Appendix

Table of Contents

A. Detailed Background and Related Works

A.1. Adversarial Robustness

Adversarial robustness indicates the vulnerability of DNN classifiers under intended perturbations by an adversary, which is commonly measured with the test accuracy under adversarial attacks [3]. Since Szegedy et al. [50] introduced the concept of adversarial attack, many effective attack methods have been proposed. For instance, the Fast Gradient Sign Method (FGSM) [16] implements the perturbation according to the gradient of the loss function *w.r.t.* the input sample. The Projected Gradient Descent (PGD) attack [36] can be viewed as a variant of FGSM. It generates the perturbation by FGSM iteratively and then projects it to the ϵ-ball of the input sample. The C&W attack [5] no longer relies on ϵ-ball as a constraint of perturbation radius, but formulates a regularization term leading to small perturbation instead. Auto-Attack (AA) [9] forms a parameter-free and user-independent ensemble of attacks for frequent pitfalls in practice like improper tuning of hyper-parameters and gradient obfuscation or masking. Due to their effectiveness, PGD, C&W and AA are commonly used to evaluate the adversarial robustness of DNNs.

With the development of adversarial attack, there are also a number of defense methods proposed to improve the adversarial robustness of DNNs, including Defense Distillation [39], Feature Squeezing [65], Input Denoising [18, 32] and Randomization [64]. However, most of these methods have been proven subsequently to rely on obfuscated gradients [2] and be ineffective against advanced adaptive attacks [54].

A.2. Adversarial Training

Currently, Adversarial Training [16, 36] is widely recognized as the most effective and practical method to acquire adversarially robust DNNs [2, 12]. Different from clean training, in the context of AT, the model is directly trained on adversarially augmented samples instead of only natural ones. Specifically, the optimization objective of AT can be defined as a min-max problem [36]:

$$\min_{\boldsymbol{\theta}} \frac{1}{n} \sum_{i=1}^{n} \max_{\|\mathbf{x}_i' - \mathbf{x}_i\|_p \leq \epsilon} \mathcal{L}(f_{\boldsymbol{\theta}}(\mathbf{x}_i'), y_i), \tag{1}$$

where n is the number of training samples and $\mathbf{x}_i'$ is the adversarial sample generated by adding the strongest perturbation to the natural one $\mathbf{x}_i$ within the ϵ-ball under L_p-norm distance, which can be used by the outer minimization of classification loss $\mathcal{L}$ to train a robust model. Research has shown that this is equivalent to optimizing an upper bound of natural risk on the original data, which implies that AT can serve as a principled defense against adversarial attacks [52].

PGD-AT [36] is the first method demonstrated to be effective for solving this $\min\max$ problem and training moderately robust DNNs [2, 59]. Based on *Danskin's Theorem* [10], PGD-AT proposed to find a constrained maximizer of the inner maximization by PGD, which is believed sufficiently close to the optimal attack, and then use the maximizer as an actual data point for the outer minimization through gradient descent. Another typical AT method, TRADES [72], proposed to decompose the robust error into natural error and boundary error to balance the trade-off between natural accuracy and robustness. Specifically, boundary error occurs when the specific data point is sufficiently close to the decision boundary that can easily cross it under slight perturbation. This is also believed as one reason for the existence of adversarial samples [3, 61]. One problem of TRADES is that the boundary error is designed to push each pair of benign and adversarial samples together, no matter whether the benign data are classified correctly or not [3]. As a follow-up work, MART [59] further investigates the influence of correctly classified and misclassified samples for adversarial robustness separately, and suggests applying robust error uniformly, plus boundary error only on misclassified samples.

A.3. Accuracy-Robustness Trade-Off in AT

One of the most important flaws of AT is an inherent trade-off between clean accuracy and adversarial robustness [3, 35, 40–42, 55, 57, 68, 72]. Specifically, the improvement in robustness achieved by AT is always at the cost of the reduction in model accuracy compared with standard training. As a consequence, AT inevitably degrades user experience under benign environments, which seriously hinders its real-world application. As shown in Figure 1, a widely recognized cause of this problem is that, to achieve predictive robustness under adversarial perturbation within the ϵ-ball of each natural data point uniformly, AT tends to learn a more complicated decision boundary than standard training [8, 13, 35, 40, 66, 68], which harms the generalization ability of the model on unseen data.

In response, it has recently become an active research area to variously craft or weight data points near the decision boundary for a better trade-off between accuracy and robustness. However, to date, there are still mixed views on the role and principle of these boundary samples in AT. **One side** believes the data closer to the current decision boundaries should be regarded as more critical and learned with larger weights or enhanced patterns. Representatively, GAIRAT [74] proposed that a natural data point closer to (or farther from) the class boundary is less (or more) robust, and the corresponding adversarial data point should be assigned with larger (or smaller) weight. MAIL [33] further revealed the unreliability of existing measures of the closeness, and proposed three types of probabilistic margin for measuring the closeness and reweighting adversarial data. **The other side** argues that reducing perturbations or weights for those data points near the decision boundary can bring significant benefits to AT generalization. For instance, MMA [11] proposed to use adaptive ϵ for adversarial perturbations to directly estimate and maximize the margin between data and the decision boundary. HAT [40] has a similar idea, which is realized by creating artificial helper-examples. TE [13] suggested that one-hot labels might be noisy for the boundary samples because they naturally lie close to the decision boundary, which makes it essentially difficult to assign high-confident one-hot labels for all perturbed samples within the ϵ-ball of them [8, 48]. So the model may try to memorize these hard samples during AT, leading to a more complicated decision boundary. All in all, with this inconclusive controversy, how to appropriately treat data points near the decision boundary is still an open question in AT.

Besides, some previous works also study the trade-off problem from other perspectives. For instance, AWP [62] proposed a double perturbation mechanism that can flatten the loss landscape by weight perturbation to improve robust generalization. EWAT [26] weighs the loss for each adversarial sample proportionally to the entropy of its prediction distribution during AT to focus on those with more uncertain labels. SOVR [25] proposed to increase logit margins of important samples by switching from cross-entropy to a new one-vs-therest loss for a considerable trade-off between accuracy and robustness. ADR [63] generates soft labels as a better guidance mechanism that accurately reflects the distribution shift under attack during AT. ReBAT [60] views AT as a dynamic mini-max game between the model trainer and the attacker, and proposes to rebalance the two players by either regularizing the trainer's capacity or improving the attack strength. PIAT [35] tunes the model parameters by interpolating them from the previous and current epochs, reducing the oscillations during the training process and moderating the change of decision boundaries. RPAT [58] suggests that the over-sufficient learning of hard adversarial samples degrades the decision boundary and contributes to the trade-off problem, and thus release it by encouraging the model perception to change smoothly with input perturbations.

At the same time, there are also some works theoretically doubting the assumption that adversarial robustness should be inherently at the cost of clean accuracy [3, 47, 67]. Representatively, Yang et al. [67] found that various common datasets are separated in class distribution with the separation larger than 2ϵ in usual, which directly indicates the existence of both robust and accurate classifiers and implies that the current trade-off problem is just an undesirable result caused by the existing AT methods themselves. Inspired by this, different from the previous works that aim at a better trade-off between accuracy

and robustness, the vision of this work is to directly harmonize them, such that the clean and robust accuracy scores can be improved concurrently.

A.4. AT with Consistency Regularization

Consistency Regularization (CR) aims to encourage a model to produce invariant representation for different variants of the same sample [14]. This is based on an assumption that slight perturbations like randomness within DNNs (*e.g.*, with Dropout) or data augmentation transformations should not modify the model prediction of the same input [75]. Typically, this idea can be realized through an additional consistency regularization term appended to the loss function, which has been widely adopted in existing Semi-Supervised Learning (SSL) algorithms [14], such as Π-model [29, 45], Temporal Ensembling [29], Mean Teacher [53] and Interpolation Consistency Training [56].

Recently, it has been found that there is an implicit connection between the goals of CR and AT [75]. Specifically, CR forces the model to give the same output distribution when the input or parameters are slightly perturbed, which covers the aim of AT when the perturbation is generated adversarially. Besides, as adversarial robustness essentially refers to model stability around naturally occurring inputs, learning to satisfy such a constraint should not inherently require labels [6], which provides the theoretical basis for the extension of CR, as an unsupervised mechanism, in the field of AT. Therefore, there have been some works exploring how to make use of CR ideas to improve adversarial robustness.

One of the most straightforward ideas is to introduce additional unlabeled data and their adversarial augmentations for consistency regularization [1, 6, 31, 37, 70], which can be viewed as a natural extension of the SSL algorithm Π-model in the context of robustness. However, these works have to rely on additional data, which not only brings extra costs in data collection and computation but also violates the conventional AT settings [38]. Subsequent works focus more on directly introducing the principle of CR into the existing AT framework. For instance, TE [13] makes use of the principle of Temporal Ensembling to maintain an ensemble prediction of each adversarial example and penalize the difference between the current prediction and the ensemble prediction. This is to regularize the predictions of adversarial examples from being over-confident, which is expected to reduce the impact of label noise during AT. Zhang et al. [75] integrates the strategy from Mean Teacher into AT to smooth the model and reduce possible overfitting. Specifically, it encourages the prediction distribution of a student model over adversarial examples to be consistent with that of a teacher model over clean samples. Liu et al. [34] and Kuang et al. [28] share similar ideas in introducing teacher-student learning strategies. Still, these methods basically transfer the existing CR methods to the AT task. Recently, Cons-AT [51] proposed a new consistency target directly from the perspective of robustness that the predictive distributions after attacking from two different augmentations of the same instance should be similar with each other. The underlying principle is that the most confusing class, and further, the most frequent attack direction, of specific samples is a kind of intrinsic information belonging to the so-called "dark" knowledge [21] and should be consistent over its different augmented variants.

Despite these previous efforts, the principle of CR in the context of AT remains highly unexplored, especially *w.r.t.* its potential on the accuracy-robustness trade-off problem. Specifically, these previous works mainly focus on the ability of CR to alleviate robust overfitting, with various theoretical explanations raised from different perspectives, such as model smoothness [7] and the flatness of the weight loss landscape [49, 62].[4] Predictably, the empirical results also show that, although contributing significantly smaller robust generalization gaps in the last epochs of AT, these works only slightly improve robustness at the best checkpoint of the model. What's worse, usually the price is to further hurt clean accuracy [13, 75]. The only aforementioned works taking the accuracy-robustness trade-off into consideration are the ones based on the Π-model [1, 6, 46]. They study the trade-off problem from the perspective of sample complexity, but again, merely under a relaxed assumption with extra data. All in all, to date, whether and how CR can benefit the accuracy-robustness trade-off in typical AT without extra data remains to be further studied.

B. Proof of Theorems

B.1. Proof of Theorem 1

Given the Gaussian model as in Definition 1, as a direct consequence of Gaussian concentration, Schmidt et al. [46] prove the following theorem, suggesting that we can learn a linear classifier achieving fixed and arbitrary classification error even with only one labeled sample.

[4] Robust overfitting refers to a phenomenon that, at certain epochs during AT (*e.g.*, after the first learning rate decay), model robustness will drop sharply, resulting in a significant gap in robust accuracy score between adversarially perturbed training and test data [24, 44].

Theorem 3 (Theorem 4 in Schmidt et al. [46]). *Let (x, y) be drawn from a (θ^*, σ)-Gaussian model with $\|\theta^*\|_2 = \sqrt{d}$ and $\sigma \leq c \cdot d^{1/4}$ where c is a universal constant. Let $\widehat{w} \in \mathbb{R}^d$ be the vector $\widehat{w} = y \cdot x$. Then with high probability, the linear classifier $f_{\widehat{w}}$ has classification error at most 1%.*

However, the following theorem shows that, to achieve the same robust generalization error under ℓ_∞ perturbation, any algorithm requires more labeled samples.

Theorem 4 (Theorem 6 in Schmidt et al. [46]). *Let g_n be any learning algorithm,* i.e.*, a function from n samples to a binary classifier f_n. Moreover, let $\sigma = c_1 \cdot d^{1/4}$, $\epsilon \geq 0$ and $\theta \in \mathbb{R}^d$ be drawn from $\mathcal{N}(0, I)$. We also draw n samples from the (θ, σ)-Gaussian model. Then the expected ℓ_∞^ϵ robust error of f_n is at least $\frac{(1-1/d)}{2}$ if*

$$n \;\leq\; c_2 \frac{\epsilon^2 \sqrt{d}}{\log d}\,, \tag{2}$$

where c_1 and c_2 are two universal constants.

Combining Theorem 3 and Theorem 4, we can see the sample complexity for robust generalization is greater than that of standard generalization by $\sqrt{d}/\log d$. As a direct corollary, given the same number of training samples, the expected gap between standard and robust generalization error can also be estimated by this.

Corollary 1. *Considering a (θ, σ)-Gaussian model with $\|\theta\|_2 = \sqrt{d}$ and $\sigma \leq c_1 \cdot d^{1/4}$, we have an expected gap between the lower bounds of standard and robust generalization error increasing with $O(\sqrt{d}/\log d)$.*

On the other hand, previous works have found that the sample complexity of robust generalization can be dramatically reduced by replacing labeled examples with additional unlabeled samples [1, 6, 37, 70]. Typically, Alayrac et al. [1] derive the following theorem, suggesting the specific number of additional samples needed.

Theorem 5 (Theorem 1 and Corollary 11 in Alayrac et al. [1]). *Consider the (θ^*, σ)-Gaussian model with $\|\theta^*\|_2 = \sqrt{d}$ and $\sigma \leq \frac{1}{32} d^{1/4}$ as in Schmidt et al. [46]. Let $\widehat{w}$ be the UAT-FT estimator as in Alayrac et al. [1]. Then with high probability, for n=1, the linear classifier $f_{\hat{w}}$ has ℓ_∞^ϵ robust classification error at most 1% if*

$$m \geq c\epsilon^2 \sqrt{d}, \tag{3}$$

where c is a fixed, universal constant.

As explained in Definition 2, Gaussian CR is a natural extension of Gaussian UAT-FT under the conventional fairness assumption of AT. The main reason for proposing Gaussian CR in this work is to highlight it is practically approximated in a completely different manner, though it follows the same theoretical principle as Gaussian UAT-FT. Therefore, along with Theorem 3, we have Corollary 2 by simply adopting Gaussian CR estimator in Theorem 5.

Corollary 2. *With the Gaussian CR estimator $\tilde{\theta}$ as in Definition 2, the generalization gap between standard and robust classifications can be filled at the cost of $O(\sqrt{d})$ additional unlabeled data points.*

Finally, through combining Corollary 1 and Corollary 2, we have Theorem 1 as in the main body.

B.2. Proof of Theorem 2

As introduced in Section 3, ICT [56] is a state-of-the-art method in the field of SSL. The following theorem in its original paper suggests that its CR term for unlabeled data can act as a regularizer on higher-order derivatives.

Theorem 6 (Theorem 1 in Verma et al. [56]). *Let $u, u' \in \mathbb{R}^d$ and f_θ be real analytic, and define $\Delta = u' - u$. Then, for any $K \in \mathbb{N}^+$, there exists a pair $(\zeta, \zeta' \in [0, \hat{\lambda}] \times [0, 1])$ such that*

$$\ell(f_\theta(\underset{\lambda}{\mathrm{Mix}}(u, u')), \underset{\lambda}{\mathrm{Mix}}(f_\theta(u), f_\theta(u'))) = \left(\sum_{k=2}^{K} \frac{(\hat{\lambda} - \hat{\lambda}^k)}{k!} \operatorname{vec}[\partial^k f_\theta(u)]^\top \Delta^{\otimes k} + E_\theta(u, u', \lambda) \right)^2, \tag{4}$$

where $E_\theta(u, u', \lambda) = \frac{1}{K!}((1-\zeta')^K \operatorname{vec}[\partial^{K+1} f_\theta(u + \zeta'\Delta)] - \hat{\lambda}(\hat{\lambda} - \zeta)^K \operatorname{vec}[\partial^{K+1} f_\theta(u + \zeta\Delta)])^\top \Delta^{\otimes K} = O(\|\Delta\|_2^K)$ and $\mathrm{Mix}_\lambda(a, b) = \lambda \cdot a + (1 - \lambda) \cdot b$. With the input normalized such that $\|\Delta\|_2 < 1$, $O(\|\Delta\|_2^K) \to 0$ as $K \to \infty$.

In this work, given $\dot{x}$ and $\ddot{x}$ as in Definition 3, let $u = \dot{x}$ and $u' = \ddot{x}$ in Theorem 7, we have:

$$\ell\Big(f_\theta(\lambda\dot{x} + \hat{\lambda}\ddot{x})\left(\lambda f_\theta(\dot{x}) + \hat{\lambda} f_\theta(\ddot{x})\right)\Big) = \left(\sum_{k=2}^{K} \frac{(\hat{\lambda} - \hat{\lambda}^k)}{k!} \operatorname{vec}[\partial^k f_\theta(\dot{x})]^\top \Delta^{\otimes k} + O(\|\Delta\|_2^K)\right)^2. \tag{5}$$

Further, to transfer this theorem to the context of AT, we introduce a limitation of a sufficiently small perturbation budget $\epsilon \to 0$ such that $\dot{x}' \to \dot{x}$ and $\ddot{x}' \to \ddot{x}$. Considering that $\dot{x}$ is augmented from x together, we have:

$$\ell\Big(f_\theta(\lambda\dot{x}' + \hat{\lambda}\ddot{x}')\left(\lambda f_\theta(\dot{x}) + \hat{\lambda} f_\theta(\ddot{x})\right)\Big) = \left(\sum_{k=2}^{K} \frac{(\hat{\lambda} - \hat{\lambda}^k)}{k!} \operatorname{vec}[\partial^k f_\theta(x)]^\top \Delta^{\otimes k} + O(\|\Delta\|_2^K)\right)^2. \tag{6}$$

Then with the $\operatorname{vec}[\partial^k f_\theta(x)] \in \mathbb{R}^{d^k}$ defined as in Verma et al. [56], we have Theorem 2 in the main body.

C. Robust Alignment Adversarial Training: Overall Risk, Objective, and Algorithms

In this section, we first formulate a new theoretical adversarial **risk** with the two main ideas proposed in the main text, and then propose a customizable surrogate **objective** for the practical optimization of this risk. Finally, under such a new surrogate objective, we suggest two specific methods that are expected to achieve a better accuracy-robustness trade-off in AT.

C.1. Constructing New Adversarial Risk for Robust Alignment

In this section, we formally define the proposed adversarial risk. For a K-class classification task ($K \geq 2$), given a dataset $\mathcal{S} = \{(\mathbf{x}_i, y_i)\}_{i=1,\dots,n}$ with $\mathbf{x}_i \in \mathbb{R}^d$ and $y_i \in \{1, ..., K\}$ respectively denoting a natural sample and its supervised label, as well as a DNN model $f_{\boldsymbol{\theta}}$ to predicts the input sample as $f_{\boldsymbol{\theta}}(\mathbf{x}_i) = \arg\max_{k=1,\dots,K} \mathbf{p}_k(\mathbf{x}_i, \boldsymbol{\theta})$, where $\mathbf{p}_k$ is the probability (*i.e.*, *softmax*) of class k in the prediction of $\mathbf{x}_i$, the standard adversarial risk [36] with respect to the *0-1 loss* [72] can be defined as follow:

$$\mathcal{R}(f_{\boldsymbol{\theta}}) = \frac{1}{n}\sum_{i=1}^{n} \max_{\|\mathbf{x}_i' - \mathbf{x}_i\|_p \leq \epsilon} \mathbb{1}(f_{\boldsymbol{\theta}}(\mathbf{x}_i') \neq y_i) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}(f_{\boldsymbol{\theta}}(\hat{\mathbf{x}}_i') \neq y_i), \tag{7}$$

where $\mathbb{1}(\cdot)$ is the indicator function and $\hat{\mathbf{x}}_i'$ is the adversarial sample generated through $\hat{\mathbf{x}}_i' = \arg\max_{\|\mathbf{x}_i' - \mathbf{x}_i\|_p \leq \epsilon} \mathbb{1}(f_{\boldsymbol{\theta}}(\mathbf{x}_i') \neq y_i)$. As this risk is defined on perturbed samples within the ϵ-ball of all natural samples no matter whether they are correctly classified or not, MART [59] proposed to reformulate the adversarial risk by dividing natural samples into the correctly classified subset $\mathcal{S}_{f_{\boldsymbol{\theta}}}^+ = \{\mathbf{x}_i : \mathbf{x}_i \in \mathcal{S}, f_{\boldsymbol{\theta}}(\mathbf{x}_i) = y_i\}$ and the misclassified subset $\mathcal{S}_{f_{\boldsymbol{\theta}}}^- = \{\mathbf{x}_i : \mathbf{x}_i \in \mathcal{S}, f_{\boldsymbol{\theta}}(\mathbf{x}_i) \neq y_i\}$ with respect to the current prediction of $f_{\boldsymbol{\theta}}$, and then defining the adversarial risk separately for them.

In this work, to utilize the phenomenon observed regarding the "boundary samples", we further propose to divide the correctly classified samples $\mathcal{S}_{f_{\boldsymbol{\theta}}}^+$ into the ones close to the classification boundary as $\mathcal{S}_{f_{\boldsymbol{\theta}}}^\circ$ and the ones away from that as $\mathcal{S}_{f_{\boldsymbol{\theta}}}^\bullet$. Formally, given a natural sample $\mathbf{x}_i \in \mathcal{S}_{f_{\boldsymbol{\theta}}}^+$ and the perturbation radius ϵ for adversarial sample $\hat{\mathbf{x}}_i'$, a *reduced adversarial sample* $\hat{\mathbf{x}}_i''$ can be denoted as:

$$\hat{\mathbf{x}}_i'' = \underset{\|\mathbf{x}_i'' - \mathbf{x}_i\|_p \leq \eta\cdot\epsilon}{\arg\max} \mathbb{1}(f_{\boldsymbol{\theta}}(\mathbf{x}_i'') \neq y_i), \tag{8}$$

where $\eta \in (0, 1)$ is a pre-defined hyper-parameter to reduce the perturbation. Then if $\exists\, \hat{\mathbf{x}}_i''$, s.t. $f_{\boldsymbol{\theta}}(\hat{\mathbf{x}}_i'') \neq y_i$, the natural sample $\mathbf{x}_i$ can be defined as a *boundary sample*, otherwise a *non-boundary sample*. Accordingly, we can define $\mathcal{S}_{f_{\boldsymbol{\theta}}}^\bullet$ and $\mathcal{S}_{f_{\boldsymbol{\theta}}}^\circ$ respectively as:

$$\mathcal{S}_{f_{\boldsymbol{\theta}}}^\bullet = \{\mathbf{x}_i : \mathbf{x}_i \in \mathcal{S}_{f_{\boldsymbol{\theta}}}^+,\ f_{\boldsymbol{\theta}}(\hat{\mathbf{x}}_i'') = y_i\}, \quad \mathcal{S}_{f_{\boldsymbol{\theta}}}^\circ = \{\mathbf{x}_i : \mathbf{x}_i \in \mathcal{S}_{f_{\boldsymbol{\theta}}}^+,\ f_{\boldsymbol{\theta}}(\hat{\mathbf{x}}_i'') \neq y_i\}. \tag{9}$$

Then we can define our adversarial risk separately for non-boundary samples $\mathcal{S}_{f_{\boldsymbol{\theta}}}^\bullet$, boundary samples $\mathcal{S}_{f_{\boldsymbol{\theta}}}^\circ$ and misclassified samples $\mathcal{S}_{f_{\boldsymbol{\theta}}}^-$. Notice that we have $\mathcal{S}_{f_{\boldsymbol{\theta}}}^\bullet \cup \mathcal{S}_{f_{\boldsymbol{\theta}}}^\circ \cup \mathcal{S}_{f_{\boldsymbol{\theta}}}^- = \mathcal{S}_{f_{\boldsymbol{\theta}}}^+ \cup \mathcal{S}_{f_{\boldsymbol{\theta}}}^- = \mathcal{S}$. First of all, we consider two kinds of risks for non-boundary samples. For one thing, as observed in Figure 2, to learn the basic adversarial robustness, it is vital that they are fully perturbed. For another, as we suggest in Section 3, additionally considering a domain interpolation consistency risk helps to learn better generalization. So for $\mathbf{x}_i \in \mathcal{S}_{f_{\boldsymbol{\theta}}}^\bullet$, given two different augmentations $\dot{\mathbf{x}}_i$ and $\ddot{\mathbf{x}}_i$, we define its adversarial risk as:

$$\mathcal{R}^\bullet(f_{\boldsymbol{\theta}}, \mathbf{x}_i) := \mathbb{1}(f_{\boldsymbol{\theta}}(\hat{\mathbf{x}}_i') \neq y_i) + \mathbb{1}\Big(f_{\boldsymbol{\theta}}\Big(\frac{\hat{\dot{\mathbf{x}}}_i' + \hat{\ddot{\mathbf{x}}}_i'}{2}\Big) \neq \Big(\frac{f_{\boldsymbol{\theta}}(\dot{\mathbf{x}}_i) + f_{\boldsymbol{\theta}}(\ddot{\mathbf{x}}_i)}{2}\Big)\Big). \tag{10}$$

In contrast, as we suggest in Section 2.1, applying perturbations of different intensities to boundary samples can hardly impact the final robustness, while moderate perturbations for them do benefit the generalization on clean data. Accordingly, we formulate the adversarial risk for $\mathbf{x}_i \in \mathcal{S}^{\circ}_{f_{\boldsymbol{\theta}}}$ as:

$$\mathcal{R}^{\circ}(f_{\boldsymbol{\theta}}, \mathbf{x}_i) := \mathbb{1}(f_{\boldsymbol{\theta}}(\hat{\mathbf{x}}_i'') \neq y_i) + \mathbb{1}\Big(f_{\boldsymbol{\theta}}\Big(\frac{\hat{\dot{\mathbf{x}}}_i' + \hat{\ddot{\mathbf{x}}}_i'}{2}\Big) \neq \Big(\frac{f_{\boldsymbol{\theta}}(\dot{\mathbf{x}}_i) + f_{\boldsymbol{\theta}}(\ddot{\mathbf{x}}_i)}{2}\Big)\Big). \tag{11}$$

Finally, this work does not involve improvement, as well as any specific requirements, for the misclassified samples. So for $\mathbf{x}_i \in \mathcal{S}^{-}_{f_{\boldsymbol{\theta}}}$, both directly using the standard adversarial risk as given in Equation (7) and adopting advanced ones such as $\mathcal{R}^{-}(f_{\boldsymbol{\theta}}, \mathbf{x}_i)^{\text{MMA}} = \mathbb{1}(f_{\boldsymbol{\theta}}(\mathbf{x}_i) \neq y_i)$ suggested by MMA and $\mathcal{R}^{-}(f_{\boldsymbol{\theta}}, \mathbf{x}_i)^{\text{MART}} = \mathbb{1}(f_{\boldsymbol{\theta}}(\hat{\mathbf{x}}_i') \neq y_i) + \mathbb{1}(f_{\boldsymbol{\theta}}(\mathbf{x}_i) \neq f_{\boldsymbol{\theta}}(\hat{\mathbf{x}}_i'))$ proposed by MART would be viable options.

Combining the three risk components above, we can obtain our novel adversarial risk involving the principle of the proposed *Robust Alignment* to be minimized as follows, based on which we end up with our new RAAT method.

$$\mathcal{R}(f_{\boldsymbol{\theta}}) := \sum_{\mathbf{x}_i \in \mathcal{S}^{\bullet}_{f_{\boldsymbol{\theta}}}} \mathcal{R}^{\bullet}(f_{\boldsymbol{\theta}}, \mathbf{x}_i) + \sum_{\mathbf{x}_i \in \mathcal{S}^{\circ}_{f_{\boldsymbol{\theta}}}} \mathcal{R}^{\circ}(f_{\boldsymbol{\theta}}, \mathbf{x}_i) + \sum_{\mathbf{x}_i \in \mathcal{S}^{-}_{f_{\boldsymbol{\theta}}}} \mathcal{R}^{-}(f_{\boldsymbol{\theta}}, \mathbf{x}_i). \tag{12}$$

C.2. Specific Method Proposed: RAAT & RAAT^{++}

We now provide our specific methods with surrogate objectives to optimize the new adversarial risk proposed above. Specifically, as the optimization over the *0-1 loss* is conceptual and not computationally tractable, we replace them with proper surrogate loss functions commonly adopted for DNNs to build practical AT algorithms. Firstly, for all the label-based indicator functions (*i.e.*, with input involving label y_i, corresponding to a supervised task), a natural surrogate function is *cross-entropy* (CE), which is the common choice for supervised classifications, as well as conventional AT methods. Furthermore, inspired by Carlini and Wagner [5], MART proposed to use *boosted cross-entropy* (BCE), which appends an additional margin term to CE to improve the decision margin of the classifier. Taking $\hat{\mathbf{x}}_i'$ for an example, with $\mathbf{p}_k(\hat{\mathbf{x}}_i', \boldsymbol{\theta})$ denoting the probability output as provided in Appendix C.1, these two surrogate losses can be denoted as:

$$\mathcal{L}^{\text{CE}}(\mathbf{p}(\hat{\mathbf{x}}_i', \boldsymbol{\theta}), y_i) = -\log\left(\mathbf{p}_{y_i}(\hat{\mathbf{x}}_i', \boldsymbol{\theta})\right), \tag{13}$$

$$\mathcal{L}^{\text{BCE}}(\mathbf{p}(\hat{\mathbf{x}}_i', \boldsymbol{\theta}), y_i) = \mathcal{L}^{\text{CE}}(\mathbf{p}(\hat{\mathbf{x}}_i', \boldsymbol{\theta}), y_i) - \log\left(1 - \max_{k \neq y_i} \mathbf{p}_k(\hat{\mathbf{x}}_i', \boldsymbol{\theta})\right). \tag{14}$$

Secondly, for the indicator function serving as a consistency regularization term in $\mathcal{R}^{\bullet}(f_{\boldsymbol{\theta}}, \mathbf{x}_i)$, we adopt *Jensen-Shannon divergence* (JS) as the surrogate loss. It is a symmetrized and smoothed variant of the commonly used *Kullback–Leibler divergence* (KL), and is suggested by Cons-AT in measuring the difference of output distributions under robust setting. Given $\dot{\mathbf{x}}_i$ and $\ddot{\mathbf{x}}_i$ as in Equation (10) and a parameter $\beta \in (0, 1)$, provided a *mixture distribution* $\mathbf{q}((\dot{\mathbf{x}}_i, \ddot{\mathbf{x}}_i), \boldsymbol{\theta})$ as:

$$\mathbf{q}((\dot{\mathbf{x}}_i, \ddot{\mathbf{x}}_i), \boldsymbol{\theta}) = \frac{1}{2}\left(\mathbf{p}((\beta \cdot \hat{\dot{\mathbf{x}}}_i' + (1-\beta) \cdot \hat{\ddot{\mathbf{x}}}_i'), \boldsymbol{\theta}) + (\beta \cdot \mathbf{p}(\dot{\mathbf{x}}_i, \boldsymbol{\theta}) + (1-\beta) \cdot \mathbf{p}(\ddot{\mathbf{x}}_i, \boldsymbol{\theta}))\right), \tag{15}$$

and $\mathcal{L}^{\text{KL}}(\cdot||\cdot)$ denoting the standard KL loss, then the JS consistency loss adopted in our method is:

$$\begin{aligned}
&\mathcal{L}^{\text{JS}}(\mathbf{p}((\beta \cdot \hat{\dot{\mathbf{x}}}_i' + (1-\beta) \cdot \hat{\ddot{\mathbf{x}}}_i'), \boldsymbol{\theta}) || (\beta \cdot \mathbf{p}(\dot{\mathbf{x}}_i, \boldsymbol{\theta}) + (1-\beta) \cdot \mathbf{p}(\ddot{\mathbf{x}}_i, \boldsymbol{\theta})))\\
&= \frac{1}{2}\left(\mathcal{L}^{\text{KL}}(\mathbf{p}((\beta \cdot \hat{\dot{\mathbf{x}}}_i' + (1-\beta) \cdot \hat{\ddot{\mathbf{x}}}_i'), \boldsymbol{\theta}) || \mathbf{q}((\dot{\mathbf{x}}_i, \ddot{\mathbf{x}}_i), \boldsymbol{\theta})) + \mathcal{L}^{\text{KL}}((\beta \cdot \mathbf{p}(\dot{\mathbf{x}}_i, \boldsymbol{\theta}) + (1-\beta) \cdot \mathbf{p}(\ddot{\mathbf{x}}_i, \boldsymbol{\theta})) || \mathbf{q}((\dot{\mathbf{x}}_i, \ddot{\mathbf{x}}_i), \boldsymbol{\theta}))\right).
\end{aligned} \tag{16}$$

The overall objective. Based on these surrogate loss functions, now we can state the overall objective function of our proposed method, ***R**obust **A**lignment **A**dversarial **T**raining* (**RAAT**). Note that for the misclassified part of the adversarial risk, according to the particular type of $\mathcal{R}^{-}(f_{\boldsymbol{\theta}}, \mathbf{x}_i)$ adopted, we can follow the corresponding original work proposing it to determine the specific surrogate loss to be used. Together with the different types of *cross-entropy*, this leaves room for our surrogate algorithms to be flexibly customized to fit different real-world practices better. Representatively, we provide basic RAAT and advanced RAAT^{++} methods, with their objectives $\mathcal{L}^{\text{RAAT}}(\mathbf{x}_i, y_i, \boldsymbol{\theta})$ and $\mathcal{L}^{\text{RAAT}^{++}}(\mathbf{x}_i, y_i, \boldsymbol{\theta})$ respectively provided in Table 4, where λ is a weight parameter to balance the supervised learning and consistency regularization term, and $\bar{\boldsymbol{\theta}}$ is a fixed copy of $\boldsymbol{\theta}$, which means the corresponding terms are just used as indicators instead of parts of the optimization objective. As suggested by Zhang et al. [71], the value of β follows *beta distribution* as $\beta \sim Beta(\gamma, \gamma)$ with the parameter $\gamma \in (0, \infty)$. Finally, through minimizing $\mathcal{L}^{\text{RAAT}}(\boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^{n} \mathcal{L}^{\text{RAAT}}(\mathbf{x}_i, y_i, \boldsymbol{\theta})$ or $\mathcal{L}^{\text{RAAT}^{++}}(\boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^{n} \mathcal{L}^{\text{RAAT}^{++}}(\mathbf{x}_i, y_i, \boldsymbol{\theta})$, we are expecting to train a DNN $\boldsymbol{\theta}$ with aligned input and latent spaces to achieve better accuracy and robustness at the same time.

Table 4. The overall optimization objectives of the proposed RAAT and RAAT^{++}, as well as the benchmarks involved in this work. Note that as mentioned in Appendix D.1, different from other methods, the adversarial sample $\hat{\mathbf{x}}'$ in TRADES is generated by maximizing its KL-divergence regularization term. Also, in practice, it is possible for Cons-AT to further utilize the benign augmentations (*i.e.*, $\dot{\mathbf{x}}_i$ and $\ddot{\mathbf{x}}_i$) in the supervised term of the loss, so do our methods.

AT Method	**Optimization Objective** ($loss$)
PGD-AT	$\mathcal{L}^{\text{CE}}(\mathbf{p}(\hat{\mathbf{x}}', \boldsymbol{\theta}), y)$
TRADES	$\mathcal{L}^{\text{CE}}(\mathbf{p}(\mathbf{x}, \boldsymbol{\theta}), y) + \lambda \cdot \mathcal{L}^{\text{KL}}(\mathbf{p}(\mathbf{x}, \boldsymbol{\theta}) \| \mathbf{p}(\hat{\mathbf{x}}', \boldsymbol{\theta}))$
MART	$\mathcal{L}^{\text{BCE}}(\mathbf{p}(\hat{\mathbf{x}}', \boldsymbol{\theta}), y) + \lambda \cdot \mathcal{L}^{\text{KL}}(\mathbf{p}(\mathbf{x}, \boldsymbol{\theta}) \| \mathbf{p}(\hat{\mathbf{x}}', \boldsymbol{\theta})) \cdot (1 - \mathbf{p}_y(\mathbf{x}, \boldsymbol{\theta}))$
Cons-AT	$\mathcal{L}^{\text{CE}}(\mathbf{p}(\hat{\mathbf{x}}', \boldsymbol{\theta}), y) + \lambda \cdot \mathcal{L}^{\text{JS}}(\mathbf{p}(\hat{\mathbf{x}}_i', \boldsymbol{\theta}) \| \mathbf{p}(\hat{\ddot{\mathbf{x}}}_i', \boldsymbol{\theta}))$
RAAT	$\mathcal{L}^{\text{CE}}(\mathbf{p}(\hat{\mathbf{x}}_i', \boldsymbol{\theta}), y_i) \cdot \mathbb{1}(h_{\bar{\boldsymbol{\theta}}}(\hat{\mathbf{x}}_i'') = y_i) + \mathcal{L}^{\text{CE}}(\mathbf{p}(\hat{\mathbf{x}}_i'', \boldsymbol{\theta}), y_i) \cdot \mathbb{1}(h_{\bar{\boldsymbol{\theta}}}(\hat{\mathbf{x}}_i'') \neq y_i)$ $+\lambda \cdot \mathcal{L}^{\text{JS}}(\mathbf{p}((\beta \cdot \hat{\mathbf{x}}_i' + (1-\beta) \cdot \hat{\ddot{\mathbf{x}}}_i'), \boldsymbol{\theta}) \| (\beta \cdot \mathbf{p}(\dot{\mathbf{x}}_i, \boldsymbol{\theta}) + (1-\beta) \cdot \mathbf{p}(\ddot{\mathbf{x}}_i, \boldsymbol{\theta}))) \cdot \mathbb{1}(h_{\bar{\boldsymbol{\theta}}}(\mathbf{x}_i) = y_i)$
RAAT^{++}	$\mathcal{L}^{\text{BCE}}(\mathbf{p}(\hat{\mathbf{x}}_i', \boldsymbol{\theta}), y_i) \cdot \mathbb{1}(h_{\bar{\boldsymbol{\theta}}}(\hat{\mathbf{x}}_i'') = y_i) + \mathcal{L}^{\text{BCE}}(\mathbf{p}(\hat{\mathbf{x}}_i'', \boldsymbol{\theta}), y_i) \cdot \mathbb{1}(h_{\bar{\boldsymbol{\theta}}}(\hat{\mathbf{x}}_i'') \neq y_i)$ $+\lambda \cdot \big(\mathcal{L}^{\text{JS}}(\mathbf{p}((\beta \cdot \hat{\mathbf{x}}_i' + (1-\beta) \cdot \hat{\ddot{\mathbf{x}}}_i'), \boldsymbol{\theta}) \| (\beta \cdot \mathbf{p}(\dot{\mathbf{x}}_i, \boldsymbol{\theta}) + (1-\beta) \cdot \mathbf{p}(\ddot{\mathbf{x}}_i, \boldsymbol{\theta}))) \cdot \mathbb{1}(h_{\bar{\boldsymbol{\theta}}}(\mathbf{x}_i) = y_i) + \mathcal{L}^{\text{KL}}(\mathbf{p}(\mathbf{x}_i, \boldsymbol{\theta}) \| \mathbf{p}(\hat{\mathbf{x}}_i', \boldsymbol{\theta})) \cdot (1 - \mathbf{p}_{y_i}(\mathbf{x}_i, \boldsymbol{\theta}))\big)$

D. Additional Experimental Details

D.1. Experimental Setup

Following Pang et al. [38], which explores a great number of hyper-parameter settings for AT, we adopt the following common settings. For outer minimization, we use SGD optimizer with momentum 0.9, batch size 128, weight decay 5×10^{-4}, initial learning rate 0.1, total 110 training epochs with learning rate decay by a factor of 0.1 at 100 and 105 epochs, respectively. The only exception is in Section 5.3, where we adopt 200 training epochs with learning rate decay at 100 and 150 epochs to strictly ensure fairness in comparison with the current SOTAs. For the inner maximization, under the ℓ_∞ threat model with maximal perturbation budget $\epsilon = 8/255$, we adopt step size $\alpha = 2/255$ with 10-step adversaries (*i.e.*, PGD-10 except for TRADES, in which the adversarial samples are generated by maximizing its KL-divergence regularization term [72]). While under the ℓ_2 threat model with maximal perturbation budget $\epsilon = 128/255$, we adopt step size $\alpha = 32/255$. Same as their original papers, we set the regularization parameter $\lambda = 6$ for TRADES and MART, as well as $\lambda = 1$ for Cons-AT and ours. Also, we use default $\eta = 0.1$ to acquire boundary samples and fix $\gamma = 0.75$ for the *beta distribution*. For data pre-processing, we normalize all natural images into [0, 1], and adopt standard data augmentations including random crop with 4-pixel zero padding and random horizontal flip with 50% of probability.

The experiments are conducted on Ubuntu 22.04 OS with Intel Xeon Platinum 8336C 32-Core 2.3GHz CPU, 512GB RAM and 8 $\times$ NVIDIA GeForce RTX 4090 GPUs, and are implemented with Python 3.8.19 and PyTorch 1.8.1+cu111.

D.2. Ablation Study

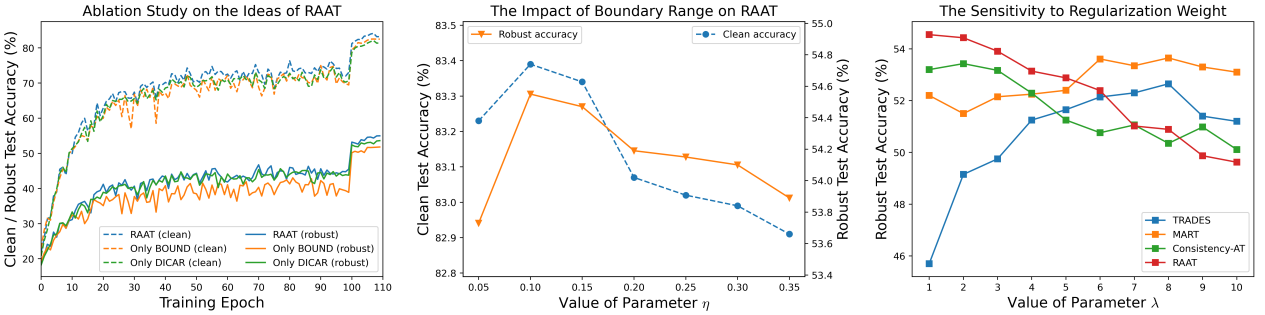
To better examine the effectiveness of the proposed defense, we investigate RAAT from the following perspectives: I) the effectiveness of the two ideas adopted by RAAT (Figure 5 (a)); II) the impact of the specific boundary range, which is determined by η (Figure 5 (b)); and III) the sensitivity to the regularization parameter λ (Figure 5 (c)).

Firstly, as we expected, both reducing perturbation for boundary samples (referred to as "BOUND") and adding the DICAR term benefit clean accuracy and robustness at the same time, concurrently contributing to the final effectiveness of RAAT in improving the current trade-off problem. Secondly, the experiments on η confirm our findings in Figure 2 (c) and (f) that an appropriate partition of boundary and non-boundary samples helps achieve better performance on both perturbed and clean data. Although η is fixed at 0.1 in our experiments for simplicity, we have no objection to fine-tuning in real-world practices. It might contain certain underlying significance and bring extra improvement especially for datasets with varying class numbers.

Finally, we explore various values of the weight λ for all the experimental methods having a regularization term. Here we focus on robustness as it is what the adversarial regularization terms mainly impact. Our results show the best λ for TRADES and MART is around 6 to 8, aligning with their original papers. In contrast, $\lambda \in [1, 3]$ suits Cons-AT and RAAT more. This is probably because, as mentioned at the end of Section 3, they inject stronger consistency-based inductive biases into models

Table 5. Experimental results on CIFAR-10, CIFAR-100, and Tiny-ImageNet datasets with ResNet-18 architecture under the ℓ_2 threat model. All the results are averages of three runs. The final row for each dataset marked with $\uparrow$ (%) shows the percentage improvement of the best of RAAT/RAAT^{++} over the best of the four benchmarks.

Dataset	Method	ResNet-18				
		Clean	PGD-10	PGD-100	C&W	AA
CIFAR-10	PGD-AT	87.76$_{\pm 0.51}$	68.82$_{\pm 0.23}$	67.72$_{\pm 0.23}$	66.64$_{\pm 0.43}$	66.36$_{\pm 0.37}$
	TRADES	83.99$_{\pm 0.28}$	68.74$_{\pm 0.32}$	68.59$_{\pm 0.63}$	67.05$_{\pm 0.55}$	65.93$_{\pm 0.30}$
	MART	84.09$_{\pm 0.19}$	68.96$_{\pm 0.37}$	68.19$_{\pm 0.21}$	67.22$_{\pm 0.43}$	66.28$_{\pm 0.27}$
	Cons-AT	88.76$_{\pm 0.41}$	70.21$_{\pm 0.49}$	69.16$_{\pm 0.45}$	68.10$_{\pm 0.42}$	67.46$_{\pm 0.26}$
	RAAT	**89.17**$_{\pm 0.26}$	69.76$_{\pm 0.35}$	68.67$_{\pm 0.17}$	**68.13**$_{\pm 0.30}$	67.95$_{\pm 0.24}$
	RAAT^{++}	86.54$_{\pm 0.33}$	**70.84**$_{\pm 0.17}$	**70.01**$_{\pm 0.15}$	69.07$_{\pm 0.28}$	**67.96**$_{\pm 0.24}$
	$\uparrow$ (%)	+0.46%	+0.90%	+1.23%	+0.04%	+0.74%
CIFAR-100	PGD-AT	65.00$_{\pm 0.41}$	42.47$_{\pm 0.29}$	41.29$_{\pm 0.18}$	40.03$_{\pm 0.17}$	39.27$_{\pm 0.39}$
	TRADES	61.25$_{\pm 0.48}$	43.52$_{\pm 0.58}$	42.96$_{\pm 0.19}$	41.22$_{\pm 0.38}$	40.15$_{\pm 0.42}$
	MART	60.08$_{\pm 0.19}$	43.92$_{\pm 0.16}$	43.28$_{\pm 0.35}$	41.63$_{\pm 0.33}$	39.85$_{\pm 0.20}$
	Cons-AT	65.14$_{\pm 0.30}$	43.16$_{\pm 0.37}$	42.09$_{\pm 0.32}$	40.99$_{\pm 0.36}$	39.92$_{\pm 0.26}$
	RAAT	**66.12**$_{\pm 0.11}$	43.47$_{\pm 0.20}$	42.22$_{\pm 0.26}$	41.33$_{\pm 0.17}$	40.25$_{\pm 0.20}$
	RAAT^{++}	62.43$_{\pm 0.09}$	**44.95**$_{\pm 0.15}$	**43.96**$_{\pm 0.12}$	**42.35**$_{\pm 0.23}$	**40.61**$_{\pm 0.15}$
	$\uparrow$ (%)	+1.50%	+2.34%	+1.57%	+1.73%	+0.90%
Tiny-ImageNet	PGD-AT	59.36$_{\pm 0.18}$	43.52$_{\pm 0.16}$	43.05$_{\pm 0.20}$	42.81$_{\pm 0.18}$	40.90$_{\pm 0.24}$
	TRADES	57.61$_{\pm 0.44}$	44.84$_{\pm 0.42}$	44.65$_{\pm 0.54}$	43.23$_{\pm 0.21}$	42.33$_{\pm 0.27}$
	MART	56.47$_{\pm 0.22}$	45.05$_{\pm 0.23}$	44.77$_{\pm 0.16}$	43.10$_{\pm 0.15}$	42.22$_{\pm 0.16}$
	Cons-AT	**61.49**$_{\pm 0.24}$	45.32$_{\pm 0.28}$	44.85$_{\pm 0.19}$	43.57$_{\pm 0.30}$	42.59$_{\pm 0.25}$
	RAAT	61.20$_{\pm 0.12}$	**45.43**$_{\pm 0.13}$	44.66$_{\pm 0.19}$	**43.82**$_{\pm 0.13}$	**42.63**$_{\pm 0.11}$
	RAAT^{++}	57.95$_{\pm 0.10}$	45.38$_{\pm 0.12}$	**44.92**$_{\pm 0.12}$	43.61$_{\pm 0.08}$	42.35$_{\pm 0.11}$
	$\uparrow$ (%)	-0.47%	+0.24%	+0.16%	+0.57%	+0.09%



(a) Ablation of two ideas. (b) Impact of boundary range. (c) Sensitivity to regularization weight.

Figure 5. Three comprehensive ablation experiments of the proposed RAAT on CIFAR-10. The blue line in (a) and the points respectively indicating $\eta = 0.1$ and $\lambda = 1$ in (b) and (c) correspond to the RAAT record under ResNet-18 in Table 1.

than individual adversarial samples-based conventional AT, making them more sensitive to inappropriately large λ values. Overall, compared to optimal cases of all these methods among the entire experimental value range, RAAT achieves the best robustness.