BayesAT: Bayes-Guided Progressive Distillation for Semi-Supervised Adversarial Training

Yimo Guo[1] **Lilin Zhang**[1] **Penglin Li**[1] **Jinhui Hao**[2] **Xianggen Liu**[1]*
[1]College of Computer Science, Sichuan University
[2]College of Physics, Sichuan University
{guoyimo, zhanglilin, lipenglin, haojinhui}@stu.scu.edu.cn
liuxianggen@scu.edu.cn

Abstract

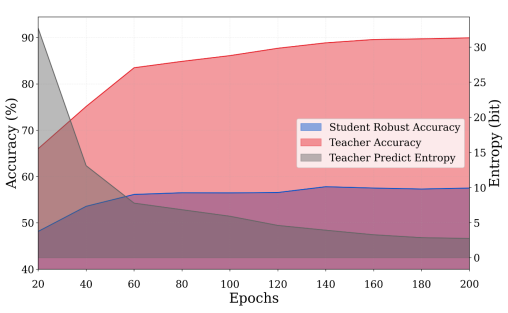
Existing semi-supervised adversarial training (SSAT) methods typically employ a teacher-student framework where a teacher model provides supervisory signals for unlabeled data. However, they rely on **static supervision**—either hard pseudo-labels or soft labels with fixed temperature—which maintains constant learning difficulty throughout training. This rigidity fails to accommodate the student's evolving capability, especially when the teacher's performance plateaus. To address this, we **reframe SSAT as knowledge distillation from a Bayes teacher** and formalize a *stage-dependent* bias-variance tradeoff. This analysis reveals a temperature-governed bias–variance tradeoff: low temperatures reduce bias, while higher temperatures control the variance term in the robust generalization bound. Guided by this insight, we propose **BayesAT**, a Bayes-guided progressive distillation framework that jointly employs a low-to-high temperature warming schedule and confidence-adaptive sample reweighting. BayesAT allows the student to first learn from sharp, high-confidence supervision to establish reliable decision boundaries, then progressively from smoother distributions that encourage exploration of inter-class relationships. Extensive experiments on CIFAR-10, CIFAR-100, and ImageNet-200 demonstrate that BayesAT consistently improves robust accuracy while maintaining natural accuracy over existing SSAT methods, highlighting the importance of dynamic, theory-driven supervision for effective knowledge transfer in semi-supervised adversarial training. Our code is available at https://github.com/Sherlockssa/BayesAT.

1 Introduction

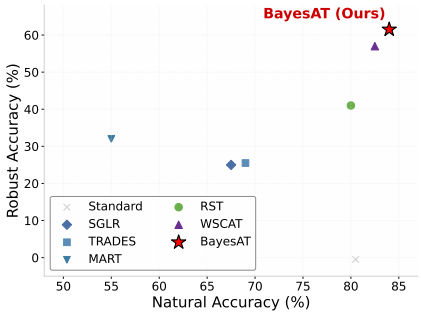
Deep neural networks remain vulnerable to adversarial perturbations [20, 6, 2, 9, 16]. Adversarial Training (AT) [14, 31, 1] is widely regarded as the most effective defense, yet its reliance on large labeled datasets incurs substantial annotation cost. Semi-Supervised Adversarial Training alleviates this issue by leveraging unlabeled data, typically via a teacher-student framework where a teacher provides pseudo-label supervision [4]. Existing approaches improve pseudo-label quality through techniques such as exponential moving averages [21] and consistency regularization [23].

Despite their differences, these methods share a common design choice: they rely on **static supervisory signals**. Whether in the form of hard pseudo-labels or soft labels with a fixed temperature, the supervision remains unchanged throughout training. This implicitly assumes a stationary learning process, while in practice the student evolves substantially. Consequently, static supervision induces a supply-demand mismatch across training stages: early training benefits from low-bias, high-confidence signals to establish stable decision boundaries, whereas later stages require smoother

*Corresponding author.



(a) Teacher performance and output smoothness



(b) Natural accuracy *vs.* Robust accuracy

Figure 1: (a) Robust accuracy and output smoothness of the teacher model during training. After sufficient training, the teacher enters a plateau and produces increasingly over-confident predictions. (b) Performance comparison on CIFAR-10. Existing SSAT methods concentrate within a constrained frontier, consistent with the limitation of static supervision.

supervision to control variance and encode inter-class structure. A fixed-temperature strategy therefore becomes inherently suboptimal over the full training trajectory.

To formalize this limitation, we revisit SSAT from a Bayesian knowledge distillation perspective and analyze the induced *bias–variance tradeoff* with respect to a Bayes teacher. The temperature parameter governs this tradeoff: low temperatures produce sharp distributions with low bias but high variance, while high temperatures yield smoother distributions that reduce variance at the cost of increased bias. Crucially, the preferred balance is *stage-dependent*, suggesting that a fixed temperature may not remain optimal throughout training. This stage dependency contrasts with standard knowledge distillation, where fixed or cooling schedules are typically sufficient, as teachers are often well-calibrated high-capacity models rather than plateaued ones.

This limitation is further amplified by a characteristic training phenomenon. As shown in Fig. 1(a), the teacher enters a performance plateau after sufficient training, where accuracy saturates and predictions become increasingly over-confident. In this regime, the teacher's approximation bias approaches a lower bound, while the over-confident predictions become hypersensitive to input perturbations, amplifying the variance of the supervision signal and making it the dominant factor in the generalization bound. Static temperature strategies fail to adapt to this shift. Figure 1(b) confirms this observation: existing SSAT methods cluster along a suboptimal frontier, unable to break through the ceiling imposed by static supervision.

Motivated by these observations, we propose **BayesAT**, a Bayes-guided progressive distillation framework that adapts supervision difficulty via a *low-to-high temperature warming schedule*. The schedule begins with low temperatures to establish reliable decision boundaries, and gradually increases to control the variance term in the generalization bound and capture richer class relationships. We further introduce a **confidence-adaptive sample reweighting** mechanism that complements this progression by prioritizing reliable samples in early stages and gradually shifting focus toward more ambiguous ones. Our contributions are summarized as follows:

- **A Bayesian Bias–variance Perspective on SSAT.** We reformulate SSAT as a distillation problem and reveal an inherent stage-dependent tradeoff in teacher supervision, explaining why static strategies are suboptimal.
- **Stage-adaptive Temperature Scheduling.** We propose a low-to-high temperature warming strategy grounded in bias–variance analysis, aligning supervision with the student's evolving needs.
- **Confidence-adaptive Reweighting.** We design a dynamic weighting mechanism that progressively shifts the learning focus from reliable samples to more ambiguous ones.
- **Consistent Empirical Improvements.** Extensive experiments on CIFAR-10, CIFAR-100, and ImageNet-200 demonstrate consistent improvements in robust accuracy over prior methods.

2 Related Work

Semi-supervised Adversarial Training. Adversarial Training [14] is an effective defense against adversarial attacks but relies heavily on labeled data. Semi-Supervised Adversarial Training alleviates

this limitation by leveraging unlabeled data, often building upon semi-supervised learning techniques such as Mean Teacher [21] and FixMatch [19]. RST [4] first demonstrated that unlabeled data can substantially improve adversarial robustness. Subsequent work has improved adversarial training objectives in the fully supervised setting. TRADES [26] introduces a principled trade-off between accuracy and robustness, while MART [22] emphasizes misclassified examples. In the semi-supervised setting, recent methods explore more effective ways to utilize unlabeled data. SGLR [13] refines pseudo-labels via self-guided label smoothing to mitigate noisy supervision, while WSCAT [30] incorporates contrastive learning to improve representation robustness under weak supervision. Beyond class-balanced benchmarks, recent works study adversarial robustness under long-tailed distributions: Zhang et al. [28] rebalance adversarial training by adaptively adjusting the perturbation intensity for each class, while Zhang et al. [29] attribute degraded worst-class robustness to a posterior mismatch and provide posterior-based probabilistic supervision that tightens the robust generalization bound. Both target supervised long-tailed settings and are complementary to our stage-adaptive supervision for semi-supervised adversarial training. Despite these advances, most SSAT methods rely on either hard pseudo-labels or fixed-temperature soft labels, resulting in largely static supervision throughout training. In contrast, we consider stage-adaptive supervision that evolves with the learning process.

Temperature Scheduling in Knowledge Distillation. Knowledge Distillation (KD) [8] transfers knowledge via softened output distributions, where temperature controls label smoothness. Most KD methods adopt fixed temperatures or simple predefined schedules (e.g., high-to-low annealing) [7, 12]. Prior work has also connected distillation to label smoothing and soft targets [18, 24], and explored teacher–student dynamics in knowledge transfer [17]. While adaptive or curriculum-based strategies have been explored, they are primarily designed for fully supervised settings with well-calibrated teachers and clean labels. In SSAT, however, pseudo-labels are inherently noisy and the teacher evolves during training, leading to different optimization dynamics and motivating alternative temperature scheduling strategies. For example, Kukleva et al. [11] introduced a cosine oscillation temperature schedule for the InfoNCE loss to alternate between instance-wise and group-wise discrimination. We discuss this temperature scheduling mechanism and its distinction from our approach in Appendix D.

Confidence-aware Sample Weighting. Sample reweighting has been widely used to improve robustness and handle noisy supervision. GAIRAT [27] assigns weights based on the geometric difficulty of samples, while MART [22] emphasizes misclassified examples. These methods demonstrate the importance of prioritizing informative or reliable samples, but are typically studied independently from label smoothing or temperature scaling. Our work highlights their complementarity and integrates both within a unified, adaptive framework.

3 BayesAT: Bayes-Guided Progressive Distillation

3.1 Semi-supervised Adversarial Training

Let $\mathcal{X} \subset \mathbb{R}^d$ denote the input data space, $\mathcal{Y} = \{y_i\}_{i\in[C]}$ the set of C classes, and let $\mathbb{P}$ be the underlying joint distribution over $\mathcal{X} \times \mathcal{Y}$. In the standard supervised adversarial training setting, each sample x is assigned to the class $y = \arg\max_{i\in[C]} \mathbb{P}(y_i|x)$, which induces the one-hot label scheme $\mathbf{y}(x) = [\mathbb{1}\{y_i = y\}]_{i\in[C]} \in \{0,1\}^C$. This one-hot label is an unbiased estimator of the Bayes class probability $\mathbf{p}^*(x) = [\mathbb{P}(y|x)]_{y\in[C]} \in [0,1]^C$ [15] as $\mathbb{E}[\mathbf{y}(x)] = \mathbf{p}^*(x)$. Let $f : \mathcal{X} \to \mathbb{R}^C$ denote the target classifier. For a data-label pair (x, y), the cross-entropy loss $\mathbf{y}(x)^\top \ell(x)$ measures the classification error of f, where $\ell(x) = -\log f(x) \in \mathbb{R}^C$. We consider the widely used ℓ_∞ threat model with perturbation budget ϵ, under which the adversarial example corresponding to x is $\hat{x} = \arg\max_{\hat{x}:\|\hat{x}-x\|_\infty\leq\epsilon} \mathbf{y}(x)^\top \ell(\hat{x})$. The population adversarial risk, which characterizes the adversarial robustness of f, is then defined as:

$$R(f) = \mathbb{E}_{(x,y)\sim\mathbb{P}}\big[\mathbf{y}(x)^\top \ell(\hat{x})\big] = \mathbb{E}_{(x,y)\sim\mathbb{P}}\big[\mathbf{p}^*(x)^\top \ell(\hat{x})\big]. \tag{1}$$

In practical semi-supervised settings, only a dataset $\mathcal{D} = \mathcal{D}_l \cup \mathcal{D}_u$ is available, where $\mathcal{D}_l = \{(x_i, y_i) \sim \mathbb{P}\}_{i=1}^{N_l}$ is the labeled set and $\mathcal{D}_u = \{x_j \sim \mathbb{P}\}_{j=1}^{N_u}$ is the unlabeled set, with $N_u \gg N_l$. Traditional adversarial training minimizes the empirical adversarial risk over the labeled dataset $\hat{R}_{\text{AT}}(f) = \mathbb{E}_{(x,y)\in\mathcal{D}_l}\big[\mathbf{y}(x)^\top \ell(\hat{x})\big]$. Since the standard adversarial training objective heavily relies on label supervision, semi-supervised adversarial training methods leverage unlabeled data by first using

a surrogate model to generate pseudo-labels y_j for each unlabeled sample $x_j \in \mathcal{D}_u$. This produces a pseudo-labeled dataset $\tilde{\mathcal{D}}_u = \{(x_j, y_j)\}_{j=1}^{N_u}$. Accordingly, the empirical risk becomes:

$$\hat{R}_{\text{SSAT}}(f) = \mathbb{E}_{(x,y)\in\mathcal{D}_l\cup\tilde{\mathcal{D}}_u}\big[\mathbf{y}(x)^\top \ell(\hat{x})\big]. \tag{2}$$

3.2 Bias-variance Tradeoff Bound on Robust Generalization Gap

We theoretically investigate how the quality of teacher-generated pseudo-labels influences the robust generalization performance of SSAT, and derive insights for method design. Specifically, we first reinterpret SSAT from the perspective of knowledge distillation, and then identify a bias-variance tradeoff that upper bounds the robust generalization gap.

Unified Formulation from the Distillation Perspective. In conventional SSAT, the surrogate model used to generate pseudo-labels can be viewed as a teacher model, while the adversarially trained model serves as the student model. Under this perspective, SSAT can be naturally reformulated as a distillation framework. Let the teacher model be $t : \mathcal{X} \to \mathbb{R}^C$, which generates labels for both labeled and unlabeled samples. The soft labels produced by teacher t are defined as:

$$\mathbf{p}_t(x;\tau) = \text{softmax}\Big(\frac{z_t(x)}{\tau}\Big) = \left[\frac{\exp(z_{t,i}(x)/\tau)}{\sum_{j\in[C]}\exp(z_{t,j}(x)/\tau)}\right]_{i\in[C]}, \tag{3}$$

where $z_t(x) \in \mathbb{R}^C$ denotes the teacher logits for input x, $z_{t,i}(x)$ is the i-th logit, and $\tau \geq 0$ is the temperature parameter controlling the sharpness of the generated labels.

Lemma 3.1 (Label Smoothness)**.** *Let $z_t(x) \in \mathbb{R}^C$ denote the teacher logits for input x. If for any $x \sim \mathbb{P}$, $i \in [C]$ and $j \in [C]$, $z_{t,i}(x) + z_{t,j}(x) \geq 0$ holds, then*

$$\nabla_\tau \|\mathbf{p}_t(x;\tau)\|_2^2 \leq 0.$$

Remark 3.1. $z_{t,i}(x) + z_{t,j}(x) \geq 0$ is natural in practice because logits can always be shifted by a constant without changing the softmax probabilities. This ensures that the softmax outputs become progressively smoother as the temperature τ increases. Intuitively, increasing τ spreads the probability mass more evenly across classes, reducing the ℓ_2 norm of the label vector. Traditional SSAT methods that use hard pseudo-labels correspond to the limit $\tau \to 0$, which is a special case of this formulation.

Under this perspective, the population risk and empirical risk of SSAT on unlabeled data can be written in the following unified form:

$$R_t(f) = \mathbb{E}_{x\sim\mathbb{P}}\big[\mathbf{p}_t(x;\tau)^\top \ell(\hat{x})\big]; \hat{R}_t(f) = \mathbb{E}_{x\in\mathcal{D}_u}\big[\mathbf{p}_t(x;\tau)^\top \ell(\hat{x})\big]. \tag{4}$$

Student Robust Generalization Bound. We next show that a Bayes teacher, which perfectly estimates $\mathbf{p}^*(x)$, yields the smallest robust generalization gap for the student model. Let $\mathbb{V}$ denote the variance of a random variable. We first show that the loss induced by the Bayes teacher, $R^*(f) = \mathbb{E}_{x\sim\mathbb{P}}[\mathbf{p}^*(x)^\top \ell(\hat{x})]$ has lower variance.

Lemma 3.2 (Loss Variance)**.** *For any f with $\ell(x) = -\log f(x)$, $\mathbb{V}[\mathbf{p}^*(x)^\top \ell(\hat{x})] \leq \mathbb{V}[\mathbf{y}(x)^\top \ell(\hat{x})]$.*

Remark 3.2. Lemma 3.2 indicates that the Bayes teacher with labeling scheme $\mathbf{p}^*(x)$ leads to lower loss variance than the one-hot labeling scheme $\mathbf{y}(x)$.

We then bound the robust generalization gap of the student model f.

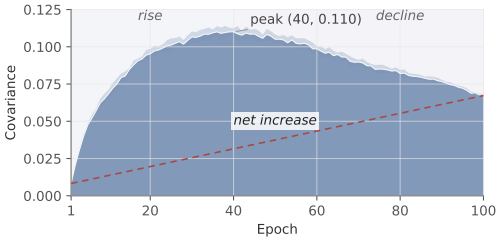


Figure 2: The covariance term $\text{Cov}(\|\mathbf{p}_t(x;\tau)\|_2, \|\ell(\hat{x})\|_2)$ on CIFAR-10 (mean over three random seeds; shaded band: standard deviation across seeds). It peaks near epoch 40 and remains above its initial value (red dashed line: net increase), exhibiting the stage-dependent variance.

Theorem 3.1 (Generalization Gap)**.** *With probability at least $1-\delta$, the following bound holds:*

$$R(f) - \hat{R}_t(f) \leq \underbrace{\mathbb{E}_{x\sim\mathbb{P}}\big[\big(\mathbf{p}^*(x) - \mathbf{p}_t(x;\tau)\big)^\top \ell(\hat{x})\big]}_{\textit{Bias}} + \underbrace{\sqrt{\frac{2\mathbb{V}[\mathbf{p}_t(x)^\top \ell(\hat{x})]\log(1/\delta)}{N_u}}}_{\textit{Variance}} + \frac{\log(1/\delta)}{3N_u}.$$

Remark 3.3. The Bayes teacher yields a tighter generalization bound because it: (i) minimizes the variance term, as shown in Lemma 3.2, and (ii) incurs zero bias term. Therefore, the closer the teacher is to the Bayes teacher, the better the robust generalization performance of the student model.

Teacher Bias-variance Tradeoff. However, obtaining an ideal teacher is challenging because $\mathbf{p}^*(x)$ is generally intractable. In practice, estimating $\mathbf{p}^*(x)$ introduces a bias–variance tradeoff in the generalization bound. We first derive the upper bound of the bias term in Theorem 3.1.

Theorem 3.2 (Bias Bound)**.** *The bias* $\mathbb{E}_{x\sim\mathbb{P}}\big[\big(\mathbf{p}^*(x)-\mathbf{p}_t(x;\tau)\big)^\top \ell(\hat{x})\big]$ *can be upper bounded as* $\mathbb{E}_{x\sim\mathbb{P}}\big[\big(\mathbf{p}^*(x)-\mathbf{p}_t(x;\tau)\big)^\top \ell(\hat{x})\big] \le \mathbb{E}_{(x,y)\in\mathcal{D}_l\cup\mathcal{D}_u}\big[\big(\mathbf{y}(x)-\mathbf{p}_t(x;\tau)\big)^\top \ell(\hat{x})\big] + L\sqrt{\frac{\log(1/\delta)}{2(N_l+N_u)}}$ *with at least* $1-\delta$ *probability, where* $L=\sup_{x\sim\mathbb{P}}\{(\mathbf{y}(x)-\mathbf{p}_t(x;\tau))^\top \ell(\hat{x})\}$.

Remark 3.4. To reduce the bias, Theorem 3.2 indicates that the teacher should be trained so that its predictions $\mathbf{p}_t(x;\tau)$ are close to the one-hot labels $\mathbf{y}(x)$ on the dataset $\mathcal{D}_l\cup\mathcal{D}_u$.

However, according to Lemma 3.2, a teacher whose outputs are closer to one-hot labels may also induce larger loss variance, enlarging the generalization gap. To better control the variance term and further achieve a more favorable bias-variance tradeoff, we then analyze the bound of $\mathbb{V}[\mathbf{p}_t(x)^\top \ell(\hat{x})]$.

Theorem 3.3 (Variance Bound)**.** *For any* t, $\mathbb{V}[\mathbf{p}_t(x)^\top \ell(\hat{x})] \le \mathbb{E}\big[\|\mathbf{p}_t(x;\tau)\|_2^2\big]\,\mathbb{E}\big[\|\ell(\hat{x})\|_2^2\big] + \mathrm{Cov}\big(\|\mathbf{p}_t(x;\tau)\|_2, \|\ell(\hat{x})\|_2\big)$*, where the equality holds only when* $R_t(f)=0$.

Remark 3.5. According to Lemma 3.1, increasing the temperature τ reduces $\mathbb{E}\big[\|\mathbf{p}_t(x;\tau)\|_2^2\big]$, which can effectively control the variance term and thus tighten the generalization bound in Theorem 3.1. Furthermore, during the optimization of the student objective $\hat{R}_t(f)=\mathbb{E}_{x\sim\mathcal{D}_u}[\mathbf{p}_t(x;\tau)^\top \ell(\hat{x})]$, the covariance term $\mathrm{Cov}\big(\|\mathbf{p}_t(x;\tau)\|_2, \|\ell(\hat{x})\|_2\big)$ exhibits a strong stage-dependent pattern rather than a globally monotonic increase. As shown in Fig. 2, the covariance rises rapidly during the early-to-middle stage, peaks at epoch 40 (~ 0.110), and then gradually decreases, while remaining substantially above its initial value at epoch 100. This stage dependence implies that a single fixed temperature need not provide the same bias–variance balance throughout the entire training trajectory. Therefore, we interpret the progressive warming of τ as a simple stage-adaptive approximation to this evolving trade-off; the concrete schedule is given in Section 3.

3.3 BayesAT Framework

Guided by the above analysis, we design **BayesAT**, which adapts the supervision produced by an imperfect teacher according to the bias–variance decomposition in Theorem 3.1. An overview of the framework is shown in Fig. 3. BayesAT operates in two phases. Phase 1: a teacher model t is trained to be reliable and then fixed. Phase 2: the student f is adversarially trained with soft labels generated by the fixed teacher.

Phase 1: Teacher Training. According to Theorem 3.2, the low-bias teacher should have low classification risk over the semi-supervised dataset. However, as the labels of the unlabeled data are unobserved, a weight-averaged model $\bar{t}$ is introduced to provide references, resulting in the Mean Teacher [21] training objective. Specifically, the objective of the teacher is defined as:

$$\mathcal{L}_{\text{teacher}} = \frac{1}{|\mathcal{D}_l|}\sum_{(x,y)\in\mathcal{D}_l}\big[-\mathbf{y}(x)^\top \log \mathbf{p}_t(x;1)\big] + \frac{\alpha}{|\mathcal{D}_u|}\sum_{x\in\mathcal{D}_u}\big[\|\mathbf{p}_t(x;1)-\mathbf{p}_{\bar{t}}(x;1)\|_2^2\big], \quad (5)$$

where $\bar{t}$ is the weight-averaged version of t, and hyper-parameter α controls the weight of consistency cost on unlabeled data. The final weight-averaged model serves as the reliable teacher.

Phase 2: Student Training. After obtaining the fixed teacher t, the student f is adversarially trained with confidence-adaptive sample reweighting and variable-temperature progressive distillation.

Confidence-adaptive Sample Reweighting. To further reduce the bias term in Theorem 3.1, we define a weight by the teacher's confidence as a practical proxy for supervision reliability to measure the sample-wise bias:

$$w(x) = \max_{i\in[C]} p_{t,i}(x;\tau=1). \quad (6)$$

As $w(x)$ reflects how close $\mathbf{p}_t(x;\tau=1)$ is to $\mathbf{y}(x)$, a training objective re-weighted by it focuses on the samples with less bias, thereby mitigating the effect of teacher bias. Its role is complementary to temperature scaling, which addresses over-concentration in the teacher distribution.

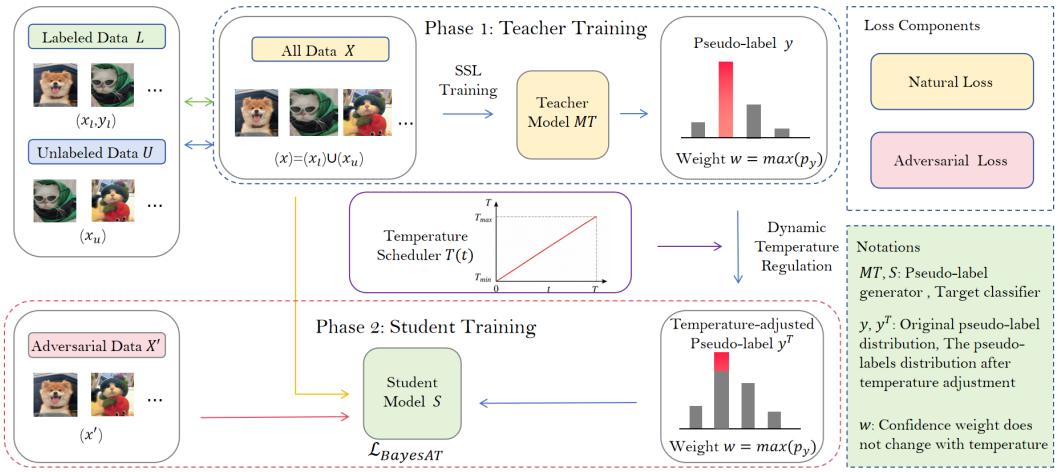


Figure 3: Overview of the BayesAT framework. Phase 1 (top) trains the teacher using labeled and unlabeled data via a Mean Teacher objective. Phase 2 (bottom) adversarially trains the student with temperature-adjusted soft pseudo-labels and confidence-adaptive reweighting. Together, these mechanisms progressively refine supervision to balance bias and variance for improved robust generalization.

Variable-temperature Progressive Distillation. Based on Theorem 3.3, a scheduler that progressively increases the temperature is designed to reduce the variance term. Let K be the total number of training epochs. For $k \in [K]$, the temperature of the k-th epoch is:

$$\tau(k) = \tau_{\min} + (\tau_{\max} - \tau_{\min}) \cdot \frac{k}{K}, \tag{7}$$

where $\tau_{\min}$ and $\tau_{\max}$ are, respectively, the pre-defined minimum and maximum temperatures. This low-to-high schedule realizes an easy-to-hard curriculum. At an early stage, a smaller τ can provide low-variance supervision with low-bias for fast boundary establishment, while later smooth signals with larger τ controlling the increasing variance and expose rich inter-class relationships.

Total Optimization. Our confidence-adaptive sample reweighting and variable-temperature progressive distillation strategies can be seamlessly plugged into existing SSAT framework, improving their robust generalization by solving the bias-variance tradeoff. Taking WSCAT [30] as the backbone, BayesAT's objective is defined as:

$$\min_f \Big\{ \frac{1}{|\mathcal{D}|} \sum_{x \in \mathcal{D}} \mathcal{L}^{(k)}_{\text{clean}}(x; t) + \frac{\lambda}{|\mathcal{D}|} \sum_{x \in \mathcal{D}} \max_{\|\hat{x} - x\|_\infty \le \epsilon} \mathcal{L}^{(k)}_{\text{robust}}(x; t) \Big\}, \tag{8}$$

where $\mathcal{L}^{(k)}_{\text{clean}}(x; t)$ and $\mathcal{L}^{(k)}_{\text{robust}}(x; t)$ are losses over clean and adversarial examples, respectively. λ is the hyper-parameter that balances standard accuracy and robustness. Let $\mathbf{p}_f(x) = \text{softmax}(f(x))$ be the predict label of f for input x. The clean loss guides f to predict the informative teacher-generated label $\mathbf{p}_t(x; \tau(k))$ (Eq. (3)) in an easy-to-hard manner (Eq. (7)), which is

$$\mathcal{L}^{(k)}_{\text{clean}}(x; t) = w(x) \cdot D_{\text{KL}}(\mathbf{p}_t(x; \tau(k)) \| \mathbf{p}_f(x)), \tag{9}$$

Let $\mathcal{N}^+_x = \{x_p | x_p \in \mathcal{D} \wedge \arg\max_y f(x_p) = \arg\max_y f(x)\}$ be the example set consists of the samples with same prediction of x given by f. The min-max problem over adversarial loss encourages f to capture the robust relations between $\hat{x}$ and $\mathbf{p}_t(x; \tau(k))$, which is defined as

$$\mathcal{L}^{(k)}_{\text{robust}}(x; t) = w(x) \cdot D_{\text{KL}}(\mathbf{p}_t(x; \tau(k)) \| \mathbf{p}_f(\hat{x})) - \bar{w} \cdot \frac{\beta}{|\mathcal{N}^+_x|} \sum_{x_p \in \mathcal{N}^+_x} \log \frac{e^{s(\hat{x}, x_p)}}{\sum_{x_n \in \mathcal{D}} e^{s(\hat{x}, x_n)}}, \tag{10}$$

where $s(\cdot, \cdot)$ is the similarity function, $\bar{w}$ is the batch average of $w(x)$ and β is a balancing coefficient. The detailed proof can be found in Appendix A and the full algorithm is provided in Appendix B. We also provide more formulation of the contrastive loss is provided in Appendix C.

4 Experimental Results

Table 1: Natural and robust accuracy (%) on CIFAR-10. Mean ± std over three seeds. Best bolded.

Method	Standard	SGLR	TRADES	MART	MT	RST	WSCAT	BayesAT
Natural	$78.04_{\pm 0.55}$	$67.33_{\pm 0.07}$	$69.13_{\pm 0.07}$	$54.48_{\pm 1.60}$	$\mathbf{86.12}_{\pm 0.48}$	$79.95_{\pm 0.71}$	$80.10_{\pm 0.67}$	$81.36_{\pm 0.13}$
FGSM	$6.55_{\pm 0.60}$	$29.47_{\pm 0.26}$	$29.72_{\pm 0.06}$	$33.60_{\pm 0.19}$	$12.60_{\pm 1.48}$	$47.95_{\pm 0.42}$	$58.02_{\pm 0.72}$	$\mathbf{61.20}_{\pm 0.21}$
PGD	$0.00_{\pm 0.00}$	$24.94_{\pm 0.05}$	$25.56_{\pm 0.06}$	$32.18_{\pm 0.18}$	$0.00_{\pm 0.00}$	$40.86_{\pm 0.08}$	$56.47_{\pm 1.07}$	$\mathbf{60.95}_{\pm 0.09}$
CW	$0.00_{\pm 0.00}$	$24.68_{\pm 0.17}$	$25.62_{\pm 0.01}$	$27.32_{\pm 0.16}$	$0.00_{\pm 0.00}$	$40.83_{\pm 0.22}$	$52.36_{\pm 0.89}$	$\mathbf{53.58}_{\pm 0.30}$
AA	$0.00_{\pm 0.00}$	$23.58_{\pm 0.15}$	$24.56_{\pm 0.09}$	$26.02_{\pm 0.29}$	$0.00_{\pm 0.00}$	$38.22_{\pm 1.27}$	$50.94_{\pm 0.68}$	$\mathbf{52.58}_{\pm 0.42}$
Mean	$16.92_{\pm 0.15}$	$34.00_{\pm 0.10}$	$34.92_{\pm 0.03}$	$34.72_{\pm 0.26}$	$19.74_{\pm 0.37}$	$49.56_{\pm 0.29}$	$59.58_{\pm 0.74}$	$\mathbf{61.94}_{\pm 0.16}$

Table 2: Natural and robust accuracy (%) on CIFAR-100. Mean ± std over three seeds. Best bolded.

Method	Standard	SGLR	TRADES	MART	MT	RST	WSCAT	BayesAT
Natural	$58.01_{\pm 0.43}$	$40.28_{\pm 0.13}$	$42.87_{\pm 0.11}$	$28.13_{\pm 0.06}$	$\mathbf{63.82}_{\pm 2.55}$	$52.82_{\pm 2.69}$	$55.27_{\pm 1.16}$	$58.41_{\pm 1.58}$
FGSM	$2.93_{\pm 0.04}$	$15.63_{\pm 0.07}$	$15.65_{\pm 0.05}$	$15.92_{\pm 0.27}$	$6.02_{\pm 1.18}$	$14.28_{\pm 1.47}$	$28.24_{\pm 0.40}$	$\mathbf{33.17}_{\pm 0.97}$
PGD	$0.01_{\pm 0.01}$	$13.32_{\pm 0.23}$	$13.17_{\pm 0.12}$	$15.48_{\pm 0.26}$	$0.04_{\pm 0.05}$	$8.44_{\pm 1.65}$	$25.28_{\pm 0.28}$	$\mathbf{30.41}_{\pm 0.67}$
CW	$0.00_{\pm 0.00}$	$12.80_{\pm 0.20}$	$12.74_{\pm 0.12}$	$12.68_{\pm 0.22}$	$0.04_{\pm 0.07}$	$7.87_{\pm 1.71}$	$23.49_{\pm 0.25}$	$\mathbf{26.34}_{\pm 0.61}$
AA	$0.00_{\pm 0.00}$	$11.83_{\pm 0.19}$	$11.80_{\pm 0.12}$	$11.99_{\pm 0.25}$	$0.00_{\pm 0.00}$	$6.19_{\pm 1.26}$	$22.44_{\pm 0.29}$	$\mathbf{24.53}_{\pm 0.57}$
Mean	$12.19_{\pm 0.09}$	$18.77_{\pm 0.11}$	$19.25_{\pm 0.06}$	$16.84_{\pm 0.18}$	$13.98_{\pm 0.72}$	$17.92_{\pm 1.36}$	$30.94_{\pm 0.45}$	$\mathbf{34.57}_{\pm 0.87}$

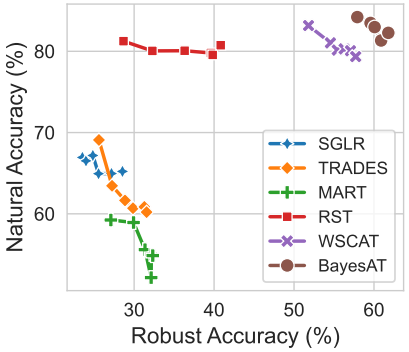


Figure 4: Natural *vs.* PGD robust accuracy on CIFAR-10 for different methods. BayesAT pushes the performance frontier further toward the top-right.

Experimental Setup. We evaluate on CIFAR-10, CIFAR-100 [10], and ImageNet-200 (a 200-class subset of ImageNet, resized to 64×64). Following WSCAT [30], we use 4k labeled and 46k unlabeled images for CIFAR-10, 10k/40k for CIFAR-100, and 10k/240k for ImageNet-200; All results are reported as mean ± std over three random seeds. Both teacher and student use WRN-28-10 [25]. The teacher is pre-trained for 100 epochs and then fixed, and the student is trained from scratch for 100 epochs. We compare against Standard [25], SGLR [13], TRADES [26], MART [22], Mean Teacher (MT) [21], RST [4], and WSCAT [30]. Models are evaluated under FGSM [6], PGD [14], , CW [3] and Auto Attack (AA) [5], with $\epsilon = 8/255$ for CIFAR and $4/255$ for ImageNet-200; we report natural, FGSM, PGD, CW, AA and their arithmetic mean. Hyperparameters are set as follows: $\tau_{\min} = 1.0$, $\tau_{\max} = 4.0$ (CIFAR-10) or 5.0(CIFAR-100) or 3.0 (ImageNet-200), linearly warmed over 100 epochs. We use SGD with Nesterov momentum, weight decay 5×10^{-4}, initial learning rate 0.1 (cosine annealing), and batch size 128. The loss weight $\lambda = 5$ for CIFAR-10 and 1 otherwise; $\beta = 0.05$. Additional implementation details and hyperparameter settings are provided in Appendix E.

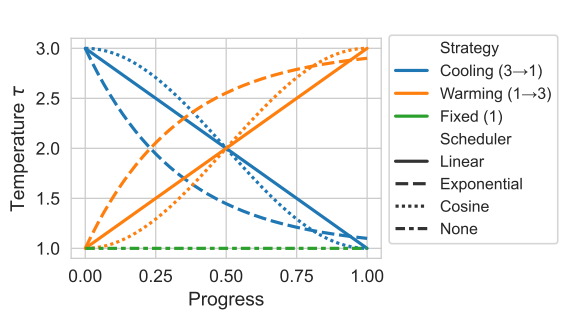
4.1 Main Results

Tables 1, 2, and 3 summarize the comparisons. Across datasets, existing SSAT methods typically lie on a constrained accuracy–robustness frontier, where further gains in robustness often come at the expense of natural accuracy. This behavior is consistent with the limitations of static supervision discussed earlier.

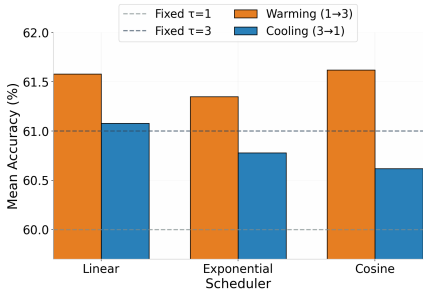
On CIFAR-10 (Table 1), strong baselines such as WSCAT already achieve a favorable balance, yet improving robustness beyond this point remains challenging. Similar trends appear on CIFAR-100 (Table 2) and ImageNet-200 (Table 3), where the trade-off becomes more pronounced as task complexity increases. In contrast, BayesAT adopts a stage-adaptive supervision strategy that dynamically adjusts temperature and sample weighting. As a result, it achieves 81.36% natural accuracy and 60.95% PGD on CIFAR-10, outperforming WSCAT by 3.83 points in robustness. On CIFAR-100, BayesAT improves PGD accuracy by 5.13 points over WSCAT (30.41% vs. 25.28%) with competitive natural accuracy (58.41%). On ImageNet-200, it reaches 31.24% natural and 14.42% PGD accuracy; the natural-accuracy difference is within seed variability, while robustness shows higher reported average improvements.

Table 3: Natural and robust accuracy (%) on ImageNet-200. Mean ± std over three seeds. Best bolded.

Method	Standard	SGLR	TRADES	MART	MT	RST	WSCAT	BayesAT
Natural	$31.05_{\pm 0.23}$	$11.31_{\pm 0.05}$	$14.97_{\pm 0.09}$	$8.28_{\pm 0.05}$	$\mathbf{33.95}_{\pm 6.29}$	$23.15_{\pm 4.50}$	$30.38_{\pm 4.21}$	$31.24_{\pm 4.19}$
FGSM	$0.19_{\pm 0.03}$	$1.88_{\pm 0.11}$	$2.19_{\pm 0.10}$	$3.90_{\pm 0.05}$	$0.51_{\pm 0.01}$	$0.66_{\pm 0.11}$	$11.23_{\pm 2.08}$	$\mathbf{15.52}_{\pm 2.07}$
PGD	$0.00_{\pm 0.00}$	$1.52_{\pm 0.05}$	$1.71_{\pm 0.06}$	$3.80_{\pm 0.04}$	$0.05_{\pm 0.05}$	$0.25_{\pm 0.04}$	$9.24_{\pm 1.59}$	$\mathbf{14.42}_{\pm 1.80}$
CW	$0.00_{\pm 0.00}$	$1.27_{\pm 0.02}$	$1.39_{\pm 0.04}$	$2.63_{\pm 0.08}$	$0.06_{\pm 0.06}$	$0.13_{\pm 0.08}$	$7.58_{\pm 1.28}$	$\mathbf{10.82}_{\pm 1.53}$
AA	$0.00_{\pm 0.00}$	$1.10_{\pm 0.02}$	$1.21_{\pm 0.04}$	$2.40_{\pm 0.12}$	$0.00_{\pm 0.00}$	$0.07_{\pm 0.06}$	$6.79_{\pm 1.19}$	$\mathbf{9.61}_{\pm 1.21}$
Mean	$6.25_{\pm 0.04}$	$3.42_{\pm 0.04}$	$4.29_{\pm 0.03}$	$4.20_{\pm 0.05}$	$6.91_{\pm 1.24}$	$4.85_{\pm 0.88}$	$13.04_{\pm 2.07}$	$\mathbf{16.32}_{\pm 2.17}$



(a) Different temperature schedules.



(b) Mean accuracy of the student for each schedule.

Figure 5: Analysis of temperature schedule design. (a) Warming, cooling, and non-linear schedules; (b) corresponding mean accuracy (%) across natural and robust.

Fig. 4 illustrates this effect: prior methods cluster along a limited frontier, while BayesAT shifts the frontier toward a more favorable accuracy–robustness trade-off.

Across different attacks, BayesAT consistently improves robustness under both weak and strong settings, indicating that adaptive supervision better captures the evolving training dynamics.

4.2 Temperature Schedule Ablation

To validate the effect of temperature scheduling, we evaluate a wide range of strategies within the same framework (soft labels + confidence weighting), varying only the schedule. Table 4 summarizes the results.As illustrated in Fig. 5a, different schedules induce distinct supervision dynamics. Fixed temperatures provide consistent supervision but do not adapt to the evolving student. Cooling schedules start with high-entropy labels and gradually sharpen them, which may hinder early-stage boundary formation due to overly smooth supervision. In contrast, warming schedules progressively increase the temperature, enabling a transition from sharp to smooth supervision.

The quantitative results reflect these differences. Fixed temperature strategies ($\tau = 1, 3, 4$) achieve comparable performance but remain within a limited range, suggesting restricted flexibility. Cooling schedules (linear, exponential, cosine decay) also cluster in a relatively narrow band of mean accuracy ($60.62\% - 61.10\%$), with consistently weaker PGD performance, indicating that decreasing temperature over time may not align well with later-stage training dynamics.

Warming schedules, on the other hand, exhibit stronger overall performance. The reverse linear strategy ($1 \to 4$) achieves the highest mean accuracy (61.88%) and best PGD performance (60.89%), while other variants (exponential and cosine) remain competitive but slightly lower. Reverse schedules ($1 \to 3, 1 \to 4$) show strong performance under weaker attacks such as FGSM, but their gains are less consistent under stronger attacks, suggesting that the timing of label smoothing plays a critical role.These observations are consistent with the analysis in Section 3: as training progresses, controlling the variance term becomes increasingly important, which favors larger temperatures in later stages. Fig. 5b further illustrates that warming schedules consistently lead to higher mean accuracy across different attacks. We provide further discussion on the design of temperature scheduling and its comparison with alternative strategies in Appendix D.

Table 4: Performance (%) of different temperature schedules on CIFAR-10 (best seed).

Strategy	Range	Natural	FGSM	PGD	CW	AA	Mean
Fixed Low τ	$\tau = 1$	80.07	58.72	57.06	52.66	51.67	60.04
Fixed Mid τ	$\tau = 3$	80.46	60.10	59.89	52.77	51.82	61.01
Fixed High τ	$\tau = 4$	79.75	59.80	60.29	51.51	50.66	60.40
Reverse Linear	$1 \to 4$	**81.33**	**61.40**	**60.89**	53.29	52.47	**61.88**
Reverse Linear	$1 \to 3$	81.26	60.97	60.07	53.65	51.83	61.56
Reverse Exponential	$1 \to 3$	81.82	60.48	59.50	53.16	51.84	61.36
Reverse Cosine	$1 \to 4$	79.50	60.06	59.61	52.37	51.51	60.61
Reverse Cosine	$1 \to 3$	80.82	61.04	60.01	**53.66**	**52.50**	61.61
Linear Decay	$5 \to 1$	81.13	59.68	58.55	53.24	51.99	60.92
Linear Decay	$4 \to 1$	81.93	59.89	57.90	53.12	51.74	61.10
Linear Decay	$3 \to 1$	81.70	60.05	57.87	53.54	52.35	61.10
Exponential Decay	$3 \to 1$	81.73	59.37	57.48	53.34	51.99	60.78
Cosine Decay	$3 \to 1$	81.63	59.33	57.50	53.00	51.62	60.62

Table 5: Component ablation on CIFAR-10 (best seed). Natural, FGSM, PGD, CW, AA accuracy and mean (%). Best in each column bolded.

Configuration	Natural	FGSM	PGD	CW	AA	Mean
Base Model (BM)	80.07	58.72	57.06	52.66	51.67	60.04
+ Soft Labels (SL)	80.95	59.72	58.01	53.09	51.96	60.75
+ Confidence Weighting (CW)	**81.41**	59.54	57.34	53.16	52.18	60.73
+ Temperature Warming (TW)	81.33	**61.40**	**60.89**	**53.29**	**52.47**	**61.88**

Overall, the results suggest that adapting temperature during training provides a more effective supervision strategy than fixed or decreasing schedules, particularly for improving robustness under stronger adversaries.

4.3 Component Analysis

We perform an ablation study to evaluate each BayesAT component: soft labels (SL), confidence weighting (CW), and temperature warming (TW). Starting from the base model (BM), we add these modules incrementally (Table 5).

Soft labels improve natural accuracy from 80.07% to 80.95% and slightly enhance adversarial robustness, reflecting variance reduction via smoother targets. Confidence weighting further increases natural accuracy to 81.41% and boosts robustness on challenging attacks (CW, AA) by down-weighting ambiguous supervision. Temperature warming provides the largest gains, notably raising FGSM and PGD accuracies to 61.40% and 60.89%, achieving the best mean robustness of 61.88%.

These two components operate along complementary axes. Temperature scheduling differentiates supervision across training stages by globally smoothing the teacher distribution, whereas confidence weighting differentiates samples within the same stage by adjusting their individual contributions. Together, they provide a unified mechanism to dynamically balance the bias-variance trade-off.

4.4 Teacher Quality Sensitivity

To evaluate sensitivity to supervision quality, we compare three teachers on CIFAR-10: (1) *Standard*: trained only on labeled data; (2) *Mean Teacher (MT)*: our default; and (3) *Ground-Truth (GT)*: direct one-hot labels. AutoAttack is omitted due to compute limits, and the mean is over four metrics.

While the weaker Standard Teacher reduces the mean from 64.27 to 61.72, MT performs comparably to the GT Teacher (64.27 vs. 64.34). This shows BayesAT has controlled sensitivity to teacher quality, and the practical MT approaches fully correct hard-label supervision.

Table 6: Teacher sensitivity on CIFAR-10 (mean $\pm$ std over 3 seeds).

Metric	Standard Teacher	Mean Teacher	GT Teacher
Natural	79.09 ± 0.20	81.36 ± 0.10	**81.93** ± 0.42
FGSM	58.31 ± 0.19	**61.20** ± 0.17	61.19 ± 0.38
PGD	58.38 ± 0.45	60.95 ± 0.07	**60.97** ± 0.34
CW	51.10 ± 0.06	**53.58** ± 0.24	53.28 ± 0.25
Mean (4 metrics)	61.72 ± 0.18	64.27 ± 0.08	**64.34** ± 0.03

5 Conclusion

We identified that static supervisory signals in semi-supervised adversarial training induce a bias–variance imbalance that limits knowledge transfer from a plateaued teacher. By reframing SSAT as distillation relative to a Bayes reference, we derived a robust generalization bound revealing a temperature-governed bias–variance tradeoff: low temperatures reduce approximation bias, whereas higher temperatures control the variance term, and the preferred balance is stage-dependent because the variance-related quantities depend on the evolving student. Guided by this analysis, we proposed BayesAT, which employs a low-to-high temperature warming schedule together with confidence-adaptive sample reweighting. These two mechanisms dynamically balance bias and variance—across training stages and across samples, respectively—without requiring explicit difficulty metrics. Experiments on CIFAR-10, CIFAR-100, and ImageNet-200 demonstrate consistent improvements in robust accuracy over existing SSAT methods, supporting the value of stage-adaptive, theory-driven supervision for effective knowledge transfer.

Limitations and Future Work. BayesAT inherits a two-phase pipeline that requires a separately pre-trained teacher, which may limit efficiency in resource-constrained settings. The linear warming schedule, though effective, relies on manually chosen temperature bounds ($\tau_{\min}$, $\tau_{\max}$); these may need tuning for new datasets or architectures. Our study focuses on ℓ_∞ perturbations for image classification; extending the bias–variance temperature framework to other threat models and modalities remains open. In addition, BayesAT relies on confidence-based supervision adjustment, which down-weights ambiguous pseudo-labels but cannot identify confidently incorrect teacher predictions, and temperature warming only redistributes the probability mass already present in the teacher output without correcting an incorrect teacher argmax. The prediction-based positive sets used by the contrastive term may likewise contain noisy samples in early training; their dynamic recomputation and the small coefficient β mitigate, but do not fully eliminate, this noise. Future work could pursue adaptive temperature schedules that automatically adjust warming speed based on online estimates of the bias–variance trade-off, as well as end-to-end joint training of teacher and student to further reduce annotation dependency. We include a discussion on societal impacts, limitations, and potential risks in Appendix F.

Acknowledgements

This work was supported by the National Key R&D Program of China(2024YFB3312503), the Science and Technology Major Project of Sichuan Province(2024ZDZX0003, 2025ZDZX0140) and the Fundamental Research Funds for the Central Universities(2026SCUQJTX001).

References

[1] Tao Bai, Jinqi Luo, Jun Zhao, Bihan Wen, and Qian Wang. Recent advances in adversarial training for adversarial robustness. In *International Joint Conference on Artificial Intelligence (IJCAI)*, 2021.

[2] Battista Biggio, Igino Corona, Davide Maiorca, Blaine Nelson, Nedim Srndić, Pavel Laskov, Giorgio Giacinto, and Fabio Roli. Evasion attacks against machine learning at test time. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases (ECML PKDD)*, pages 387–402. Springer, 2013.

[3] Nicholas Carlini and David Wagner. Towards evaluating the robustness of neural networks. In *IEEE Symposium on Security and Privacy (S&P)*, pages 39–57, 2017.

[4] Yair Carmon, Aditi Raghunathan, Ludwig Schmidt, John C. Duchi, and Percy S. Liang. Unlabeled data improves adversarial robustness. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 32, 2019.

[5] Francesco Croce and Matthias Hein. Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks. In *International Conference on Machine Learning (ICML)*, pages 2206–2216. PMLR, 2020.

[6] Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In *International Conference on Learning Representations (ICLR)*, 2015.

[7] Haorui He, Yuanzhe Ren, Zheng Li, and Jing Xue. Adaptive knowledge distillation for efficient relation classification. In Elias Pimenidis, Plamen Angelov, Chrisina Jayne, Antonios Papaleonidas, and Mehmet Aydin, editors, *International Conference on Artificial Neural Networks (ICANN)*, volume 13532 of *Lecture Notes in Computer Science*, pages 148–158, Cham, 2022. Springer Nature Switzerland.

[8] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. In *NIPS 2014 Deep Learning and Representation Learning Workshop*, 2015.

[9] Andrew Ilyas, Shibani Santurkar, Dimitris Tsipras, Logan Engstrom, Brandon Tran, and Aleksander Madry. Adversarial examples are not bugs, they are features. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 32, 2019.

[10] Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009.

[11] Anna Kukleva, Moritz Böhle, Bernt Schiele, Hilde Kuehne, and Christian Rupprecht. Temperature schedules for self-supervised contrastive methods on long-tail data. In *International Conference on Learning Representations (ICLR)*, 2023.

[12] Zheng Li, Xiang Li, Lingfeng Yang, Borui Zhao, Renjie Song, Lei Luo, Jun Li, and Jian Yang. Curriculum temperature for knowledge distillation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 1504–1512, 2023.

[13] Zhuorong Li, Daiwei Yu, Lina Wei, Canghong Jin, Yun Zhang, and Sixian Chan. Soften to defend: Towards adversarial robustness via self-guided label refinement. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 24776–24785, 2024.

[14] 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 (ICLR)*, 2018.

[15] Aditya Krishna Menon, Ankit Singh Rawat, Sashank J. Reddi, Seungyeon Kim, and Sanjiv Kumar. A statistical perspective on distillation. In *International Conference on Machine Learning (ICML)*, pages 7632–7642, 2021.

[16] David J. Miller, Zhen Xiang, and George Kesidis. Adversarial learning targeting deep neural network classification: A comprehensive review of defenses against attacks. *Proceedings of the IEEE*, 108(3):402–433, 2020.

[17] Seyed Iman Mirzadeh, Mehrdad Farajtabar, Ang Li, Nir Levine, Akihiro Matsukawa, and Hassan Ghasemzadeh. Improved knowledge distillation via teacher assistant. In *AAAI Conference on Artificial Intelligence (AAAI)*, volume 34, pages 5191–5198, 2020.

[18] Rafael Müller, Simon Kornblith, and Geoffrey Hinton. When does label smoothing help? In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 32, 2019.

[19] 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. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, pages 596–608, 2020.

[20] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. In *International Conference on Learning Representations (ICLR)*, 2014.

[21] Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 30, 2017.

[22] Yisen Wang, Difan Zou, Jinfeng Yi, James Bailey, Xingjun Ma, and Quanquan Gu. Improving adversarial robustness requires revisiting misclassified examples. In *International Conference on Learning Representations (ICLR)*, 2020.

[23] Qizhe Xie, Zihang Dai, Eduard Hovy, Minh-Thang Luong, and Quoc V. Le. Unsupervised data augmentation for consistency training. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, pages 6256–6268, 2020.

[24] Li Yuan, Francis E.H. Tay, Guilin Li, Tao Wang, and Jiashi Feng. Revisiting knowledge distillation via label smoothing regularization. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3903–3911, 2020.

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

[26] Hongyang Zhang, Yaodong Yu, Jiantao Jiao, Eric P. Xing, Laurent El Ghaoui, and Michael I. Jordan. Theoretically principled trade-off between robustness and accuracy. In *International Conference on Machine Learning (ICML)*, pages 7472–7482, 2019.

[27] Jingfeng Zhang, Xilie Xu, Bo Han, Gang Niu, Lizhen Cui, Massashi Sugiyama, and Mohan Kankanhalli. Geometry-aware instance-reweighted adversarial training. In *International Conference on Learning Representations (ICLR)*, 2021.

[28] Lilin Zhang, Yimo Guo, Yue Li, Jiancheng Shi, and Xianggen Liu. Taming the long tail: Rebalancing adversarial training via adaptive perturbation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 34900–34907, June 2026.

[29] Lilin Zhang, Yue Li, Jiancheng Shi, Jiancheng Lv, and Xianggen Liu. Posterior mismatch matters: Adversarial training for long-tailed robustness. In *International Conference on Machine Learning (ICML)*, pages 160265–160280. PMLR, 2026.

[30] Lilin Zhang, Chengpei Wu, and Ning Yang. Weakly supervised contrastive adversarial training for learning robust features from semi-supervised data. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 25718–25727, 2025.

[31] Weimin Zhao, Sanaa Alwidian, and Qusay H. Mahmoud. Adversarial training methods for deep learning: A systematic review. *Algorithms*, 15(8):283, 2022.

A Detailed Proofs

A.1 Proof of Lemma 3.1

Proof. By the definition of l_2 norm,

$$\nabla_\tau \|\mathbf{p}_t(x;\tau)\|_2^2 = \nabla_\tau \sum_{i\in[C]} \big(\mathbf{p}_{t,i}(x;\tau)\big)^2 = \sum_{i\in[C]} 2\,\mathbf{p}_{t,i}(x;\tau)\cdot\nabla_\tau \mathbf{p}_{t,i}(x;\tau).$$

As $\mathbf{p}_t(x;\tau) = \text{softmax}\Big(\frac{z_t(x)}{\tau}\Big) = \left[\frac{\exp(z_{t,i}(x)/\tau)}{\sum_{j\in[C]}\exp(z_{t,j}(x)/\tau)}\right]_{i\in[C]}$ (Eq. (3)), the gradient

$$\nabla_\tau \mathbf{p}_{t,i}(x;\tau) = \frac{-\tau^{-2}\sum_{j\in[C]}\big\{\exp\big((z_{t,i}(x)+z_{t,j}(x))/\tau\big)\big(z_{t,i}(x)+z_{t,j}(x)\big)\big\}}{\big(\sum_{j\in[C]}\exp(z_{t,j}(x)/\tau)\big)^2}.$$

Since for any $x\sim\mathbb{P}$, $i\in[C]$ and $j\in[C]$, $z_{t,i}(x)+z_{t,j}(x)\geq 0$ holds, $\nabla_\tau \mathbf{p}_{t,i}(x;\tau)\leq 0$. Therefore,

$$\nabla_\tau \|\mathbf{p}_t(x;\tau)\|_2^2 \leq 0.$$

□

A.2 Proof of Lemma 3.2

Proof. By the definition of variance,

$$\begin{aligned}
\mathbb{V}[\mathbf{p}^*(x)^\top \ell(\hat{x})] &= \mathbb{E}_{x\sim\mathbb{P}}\big[\big(\mathbf{p}^*(x)^\top\ell(\hat{x})\big)^2\big] - \big(\mathbb{E}_{x\sim\mathbb{P}}\big[\mathbf{p}^*(x)^\top\ell(\hat{x})\big]\big)^2 \\
&= \mathbb{E}_{x\sim\mathbb{P}}\big[\big(\mathbb{E}_{y\sim\mathbb{P}}[\mathbf{y}(x)^\top\ell(\hat{x})]\big)^2\big] - \big(\mathbb{E}_{x\sim\mathbb{P}}\big[\mathbb{E}_{y\sim\mathbb{P}}[\mathbf{y}(x)^\top\ell(\hat{x})]\big]\big)^2 \\
&= \mathbb{E}_{x\sim\mathbb{P}}\big[\big(\mathbb{E}_{y\sim\mathbb{P}}[\mathbf{y}(x)^\top\ell(\hat{x})]\big)^2\big] - \big(\mathbb{E}_{(x,y)\sim\mathbb{P}}\big[\mathbf{y}(x)^\top\ell(\hat{x})\big]\big)^2,
\end{aligned}$$

and

$$\begin{aligned}
\mathbb{V}[\mathbf{y}(x)^\top \ell(\hat{x})] &= \mathbb{E}_{(x,y)\sim\mathbb{P}}\big[\big(\mathbf{y}(x)^\top\ell(\hat{x})\big)^2\big] - \big(\mathbb{E}_{(x,y)\sim\mathbb{P}}[\mathbf{y}(x)^\top\ell(\hat{x})]\big)^2 \\
&= \mathbb{E}_{x\sim\mathbb{P}}\big[\mathbb{E}_{y\sim\mathbb{P}}[\big(\mathbf{y}(x)^\top\ell(\hat{x})\big)^2]\big] - \big(\mathbb{E}_{(x,y)\sim\mathbb{P}}[\mathbf{y}(x)^\top\ell(\hat{x})]\big)^2.
\end{aligned}$$

Then, we get the difference between the above two as:

$$\begin{aligned}
\mathbb{V}[\mathbf{p}^*(x)^\top \ell(\hat{x})] - \mathbb{V}[\mathbf{y}(x)^\top \ell(\hat{x})] &= \mathbb{E}_{x\sim\mathbb{P}}\big[\big(\mathbb{E}_{y\sim\mathbb{P}}[\mathbf{y}(x)^\top\ell(\hat{x})]\big)^2 - \mathbb{E}_{y\sim\mathbb{P}}\big[\big(\mathbf{y}(x)^\top\ell(\hat{x})\big)^2\big]\big] \\
&= -\mathbb{E}_{x\sim\mathbb{P}}\big[\mathbb{V}[\mathbf{y}(x)^\top\ell(\hat{x})]\big] \leq 0.
\end{aligned}$$

Therefore, $\mathbb{V}[\mathbf{p}^*(x)^\top \ell(\hat{x})] \leq \mathbb{V}[\mathbf{y}(x)^\top \ell(\hat{x})]$ and the equality holds when $\mathbf{y}(x)^\top\ell(\hat{x})$ is a constant over $\mathbb{P}$. □

A.3 Proof of Theorem 3.1

Proof. According to Bennett's inequality, it holds with a probability of at least $1-\delta$ that

$$\begin{aligned}
R_t(f) - \hat{R}_t(f) =& \mathbb{E}_{(x,y)\sim\mathbb{P}}\big[\mathbf{p}_t(x)^\top\ell(\hat{x})\big] - \mathbb{E}_{(x,y)\in\mathcal{D}}\big[\mathbf{p}_t(x)^\top\ell(\hat{x})\big] \\
\leq& \sqrt{\frac{2\mathbb{V}[\mathbf{p}_t(x;\tau)^\top\ell(\hat{x})]\log(1/\delta)}{N}} + \frac{\log(1/\delta)}{3N}.
\end{aligned}$$

Therefore,

$$\begin{aligned}
R(f) - \hat{R}_t(f) =& R(f) - R_t(f) + R_t(f) - \hat{R}_t(f) \\
=& \mathbb{E}_{(x,y)\sim\mathbb{P}}\big[\mathbf{p}^*(x)^\top\ell(\hat{x})\big] - \mathbb{E}_{(x,y)\sim\mathbb{P}}\big[\mathbf{p}_t(x)^\top\ell(\hat{x})\big] + R_t(f) - \hat{R}_t(f) \\
\leq& \mathbb{E}_{(x,y)\sim\mathbb{P}}\big[\big(\mathbf{p}^*(x) - \mathbf{p}_t(x;\tau)\big)^\top\ell(\hat{x})\big] + \sqrt{\frac{2\mathbb{V}[\mathbf{p}_t(x;\tau)^\top\ell(\hat{x})]\log(1/\delta)}{N}} + \frac{\log(1/\delta)}{3N}
\end{aligned}$$

holds with a probability of at least $1-\delta$. □

A.4 Proof of Theorem 3.2

Proof. Since the one-hot label is an unbiased estimator of the Bayes class probability, i.e., $\mathbb{E}[\mathbf{y}(x)] = \mathbf{p}^*(x)$, we can get that

$$\begin{aligned}\mathbb{E}_{x\sim\mathbb{P}}\big[\big(\mathbf{p}^*(x)-\mathbf{p}_t(x;\tau)\big)^\top \ell(\hat{x})\big] &= \mathbb{E}_{(x,y)\sim\mathbb{P}}\big[\big(\mathbf{y}(x)-\mathbf{p}_t(x;\tau)\big)^\top \ell(\hat{x})\big] \\ &\le \mathbb{E}_{(x,y)\in\mathcal{D}_l\cup\mathcal{D}_u}\big[\big(\mathbf{y}(x)-\mathbf{p}_t(x;\tau)\big)^\top \ell(\hat{x})\big] + L\sqrt{\frac{\log(1/\delta)}{2(N_l+N_u)}},\end{aligned}$$

where the last line holds for at least $1-\delta$ probability according to the Hoeffding's Inequality. □

A.5 Proof of Theorem 3.3

Proof. According to the Cosine Theorem,

$$\mathbf{p}_t(x;\tau)^\top \ell(\hat{x}) = \|\mathbf{p}_t(x;\tau)\|_2 \cdot \|\ell(\hat{x})\|_2 \cdot \cos\theta_{\mathbf{p}_t(x;\tau),\ell(\hat{x})} \le \|\mathbf{p}_t(x;\tau)\|_2 \cdot \|\ell(\hat{x})\|_2,$$

where $\theta_{\mathbf{p}_t(x;\tau),\ell(\hat{x})}$ represents the angle between $\mathbf{p}_t(x;\tau)$ and $\ell(\hat{x})$. By the definition of variance,

$$\begin{aligned}\mathbb{V}[\mathbf{p}_t(x;\tau)^\top \ell(\hat{x})] &= \mathbb{E}\big[\big(\mathbf{p}_t(x;\tau)^\top \ell(\hat{x})\big)^2\big] - \big(\mathbb{E}[\mathbf{p}_t(x;\tau)^\top \ell(\hat{x})]\big)^2 \\ &\le \mathbb{E}\big[\|\mathbf{p}_t(x;\tau)\|_2^2 \cdot \|\ell(\hat{x})\|_2^2\big] \\ &= \mathbb{E}\big[\|\mathbf{p}_t(x;\tau)\|_2^2\big]\,\mathbb{E}\big[\|\ell(\hat{x})\|_2^2\big] + \mathrm{Cov}\big(\|\mathbf{p}_t(x;\tau)\|_2, \|\ell(\hat{x})\|_2\big),\end{aligned}$$

where the equality holds only when $R_t(f) = \mathbb{E}[\mathbf{p}_t(x;\tau)^\top \ell(\hat{x})] = 0$. □

B Detailed Algorithm

Algorithm 1 provides the complete pseudo-code for BayesAT training.

Algorithm 1 BayesAT

Require: Labeled data $\mathcal{D}_l$, unlabeled data $\mathcal{D}_u$, teacher T, student S, temperature bounds $\tau_{\min}$, $\tau_{\max}$, loss weights λ, β, epochs E, batch size B
Ensure: Trained student model S

// Phase 1: Teacher pre-training
Train teacher T on $\mathcal{D}_l$ and $\mathcal{D}_u$ for 100 epochs
// Phase 2: Student training with BayesAT
for $e = 1$ **to** E **do**
  $\tau \leftarrow \tau_{\min} + (\tau_{\max} - \tau_{\min}) \cdot e/E$ ▷ Linear warming
  for each batch $\mathcal{B} \subset \mathcal{D}_l \cup \mathcal{D}_u$ of size B **do**
    Compute teacher logits $z_T(x)$ for all $x \in \mathcal{B}$
    Compute soft labels: $p_T(x;\tau) \leftarrow \mathrm{softmax}(z_T(x)/\tau)$
    Compute raw confidence weights: $w(x) \leftarrow \max_i p_T^{(i)}(x;\tau=1)$
    $\bar{w} \leftarrow \frac{1}{B}\sum_{x\in\mathcal{B}} w(x)$
    // Generate adversarial examples
    Initialize $x' \leftarrow x + \mathrm{uniform}(-\epsilon, \epsilon)$
    for $k = 1$ **to** 10 **do**
      $g \leftarrow \nabla_{x'}[D_{\mathrm{KL}}(p_S(x')\|p_T(x;\tau)) + \beta \cdot \ell_{\mathrm{con}}(z', z)]$
      $x' \leftarrow \Pi_{\mathcal{B}_\epsilon(x)}[x' + \alpha \cdot \mathrm{sign}(g)]$
    end for
    // Compute losses
    $\mathcal{L}_{\mathrm{nat}} \leftarrow \frac{1}{B}\sum_x w(x) D_{\mathrm{KL}}(p_S(x)\|p_T(x;\tau))$
    $\mathcal{L}_{\mathrm{rob}} \leftarrow \frac{1}{B}\sum_x w(x) D_{\mathrm{KL}}(p_S(x')\|p_T(x;\tau))$
    $\mathcal{L}_{\mathrm{total}} \leftarrow \mathcal{L}_{\mathrm{nat}} + \lambda\mathcal{L}_{\mathrm{rob}} + \beta \cdot \bar{w} \cdot \ell_{\mathrm{con}}(z, z')$
    Update S using $\nabla_S \mathcal{L}_{\mathrm{total}}$
  end for
end for
return S

C Contrastive Loss Configuration

BayesAT incorporates a weakly supervised contrastive loss ℓ_{con} in both adversarial example generation and the total training objective. This loss is adapted from WSCAT [30] and serves to encourage complete perturbation of non-robust features by maximizing the distributional discrepancy between natural and adversarial embeddings.

Given a natural sample x and its adversarial counterpart x', with corresponding feature embeddings $z = f(x)$ and $z' = f(x')$, the contrastive loss is defined as:

$$\ell_{\text{con}}(z', z) = -\frac{1}{|\mathcal{N}_x^+|} \sum_{x_p \in \mathcal{N}_x^+} \log \frac{\exp(s(z', z_p))}{\sum_{x_n \in \mathcal{B}} \exp(s(z', z_n))}, \tag{11}$$

where $s(\cdot, \cdot)$ denotes the cosine similarity function, $\mathcal{B}$ is the current training batch, and $\mathcal{N}_x^+$ is the positive sample set defined as:

$$\mathcal{N}_x^+ = \{x_p \mid x_p \in \mathcal{B} \ \wedge \ \arg\max_y C(x_p) = \arg\max_y C(x)\}. \tag{12}$$

Here, $C = g \circ f$ denotes the target classifier (student). The positive set $\mathcal{N}_x^+$ consists of samples in the current batch that share the same predicted class as x, where the predictions are produced by the current student model. This design yields three key properties:

Weak Supervision. The positive set relies solely on model predictions rather than ground-truth labels, making it applicable to both labeled and unlabeled data without requiring pseudo-label pre-generation.

Dynamic Adaptation. As the student model C is continuously updated during training, the positive sets evolve accordingly. This dynamic selection reflects the current model's understanding of class structures and progressively captures more accurate feature-level correlations.

Correlation Blocking. Maximizing $\ell_{\text{con}}(z', z)$ during AE generation encourages z' to become dissimilar to the embeddings of samples predicted as the same class. This disrupts correlations between non-robust features and labels, promoting more complete adversarial perturbation.

The contrastive temperature κ is set to 0.07 following standard practice. For computational efficiency, the positive set and the summation in ℓ_{con} are computed over the current batch of size 128 rather than the entire dataset. This batched approximation provides a sufficient signal for correlation blocking while maintaining training efficiency.

In the complete AE generation (inner maximization), the contrastive loss is combined with the KL divergence:

$$x' = \arg\max_{x' \in \mathcal{B}_\epsilon(x)} \Big\{ D_{\text{KL}}(p_S(x') \| p_T(x; \tau(t))) + \beta \cdot \ell_{\text{con}}(z', z) \Big\}, \tag{13}$$

where $\beta = 0.05$ controls the weight of the contrastive term. In the total training objective (outer minimization), the contrastive loss appears as an additional regularization weighted by the batch-averaged confidence $\bar{w}$:

$$\mathcal{L}_{\text{total}} = \frac{1}{|\mathcal{D}|} \sum_{x \in \mathcal{D}} w(x) \Big[D_{\text{KL}}(p_S(x) \| p_T(x; \tau(t))) + \lambda \cdot D_{\text{KL}}(p_S(x') \| p_T(x; \tau(t))) \Big] + \beta \cdot \bar{w} \cdot \ell_{\text{con}}(z, z'). \tag{14}$$

This dual role of the contrastive loss—promoting complete perturbation in AE generation while regularizing feature learning in defense—is central to BayesAT's ability to learn robust features from semi-supervised data.

Reliability of Prediction-based Positive Sets. The contrastive term primarily guides adversarial example generation by encouraging the adversarial representation to deviate from features associated with the current predicted class. If early predictions are noisy, the positive set $\mathcal{N}_x^+$ may contain

incorrect samples, potentially introducing noisy semantic structure. This effect is mitigated by three factors: (1) *Dynamic grouping*: positive sets are recomputed from the current predictions throughout training, so early errors are not permanently fixed; (2) *Controlled regularization strength*: the outer contrastive term is weighted by $\beta\bar{w}$ with $\beta = 0.05$, acting as an auxiliary regularizer rather than the dominant training signal; (3) *Teacher-guided primary supervision*: the main clean and adversarial distillation losses remain supervised by the pretrained teacher. The early prediction noise cannot be fully eliminated theoretically, but these constraints prevent persistent error reinforcement.

D Discussion on Temperature Schedule Design

The choice between warming and cooling depends critically on the teacher's state and the learning context. In standard knowledge distillation (KD), the teacher is typically a large, well-trained model with rich dark knowledge embedded in its softmax outputs. The student, being a compact model, needs to capture this structural information early in training. This motivates the conventional "high→low" cooling schedule: starting with a high temperature reveals inter-class similarities, while later reducing the temperature sharpens decision boundaries.

In contrast, the SSAT scenario presents a fundamentally different dynamic. The teacher model has already plateaued; its outputs are over-confident and sharp. Providing high-temperature (smooth) labels early would be wasteful, as the student lacks the basic decision boundaries to interpret them. Instead, the student benefits from a "low→high" warming schedule: early sharp labels establish robust boundaries, and later smooth labels enable fine-grained exploration of inter-class relationships. Our experiments confirm that warming consistently outperforms cooling across all datasets, validating this hypothesis. We recommend warming schedules for any semi-supervised setting where the teacher's outputs are inherently sharp or plateaued.

Comparison with the oscillating schedule of Kukleva et al. Kukleva et al. [11] apply temperature to the InfoNCE contrastive loss, where the cosine oscillation between low and high temperatures encourages alternating between instance-wise and group-wise discrimination. In contrast, our temperature is applied to teacher-generated class-probability targets in semi-supervised adversarial training. The underlying optimization dynamics are fundamentally different: SSAT requires a directional transition from sharp supervision (for initial boundary establishment) to smooth supervision (for variance control), rather than a periodic oscillation.

To empirically validate this distinction, we implemented a cosine schedule oscillating between $\tau = 1$ and $\tau = 4$ with one complete cycle every 20 epochs. As shown in Table 7, the oscillating schedule is consistently weaker than our directional linear warming across all evaluated metrics. This suggests that, within the tested configurations, a monotonic warming trajectory is better aligned with SSAT training dynamics than periodic oscillation.

Table 7: Comparison with the oscillating temperature schedule on CIFAR-10 (best seed).

Strategy	Range	Period	Natural	FGSM	PGD	CW	Mean
Cosine Oscillation	1↔4	20 epochs	78.93	57.97	56.96	51.52	61.34
Linear Warming	1→4	100 epochs	**81.33**	**61.40**	**60.89**	**53.29**	**64.22**

Scope of the Theoretical Claim. While our bias-variance analysis motivates stage-dependent adaptation, it does not rigorously derive a globally monotonic increasing schedule. The linear warming strategy should be understood as a practical approximation that is empirically validated in Section 4.2, rather than a direct mathematical consequence of Theorem 3.3.

E Additional Experimental Details

Implementation Framework. All experiments are implemented in PyTorch 2.0. CIFAR-10/100 and ImageNet-200 experiments run on a single NVIDIA L20 GPU with 48 GB memory.

Data Augmentation. For CIFAR-10/100 and ImageNet-200, we apply random horizontal flip and random crop with 4-pixel padding. All images are normalized using dataset statistics.

Teacher Training Details. The teacher is trained for 100 epochs using the mean-teacher objective with both labeled and unlabeled data, where the supervised loss is computed on the labeled subset and the consistency loss on the unlabeled subset. We use SGD with momentum 0.9, weight decay 5×10^{-4}, and initial learning rate 0.1 decayed via cosine annealing. Adversarial examples for training are generated with PGD ($\epsilon = 8/255$, step size 2/255).

Contrastive Loss Configuration. The contrastive temperature κ is set to 0.07. Positive samples for a given anchor are those in the same batch that share the student's predicted class. The batch size for contrastive computation is 128.

Hyperparameter Sensitivity. We tested $\tau_{\max} \in \{2.0, 3.0, 4.0, 5.0\}$ on CIFAR-10. $\tau_{\max} = 4.0$ achieved the best PGD accuracy. The loss weight λ was varied in $\{1, 3, 5, 7\}$; $\lambda = 5$ performed best for CIFAR-10, while $\lambda = 1$ sufficed for CIFAR-100 and ImageNet-200. The method is robust to moderate variations in β (± 0.02).

Computational Cost. BayesAT adds minimal overhead compared to WSCAT. The additional computations include temperature scaling (negligible) and confidence weight calculation (already computed by teacher). Training time per epoch on CIFAR-10 is approximately 652 seconds, comparable to WSCAT (612 seconds).

F Broader Societal Impacts

Our work introduces BayesAT, a method designed to improve the robustness of deep neural networks against adversarial attacks in semi-supervised settings. We discuss both the potential positive societal impacts and the possible risks associated with this line of research.

Positive Impacts. The primary goal of BayesAT is to significantly enhance the adversarial robustness of image classifiers while reducing the reliance on large-scale manually annotated data. Improved adversarial robustness can directly benefit security-sensitive applications, such as autonomous driving perception systems, medical image analysis, and facial verification mechanisms, by making them more resistant to malicious attacks and manipulations. Furthermore, by effectively leveraging unlabeled data, our approach lowers the annotation barrier, which can democratize the development of robust AI models in resource-constrained domains (e.g., rare disease diagnosis or environmental monitoring) where labeled data is scarce. Theoretically, a more reliable model also fosters appropriate trust in AI systems during critical decision-making processes.

Potential Negative Impacts and Mitigations. Despite its benefits, the advancement of adversarial defense mechanisms inherently follows a dual-use nature. Specifically:

1. **Arms Race with Attackers:** The robust features learned by BayesAT might be reverse-engineered to design stronger, more targeted adaptive attacks. We believe that publicly disclosing defense mechanisms is essential for the community to collectively scrutinize and develop countermeasures, ultimately leading to long-term safety. To mitigate risks, we recommend that any deployment of models trained with BayesAT should be accompanied by continuous adversarial monitoring and online robustness certification tools.
2. **Fairness and Bias:** Our empirical evaluation relies on standard academic benchmarks (CIFAR-10/100, ImageNet-200). We did not explicitly test for performance disparities across different demographic subgroups or under-represented categories within these datasets, which is a limitation. If deployed in the real world, semi-supervised learning on biased unlabeled data could potentially amplify existing data biases. We encourage practitioners to rigorously audit quantized class-wise fairness metrics and confidence calibration before deploying our method in sensitive societal contexts.
3. **Computational Environmental Cost:** While BayesAT adds minimal overhead compared to the baseline WSCAT (as quantified in Appendix E), adversarial training remains significantly more computationally intensive than standard training. This leads to increased energy consumption and carbon footprint. Such costs should be weighed against the safety benefits, and efforts should be made to utilize energy-efficient hardware and carbon-neutral cloud services.

As foundational robustness research, BayesAT does not directly target specific harmful applications, but we urge future adopters to consider these implications and implement necessary safeguards, such as gated model releases or fairness constraints, when transitioning to high-stakes environments.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction accurately state the contributions (bias-variance analysis, progressive warming, confidence reweighting, SOTA results) and scope (semi-supervised adversarial training for image classification), which are fully supported by the theory in Sec. 3 and experiments in Sec. 4.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The paper includes a dedicated "Limitations and future work" paragraph in Sec. 5 (Conclusion), discussing the two-phase pipeline, reliance on manually chosen temperature bounds, and the focus on ℓ_∞ perturbations, with suggestions for future improvements.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes]

Justification: All lemmas and theorems (Lemma 1-2, Theorem 1-3) are clearly stated with required assumptions, and complete proofs are provided in Appendix A.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: Section 4.1 and Appendix E specify datasets, data splits, model architectures (WRN-28-10), attack parameters (ϵ, step size), optimizer (SGD with momentum, cosine annealing), hyperparameters ($\tau_{\min}, \tau_{\max}, \lambda, \beta$), and data augmentation, which are sufficient for reproduction.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).

(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification:Yes, the code is open-sourced at https://github.com/Sherlockssa/BayesAT with detailed instructions.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: Section 4.1 and Appendix E describe the data splits (e.g., 4k/46k for CIFAR-10), all hyperparameters (learning rate, batch size, λ, $\tau_{\min}$, $\tau_{\max}$, etc.), optimizer (SGD with Nesterov momentum, cosine annealing), and how hyperparameter ranges were tested.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: Table 1, Table 2 and Table 3 report mean $\pm$ standard deviation over three random seeds. The source of variability (random seeds) is explicitly stated in Sec. 4.1.

   Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Appendix E states that all experiments ran on a single NVIDIA L20 GPU and reports the training time per epoch (e.g., 652s/epoch on CIFAR-10) compared to the baseline.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The paper does not involve human subjects, sensitive data, or any practice that conflicts with the Code of Ethics. All used datasets are standard benchmarks.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [Yes]

Justification: The paper now includes an appendix on Broader Societal Impacts (Appendix F) that discusses positive aspects (improved security, reduced annotation cost) and potential negative aspects (arms race with attackers, fairness concerns, environmental cost), along with suggested mitigation strategies.

Guidelines:

- The answer [N/A] means that there is no societal impact of the work performed.
- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

Answer: [N/A]

Justification: The paper does not release any new datasets or models that carry a high risk of misuse; it presents a training algorithm for existing benchmarks.

Guidelines:

- The answer [N/A] means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [No]

Justification: The paper cites the original publications for datasets (CIFAR-10/100, ImageNet) and methods, but does not explicitly state the licenses (e.g., CC BY 4.0, MIT) or terms of use for these assets.

Guidelines:

- The answer [N/A] means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [N/A]Yes

Justification: The paper releases the implementation of BayesAT as a public GitHub repository with a README, configuration files, and instructions for reproducing the main results.

Guidelines:

- The answer [N/A] means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [N/A]

Justification: The paper does not involve crowdsourcing or human subjects research.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [N/A]

Justification: The paper does not involve human subjects.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [N/A]

Justification: The core methodology does not involve LLMs as an important or non-standard component.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.