight perturbation helps robust generalization. *Advances in neural information processing systems*, 33:2958–2969, 2020. 1, 7

Yu-Yu Wu, Hung-Jui Wang, and Shang-Tse Chen. Annealing self-distillation rectification improves adversarial training. In *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=eT6oLkm1cm. 7, 10, 14

Qizhe Xie, Zihang Dai, Eduard Hovy, Thang Luong, and Quoc Le. Unsupervised data augmentation for consistency training. *Advances in neural information processing systems*, 33:6256–6268, 2020. 9

Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. In *British Machine Vision Conference (BMVC)*, 2016. 4

Runtian Zhai, Tianle Cai, Di He, Chen Dan, Kun He, John Hopcroft, and Liwei Wang. Adversarially robust generalization just requires more unlabeled data. *arXiv preprint arXiv:1906.00555*, 2019. 1, 3, 10

Chaoning Zhang, Kang Zhang, Chenshuang Zhang, Axi Niu, Jiu Feng, Chang D. Yoo, and In So Kweon. Decoupled adversarial contrastive learning for self-supervised adversarial robustness. In Shai Avidan, Gabriel Brostow, Moustapha Cissé, Giovanni Maria Farinella, and Tal Hassner (eds.), *Computer Vision – ECCV 2022*, pp. 725–742, Cham, 2022a. Springer Nature Switzerland. 10

Hongyang Zhang, Yaodong Yu, Jiantao Jiao, Eric Xing, Laurent El Ghaoui, and Michael Jordan. Theoretically principled trade-off between robustness and accuracy. In *International conference on machine learning*, pp. 7472–7482. PMLR, 2019. 1, 7, 10, 15

Jingfeng Zhang, Jianing Zhu, Gang Niu, Bo Han, Masashi Sugiyama, and Mohan Kankanhalli. Geometry-aware instance-reweighted adversarial training. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=iAX016Cz8ub. 1, 10

Jingfeng Zhang, Xilie Xu, Bo Han, Tongliang Liu, Lizhen Cui, Gang Niu, and Masashi Sugiyama. Noilin: Improving adversarial training and correcting stereotype of noisy labels. *Transactions on Machine Learning Research*, 2022b. ISSN 2835-8856. URL https://openreview.net/forum?id=zlQXV7xtZs. 10

Jianing Zhu, Jingfeng Zhang, Bo Han, Tongliang Liu, Gang Niu, Hongxia Yang, Mohan Kankanhalli, and Masashi Sugiyama. Understanding the interaction of adversarial training with noisy labels. *arXiv preprint arXiv:2102.03482*, 2021. 10

Bojia Zi, Shihao Zhao, Xingjun Ma, and Yu-Gang Jiang. 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. 5, 8

APPENDIX

A IMPLEMENTATION DETAILS

In this section, we present the specifics of our implementation, covering the computing infrastructure used and the training process details.

A.1 COMPUTING INFRASTRUCTURE

All experiments were conducted on a personal computer with an 8-core CPU and an NVIDIA RTX3090 GPU. The operating system was Ubuntu 20.04, and we implemented the system using Python 3.7. For efficient computations, we leveraged CUDA 11.3 and employed PyTorch 1.11.0 as our deep learning framework.

A.2 TRAINING DETAILS

Our SNORD approach involves training a pseudo label generator through semi-supervised learning (SSL) and a robust model adversarial trained by our noise-aware manner. We trained on CIFAR-10/100 images of size 32×32 and on TinyImageNet-200 images of size 64×64. The data generation method is detailed in Section 4.1 of the main manuscript, where we adopt the data split from (Rice et al., 2020) for CIFAR-10/100 and rely on the official split for TinyImageNet-200. We constructed class-balanced SSL datasets following established methods such as (Sohn et al., 2020; Jiang et al., 2020; Luo et al., 2023).

Training Pseudo Label Generator. In our experiments, the SSL-trained model was trained using the FixMatch algorithm (Sohn et al., 2020) with a ResNet-18 (He et al., 2016) backbone. The hyperparameters for training our pseudo label generator are detailed in Table A.1. In this table, τ represents the confidence threshold, and μ is the ratio of unlabeled data to labeled data. The training incorporated the SGD optimizer with an initial learning rate of 0.03, momentum of 0.9, and Nesterov momentum enabled, consistent for the three datasets. The weight decay for CIFAR-10, CIFAR-100, and TinyImageNet-200 were set to 1e-3, 2e-4, and 1e-4, respectively. We employed the original cosine annealing scheduler for adaptive learning rate adjustment. Other parameters and data augmentation methods remained consistent with those in the original paper.

	CIFAR-10	CIFAR-100	TinyImageNet-200
τ		0.95	
μ		5	
B		64	
lr		0.03	
weight decay	0.001	0.0002	0.0001

Table A.1: Hyperparameters for training the pseudo label generator on the three datasets.

Training Robust Model. After detailing the pseudo label generator's training, we proceed to explain the adversarial training of the robust model. The robust model's training commenced at the 128th and 256th epochs of the pseudo label generator on CIFAR and TinyImageNet datasets, respectively, as the standard model's performance was satisfactory.

The robust model was trained for 200 epochs for the CIFAR-10/100 and 80 epochs for the TinyImageNet-200 following existing work (Pang et al., 2021; Wu et al., 2024). We optimize the model with the SGD optimizer, initialized with a learning rate of 0.1 and a weight decay of 5e-4, in line with prior studies (Pang et al., 2021). The learning rate was managed through a piecewise learning rate scheduler, with a reduction factor of 0.1 at the 50% and 75% epochs of the total robust model training duration (Rice et al., 2020). Data augmentation included basic techniques such as random cropping with padding and random horizontal flipping, consistent with prior studies (Pang et al., 2021; Rice et al., 2020; Jiang et al., 2020).

Notably, we set the noise-aware label rectification parameter λ in equation (2) of the main paper to 0.5. The parameter's sensitivity was evaluated in Figure 4 of the main manuscript. We set the β parameter in

equation (5) of the main paper to 6, following TRADES (Zhang et al., 2019). Additionally, we applied the proposed ORD method from the standard model to the robust model every 5 epochs, indicating that we update the robust model once after 5 standard model updates. This strategy was chosen as we observed that allowing the standard model to update for multiple epochs enhanced label diversity and overall performance.

Ensuring Fair Comparison. To ensure consistency, each experimental setting, including all the baseline methods, was executed once with the same random seed, as the observed variation on two major robust evaluation metrics was minimal as shown in Table A.2. Our reported results, following well-established protocols (Rice et al., 2020), are derived from the best checkpoint, selected based on the highest Robustness Accuracy (RA) on the validation set.

Methods	CIFAR-10 5%			CIFAR-100 5%		
	SA	RA	AA	SA	RA	AA
Run 1	81.66	52.91	49.25	48.29	27.25	23.30
Run 2	81.27	52.90	49.29	48.09	27.05	23.42

Table A.2: Performance variation with different random seeds.

Note that to accelerate the whole training process, we harnessed automatic mixed precision (AMP) training. Additionally, we implemented early stopping as deemed necessary within our approach. Conversely, for other methods, we utilized their original implementations with 32-bit full precision training. This approach was implemented to efficiently manage computational resources while ensuring fair comparison across methods.

The overall training time of our SNORD method, including both the pseudo label generator and the robust model, for CIFAR-10/100 amounted to approximately 32 hours. This timeframe is quite comparable to the requirements of our baseline methods ACL (Jiang et al., 2020) and DynACL++ (Luo et al., 2023), which involve around 30 hours for adversarial pretraining and 2 to 3 hours for adversarial fine-tuning. As for the TinyImageNet-200 dataset, due to its greater complexity, the training time of SNORD extended to approximately 3.5 days to effectively train the model.