# COMPASS: DUAL RECALIBRATION FOR LONG-TAILED ADVERSARIAL TRAINING

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Adversarial training (AT) has demonstrated remarkable success on balanced datasets, yet its effectiveness under long-tailed distributions remains challenging, as it necessitates simultaneously improving clean accuracy, overall robustness, and tail-class robustness. Existing methods for long-tailed robustness commonly mitigate class imbalance via prior-based corrections or class-wise perturbation schedules. Nevertheless, they frequently overlook two critical issues: whether a sample is sufficiently reliable in the clean domain to warrant strong adversarial supervision, and whether the final classifier still suffers from bias inherited from the long-tailed training prior. To address these issues, we propose COMPASS, a framework that navigates long-tailed AT through reliable robust supervision and classifier recalibration. COMPASS consists of two modules. Robust Supervision Recalibration (RSR) strengthens adversarial supervision only when class-level robust deficiency is supported by sample-level clean reliability, avoiding excessive perturbations on unstable tail samples. Train-only Classifier Recalibration (TCR) then freezes robust representations and interpolates the original adversarial classifier with a class-balanced recalibrated classifier selected solely from training data. Extensive experiments show that COMPASS substantially improves tail-class robustness and achieves competitive overall performance.

*AI Use Statement.* Generative AI assisted with code implementation and language refinement. The authors reviewed and verified all AI-assisted content.

*Reproducibility Statement.* Experimental settings and evaluation protocols are fully documented, and the source code is included in the supplementary material.

## 1 INTRODUCTION

Adversarial training (AT) is a cornerstone approach for improving model robustness against imperceptible input perturbations (Losch et al., 2024; Waseda et al., 2025; Tian & Mao, 2025). Its effectiveness is well established under balanced class distributions, where clean and adversarial objectives are closely aligned across categories. However, class imbalance is common in practical recognition systems: a few head classes dominate the sample population, while numerous tail classes are each supported by only sparse observations (Wang et al., 2024a; Hasegawa & Sato, 2024; Zhao et al., 2024b; Zhou et al., 2024). Hence, a practical model should simultaneously preserve overall clean accuracy and robustness, and prevent the clean or robust performance of tail classes from degrading.

Existing long-tailed AT approaches mainly focus on two perspectives. Prior-correction approaches reduce head-class bias via logit adjustment, sample reweighting, or classifier retraining (Ren et al., 2020; Sun et al., 2025). Class-aware robust methods tailor perturbation budgets, distillation targets, or regularization strength based on class-level statistics (Cho et al., 2025; Zhang et al., 2026). Overall, these methods largely characterize robust difficulty by class frequency or per-class robustness. Such simplification conflates two failure modes. When the local clean structure is reliable, adversarial perturbations can effectively regularize decision boundaries. However, for samples with low confidence, whose representations remain fragile, the same perturbations may instead amplify uncertainty and hurt generalization. Furthermore, even with a robust feature backbone, the final classifier can still inherit the empirical long-tailed prior and favor head classes under balanced evaluation.

In this work, we propose COMPASS, a framework that addresses the two aforementioned mismatches: robust supervision should be amplified only when the underlying clean prediction structure is trustworthy, and the learned robust representation should be converted into class decisions without inheriting long-tailed prior bias.



Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

We instantiate this principle through two coupled modules. Robust Supervision Recalibration (RSR) operates during training and makes robust supervision conditional on clean reliability. It first identifies classes with delayed robust learning, and then assigns larger perturbation budgets through samples with stable clean margins and adjusts robust weights according to aggregated clean–robust gaps. In this way, RSR strengthens robustness where it is needed while avoiding aggressive adversarial supervision on uncertain samples. Train-Only Classifier Recalibration (TCR) acts after robust representation learning. It freezes the backbone learned with RSR and refits only the final classifier using class-balanced training samples. Rather than replacing the original adversarial classifier, TCR interpolates between the original and recalibrated classifiers, with the interpolation coefficient selected on a split derived from the training set. This corrects residual long-tailed prior bias without discarding the robust decision structure learned during training.

Our contributions are as follows:

- We identify a clean-reliability mismatch in long-tailed AT: under-robust classes require stronger supervision, but unstable clean samples within these classes can be harmed by aggressive adversarial perturbations.
- We propose RSR, which combines class-level robust-gap feedback with sample-level clean-reliability gating, so that stronger adversarial supervision is mainly assigned to under-robust classes through trustworthy samples.
- We introduce TCR as a train-only classifier recalibration step that corrects residual long-tailed classifier bias without discarding the adversarial classifier learned during robust training.
- Through extensive experiments, we show that RSR and TCR address complementary failure modes: RSR realizes reliability-aware robust representation learning, while TCR converts the learned representation into less biased class decisions.

## 2 Related Work

**Adversarial training.** AT typically minimizes the worst-case loss within a bounded perturbation set (Goodfellow et al., 2015; Zhao et al., 2024a; 2026). Many follow-up works have made AT more effective, by improving the inner maximization (Pang et al., 2021; Jia et al., 2022), reducing training cost (Rebuffi et al., 2021; Tsiligkaridis & Roberts, 2022), using extra data or generative models (Gowal et al., 2021; Rade & Moosavi-Dezfooli, 2021; Wang et al., 2023), and applying regularization to inputs, features, or weights (Li et al., 2022; Herrmann et al., 2022; Mao et al., 2022; Yin et al., 2023; Zhu et al., 2023; Hua et al., 2024). Complementary studies have examined robust generalization through label noise, initialization, architecture, and optimization dynamics (Suzuki et al., 2023; Jain & Dutta, 2024; Li et al., 2024; Wu et al., 2024; Wang et al., 2024b; Li & Li, 2025). Although these advances have improved AT, most are designed for balanced regimes and do not explicitly model how reliable each long-tailed sample is for adversarial supervision.

**Long-tailed AT.** Long-tailed AT should address two biases simultaneously: the classifier bias induced by imbalanced class frequencies and the robustness bias induced by uneven adversarial supervision (Xian et al., 2026; Kang et al., 2019; Menon et al., 2020). Existing approaches typically target one of these two biases. Prior-correction methods, such as RoBal (Wu et al., 2021) and BSL (Yue et al., 2024), adjust logits or class priors to reduce the dominance of head classes. Equalization-based methods, including REAT (Li et al., 2023) and TAET (Yu-Hang et al., 2025), rebalance class-wise adversarial optimization. Recent methods modulate robust supervision more directly. UDR introduces distributional robustness (Bui et al., 2022), CFA adapts perturbation radii and robust coefficients based on class-wise robust accuracy (Wei et al., 2023), DAFA assigns perturbations with distance-aware feedback (Lee et al., 2024), self-distilled long-tailed AT transfers balanced robust knowledge from a teacher (Cho et al., 2025), and RLT constructs long-tailed perturbation priors and schedules (Zhang et al., 2026). These methods collectively show that robust training should be class-aware. However, class-level statistics alone cannot determine whether an individual sample is reliable enough to withstand strong adversarial pressure.

In contrast, COMPASS disentangles class-level robust-gap feedback from sample-level clean-reliability gating. The former identifies which classes require stronger robust supervision, while the latter determines which samples should receive such supervision. Additionally, it introduces a

train-only classifier recalibration method to mitigate the long-tailed prior bias inherited from robust representation learning.

## 3 METHOD

This section presents our proposed COMPASS. We first introduce Robust Supervision Recalibration (RSR), which modulates per-sample and per-class adversarial supervision during training. We then describe Train-Only Classifier Recalibration (TCR), which freezes learned representations and recalibrates the final classifier on resamples training data to remove long-tailed prior bias.

### ROBUST SUPERVISION RECALIBRATION (RSR)

Let $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{N}$ be a long-tailed training set with $C$ classes and class counts $\{n_c\}_{c=1}^{C}$. RSR can be formulated as a calibrated AT objective:

$$\mathcal{L}_{\mathrm{RSR}} = \mathbb{E}\left[\ell_{\mathrm{bal}}(f_\theta(x), y)\right] + \mathbb{E}\left[\beta_y^{(t)} \ell_{\mathcal{B}}(f_\theta, x, y, x^{adv})\right]. \tag{1}$$

Here, $\ell_{\mathrm{bal}}$ is a prior-calibrated classification loss that reduces the empirical class bias,

$$\ell_{\mathrm{bal}}(z, y) = -\log \frac{\exp(z_y + b_y)}{\sum_{j=1}^{C} \exp(z_j + b_j)}, \tag{2}$$

where logits $z = f_\theta(x)$, $b_c = \log(\pi_c)$ and $\pi_c = n_c/N$. $\beta_y^{(t)}$ denotes a class-dependent coefficient that controls the strength of robust supervision at training step $t$, and $\ell_{\mathcal{B}}$ is the robust loss inherited from the base adversarial learner (e.g., PGD-AT, TRADES, or AWP). The adversarial example $x^{adv}$ is generated using a calibrated perturbation budget:

$$x^{adv} \in \arg\max_{x' \in \mathcal{S}(x_i)} \phi_{\mathcal{B}}(f_\theta, x_i, y_i, x'), \tag{3}$$

where $\mathcal{S}(x_i) = \{x' : \|x' - x_i\|_\infty \le \epsilon_i^{(t)}\}$ defines the allowed perturbation set, and $\phi_{\mathcal{B}}$ is the attack objective used by the base learner to find the worst-case perturbation. RSR determines two terms in Eq. (1): the perturbation budget $\epsilon_i^{(t)}$ and the robust weight $\beta_c^{(t)}$.

*Reliability-Guided Budget Allocation.* This branch aims to assign strong perturbations to samples that are both from under-robust classes and cleanly trustworthy.

$$\epsilon_i^{(t)} = \epsilon q_{y_i,\epsilon}^{(t)} \left[s_{\min} + (1 - s_{\min}) r_i^{(t)}\right]. \tag{4}$$

Here, the class-level scale $q_{y_i,\epsilon}^{(t)}$ determines how much attack budget should be allocated to class $y_i$, while the sample-level gate $r_i^{(t)}$ determines how much of that class budget should be applied to sample $i$. For calculating $q_{c,\epsilon}^{(t)}$, RSR monitors class-wise robust accuracy on training set. Let $a_c^{rob,(t)}$ denote the robust accuracy of class $c$, and let $\bar{a}^{rob,(t)}$ be the macro average over classes, $q_{c,\epsilon}^{(t)}$ is formulated as

$$q_{c,\epsilon}^{(t)} = \max\left(1 + \lambda g_c^{rob,(t)}, q_\epsilon^{max}\right). \tag{5}$$

It increases the class-level attack budget for classes with delayed robust learning.

$$g_c^{rob,(t)} = \max(\bar{a}^{rob,(t)} - a_c^{rob,(t)}, 0), \tag{6}$$

$g_c^{rob,(t)}$ measures how far class $c$ falls behind the average robust performance.

For calculating $r_i^{(t)}$, given $(x_i, y_i)$ and logits $z_i = f_\theta(x_i)$, we calculate the clean margin as

$$m_i = z_{i,y_i} - \max_{k \neq y_i} z_{i,k}, \tag{7}$$

where a positive value indicates that the true class logit exceeds the strongest competing logit. Since margins are not directly comparable across long-tailed classes, RSR maintains a class-wise margin

threshold $\tau_c^{(t)}$. Let $\mathcal{I}_c^{(t)} = \{i : y_i = c\}$ denote the samples of class $c$ in the current mini-batch, the threshold is updated by

$$\tau_c^{(t)} = \mu_\tau \tau_c^{(t-1)} + (1 - \mu_\tau) \frac{1}{|\mathcal{I}_c^{(t)}|} \sum_{i \in \mathcal{I}_c^{(t)}} m_i, \tag{8}$$

where $\mu_\tau$ is the EMA momentum. The reliability score is then expressed as

$$r_i^{(t)} = \sigma\left(m_i - \tau_{y_i}^{(t)}\right), \tag{9}$$

where $\sigma(\cdot)$ is the sigmoid function. A sample receives a high trust score only when its clean margin is strong relative to the historical margin level of its class.

*Gap-Guided Robust Weight Calibration.* This branch adjusts the strength of the robust loss according to the current clean–robust gap, so that robust supervision is increased when robustness falls behind natural recognition. The class-dependent weight $\beta_c^{(t)}$ can be expressed as

$$\beta_c^{(t)} = \beta \tilde{q}_{c,\beta}^{(t)}, \tag{10}$$

where $\beta$ is the base robust coefficient. $\tilde{q}_{c,\beta}^{(t)}$ is the adaptive scale assigned to class $c$:

$$\tilde{q}_{c,\beta}^{(t)} = \max\Big(\big(1 + \Delta_{all}^{(t)}\big)\big(1 + \Delta_{tail}^{(t)} \chi_c\big),\, \tilde{q}_\beta^{\max}\Big), \tag{11}$$

where $\chi_c \in \{0, 1\}$ indicates whether class $c$ is a tail class. Let $A_{all}^{clean,(t)}$ and $A_{all}^{rob,(t)}$ denote the clean and robust accuracies over all classes of the training set, $A_{tail}^{clean,(t)}$ and $A_{tail}^{rob,(t)}$ denote the corresponding accuracies restricted to tail classes, the all-class and tail-class gaps are

$$\Delta_{all}^{(t)} = \left[A_{all}^{clean,(t)} - A_{all}^{rob,(t)}\right]_+, \quad \Delta_{tail}^{(t)} = \left[A_{tail}^{clean,(t)} - A_{tail}^{rob,(t)}\right]_+. \tag{12}$$

TRAIN-ONLY CLASSIFIER RECALIBRATION (TCR)

RSR aims to learn sharper classification boundaries, yet it is still trained on imbalanced data and therefore remains subject to the long-tailed prior. To mitigate this, TCR seeks a combination coefficient that interpolates between the weights of a balanced-sampling fine-tuned model and those of the RSR output model. Let the trained model be decomposed as

$$p(y|x) = \mathrm{softmax}(W f_\theta(x) + b), \tag{13}$$

where $f_\theta$ is the representation learning module and $(W, b)$ is the final linear classifier. TCR splits the training data into a class-balanced subset $\mathcal{D}_{recal}$ and a selection subset $\mathcal{D}_{sel}$. It freezes $\theta$ and optimizes only $(W, b)$ over $\mathcal{D}_{recal}$:

$$(W^*, b^*) = \arg\min_{W,b} \mathbb{E}_{(x,y)\sim\mathcal{D}_{recal}^{bal}} \left[\ell_{\mathcal{B}}(W f_\theta(x) + b, y)\right]. \tag{14}$$

Fully replacing the original classifier can over-correct the head-tail balance. TCR therefore interpolates between the original classifier $(W_0, b_0)$ and the recalibrated classifier $(W^*, b^*)$:

$$W_\alpha = (1 - \alpha) W_0 + \alpha W^*, \quad b_\alpha = (1 - \alpha) b_0 + \alpha b^*. \tag{15}$$

For each candidate $\alpha \in \mathcal{A}$, TCR evaluates the interpolated classifier on $\mathcal{D}_{sel}$ and chooses

$$\alpha^* = \arg\max_\alpha H_{\mathcal{D}_{sel}}(\alpha), \tag{16}$$

where $H$ is the harmonic mean of four train-selection metrics:

$$H(\alpha) = \frac{4}{\frac{1}{A_{all}^{clean}(\alpha)} + \frac{1}{A_{all}^{rob}(\alpha)} + \frac{1}{A_{tail}^{clean}(\alpha)} + \frac{1}{A_{tail}^{rob}(\alpha)}}. \tag{17}$$

The final evaluation on the test set is performed using $W_{\alpha^*}$ and $b_{\alpha^*}$.

**Algorithm 1** COMPASS: Navigating Long-Tailed Adversarial Training

**Input:** long-tailed training set $\mathcal{D}$, class counts $\{n_c\}_{c=1}^{C}$, base adversarial learner $\mathcal{B}$, candidate set $\mathcal{A}$, total epochs $T$.
**Output:** final model with recalibrated classifier $(f_\theta, W_{\alpha^*}, b_{\alpha^*})$.
*Stage I: Robust Supervision Recalibration (RSR)*
**for** epoch $t = 1, \ldots, T$ **do**
  Compute class robust gaps $g_c^{rob}$ and aggregate gaps $(\Delta_{all}, \Delta_{tail})$ by Eqs. (6) and (12).
  Form class-level perturbation scale $q_{c,\epsilon}^{(t)}$ and robust weights $\beta_c^{(t)}$ by Eqs. (5) and (10).
  **for** mini-batch $(x, y) \in \mathcal{D}$ **do**
    Compute clean margins $m_i$ and update class thresholds $\tau_{y_i}$ by Eq. (8).
    Estimate reliability scores $r_i^{(t)}$ by Eq. (9).
    Construct perturbation budgets $\epsilon_i^{(t)}$ by Eq. (4).
    Generate adversarial examples $x^{adv}$ with learner $\mathcal{B}$ under budgets $\epsilon_i^{(t)}$.
    Update the model by minimizing Eq. (1) with class weights $\beta_{y_i}^{(t)}$.
  **end for**
**end for**
*Stage II: Train-only Classifier Recalibration (TCR)*
Freeze the learned representation $f_\theta$.
Fit a class-balanced classifier $(W^*, b^*)$ on the recalibration subset by Eq. (14).
**for** candidate $\alpha \in \mathcal{A}$ **do**
  Interpolate $(W_\alpha, b_\alpha)$ by Eq. (15).
  Score $(W_\alpha, b_\alpha)$ on the train-selection subset by Eq. (17).
**end for**
Select $\alpha^*$ by Eq. (16) and return $(f_\theta, W_{\alpha^*}, b_{\alpha^*})$.

Table 1: Comparative results on ResNet. Dedicated LT-AT methods are typically built upon TRADES. Light gray cells indicate the best result within the same base learner, and dark gray bold cells indicate the best result among all methods.

| Dataset | Metric | Dedicated LT-AT | | | | PGD-AT | | | | | | TRADES | | | | | | AWP | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | REAT | TAET | RoBal | BSL | Base | +UDR | +CFA | +DAFA | +RLT | +Ours | Base | +UDR | +CFA | +DAFA | +RLT | +Ours | Base | +UDR | +CFA | +DAFA | +RLT | +Ours |
| C10-LT | Nat.(all) | 64.8% | 66.0% | 72.9% | 64.9% | 55.7% | 57.1% | 55.6% | 64.1% | 55.9% | **73.8%** | 53.8% | 54.8% | 53.3% | 62.8% | 55.1% | 68.6% | 49.4% | 42.9% | 53.0% | 54.1% | 57.7% | 69.3% |
| | Nat.(tail) | 57.1% | 58.7% | 68.7% | 57.1% | 45.6% | 47.2% | 45.4% | 56.9% | 45.8% | **71.4%** | 43.1% | 44.2% | 42.5% | 55.7% | 44.8% | 65.6% | 37.6% | 29.6% | 42.0% | 44.5% | 44.4% | 67.9% |
| | PGD(all) | 33.0% | 33.1% | 35.4% | 33.5% | 26.4% | 29.5% | 29.3% | 30.5% | 30.7% | 34.0% | 30.2% | 29.7% | 30.9% | 31.6% | 31.9% | 38.0% | 29.3% | 28.5% | 32.4% | 32.6% | 36.7% | **38.3%** |
| | PGD(tail) | 21.0% | 21.4% | 27.9% | 21.6% | 13.1% | 16.7% | 15.9% | 21.3% | 18.5% | 30.9% | 16.5% | 16.2% | 17.0% | 21.9% | 19.2% | 36.3% | 15.3% | 14.7% | 18.7% | 22.6% | 27.1% | **36.5%** |
| | C&W(all) | 31.9% | 31.7% | 34.5% | 32.6% | 29.6% | 29.3% | 29.0% | 31.4% | 30.7% | 34.6% | 29.4% | 29.4% | 29.9% | 30.4% | 30.8% | 36.1% | 28.0% | 27.7% | 31.3% | 31.5% | 30.1% | **36.2%** |
| | C&W(tail) | 19.9% | 19.8% | 26.9% | 20.4% | 16.1% | 16.4% | 15.3% | 22.6% | 18.5% | 31.6% | 15.7% | 15.7% | 15.8% | 20.7% | 18.1% | 33.9% | 13.8% | 13.8% | 17.3% | 21.2% | 16.3% | **34.3%** |
| | AA(all) | 29.9% | 29.9% | 32.1% | 30.5% | 27.5% | 27.2% | 27.3% | 28.5% | 27.9% | 30.1% | 28.2% | 27.7% | 29.0% | 28.7% | 29.5% | **34.1%** | 27.2% | 26.4% | 29.8% | 29.6% | 28.5% | 33.7% |
| | AA(tail) | 17.8% | 17.9% | 24.3% | 18.3% | 13.8% | 14.2% | 13.5% | 19.6% | 15.5% | 26.7% | 14.5% | 14.1% | 14.8% | 19.1% | 16.8% | 31.5% | 12.9% | 12.4% | 15.7% | 19.5% | 14.5% | **31.9%** |
| C100-LT | Nat.(all) | 46.2% | 45.9% | 45.6% | 46.5% | 43.1% | 42.4% | 43.0% | 44.2% | 45.7% | 50.6% | 43.7% | 41.3% | 43.4% | 44.7% | 45.2% | 48.8% | 44.8% | 42.5% | 46.8% | 44.2% | 45.7% | **54.7%** |
| | Nat.(tail) | 43.4% | 42.6% | 44.9% | 43.8% | 38.6% | 38.1% | 38.4% | 40.4% | 41.7% | 49.4% | 38.7% | 36.4% | 37.9% | 40.8% | 40.6% | 48.2% | 39.4% | 37.2% | 41.8% | 38.4% | 39.7% | **53.8%** |
| | PGD(all) | 20.8% | 19.1% | 19.9% | 20.7% | 15.4% | 16.3% | 17.7% | 16.8% | 17.9% | 16.4% | 20.1% | 18.8% | 20.8% | 19.8% | 20.8% | 21.4% | 21.4% | 20.8% | 22.0% | 22.9% | **23.0%** | 22.1% |
| | PGD(tail) | 19.6% | 17.6% | 19.8% | 19.4% | 13.3% | 14.5% | 15.6% | 15.6% | 16.3% | 16.4% | 17.6% | 16.4% | 17.7% | 18.3% | 18.6% | **21.8%** | 18.2% | 18.1% | 19.2% | 20.4% | 19.9% | 21.4% |
| | C&W(all) | 18.9% | 17.4% | 17.8% | 19.0% | 17.7% | 16.3% | 17.2% | 16.5% | 17.6% | 17.2% | 18.4% | 17.1% | 18.8% | 17.8% | 18.9% | 19.7% | 18.9% | 18.9% | 20.2% | 20.7% | **21.5%** | **21.5%** |
| | C&W(tail) | 17.8% | 16.0% | 17.8% | 17.6% | 15.4% | 14.5% | 15.1% | 15.5% | 15.9% | 17.0% | 15.9% | 15.1% | 16.2% | 16.7% | 17.1% | 19.7% | 16.2% | 16.2% | 17.4% | 18.3% | 18.8% | **21.2%** |
| | AA(all) | 17.3% | 15.9% | 16.6% | 17.3% | 16.0% | 14.7% | 15.5% | 14.7% | 15.5% | 15.0% | 17.3% | 15.8% | 17.5% | 16.5% | 17.4% | 17.9% | 17.5% | 17.0% | 18.2% | 18.9% | 19.3% | **19.7%** |
| | AA(tail) | 16.2% | 14.7% | 16.6% | 16.2% | 13.8% | 13.2% | 13.7% | 14.0% | 14.1% | 14.9% | 15.0% | 13.9% | 15.1% | 15.4% | 15.6% | 17.6% | 15.0% | 14.7% | 15.7% | 16.8% | 17.0% | **17.8%** |
| T-LT | Nat.(all) | 37.8% | 37.7% | 37.2% | 38.9% | 34.5% | 34.6% | 36.9% | 36.4% | 37.3% | 40.8% | 35.9% | 34.0% | 37.6% | 36.1% | 36.3% | 40.6% | 33.9% | 30.1% | 38.4% | 34.9% | 34.4% | **42.0%** |
| | Nat.(tail) | 36.0% | 34.8% | 36.1% | 37.0% | 29.8% | 29.6% | 32.1% | 32.2% | 32.8% | 40.4% | 30.8% | 28.9% | 32.4% | 32.0% | 31.1% | 39.5% | 28.0% | 24.6% | 32.7% | 29.4% | 28.1% | **41.4%** |
| | PGD(all) | 13.6% | 12.1% | 13.5% | 14.2% | 9.0% | 12.6% | 12.2% | 12.2% | 12.6% | 13.1% | 13.8% | 12.5% | 14.9% | 13.4% | 13.7% | 14.7% | 16.7% | 13.5% | 16.0% | 16.9% | 16.5% | **17.1%** |
| | PGD(tail) | 12.8% | 10.7% | 12.8% | 13.2% | 7.2% | 9.9% | 9.8% | 10.4% | 10.4% | 12.3% | 10.9% | 9.8% | 11.7% | 11.4% | 11.2% | 13.4% | 12.9% | 10.3% | 12.6% | 13.8% | 12.6% | **15.5%** |
| | C&W(all) | 11.2% | 10.2% | 10.7% | 11.7% | 12.3% | 11.5% | 11.6% | 11.8% | 12.2% | 12.2% | 11.7% | 10.1% | 12.0% | 10.7% | 11.2% | 12.3% | 14.6% | 11.7% | 14.1% | 14.5% | **14.8%** | 14.7% |
| | C&W(tail) | 10.3% | 9.0% | 10.0% | 10.7% | 9.6% | 8.8% | 9.4% | 9.9% | 9.8% | 11.3% | 9.2% | 8.0% | 9.4% | 9.1% | 9.0% | 11.7% | 11.3% | 8.9% | 11.0% | 11.7% | 11.3% | **13.0%** |
| | AA(all) | 10.3% | 9.3% | 9.7% | 10.6% | 10.9% | 10.2% | 10.4% | 10.4% | 11.0% | 10.6% | 10.8% | 9.4% | 11.3% | 9.8% | 10.3% | 11.2% | 13.0% | 10.4% | 12.5% | **13.2%** | **13.2%** | 12.3% |
| | AA(tail) | 9.5% | 8.3% | 9.0% | 9.7% | 8.6% | 7.8% | 8.3% | 8.8% | 9.0% | 9.8% | 8.4% | 7.5% | 8.8% | 8.3% | 8.3% | 10.6% | 10.0% | 7.7% | 9.8% | 10.7% | 10.0% | **11.5%** |

# 4 EXPERIMENTS

## EXPERIMENTAL DETAILS

**Datasets.** In line with RLT(Zhang et al., 2026), we employ the exponential long-tailed sampling strategy (Cao et al., 2019) to derive C10-LT, C100-LT, and T-LT from the original CIFAR10 Krizhevsky (2009), CIFAR100 Krizhevsky (2009), and TinyImageNet Le et al. (2015), respectively. The imbalance ratio is set to 50 for C10-LT and to 10 for the other two datasets. The test splits are kept intact, preserving the original balanced class frequencies.

Table 2: Comparative results on PreActResNet.

| Dataset | Metric | Dedicated LT-AT | | | | PGD-AT | | | | | | TRADES | | | | | | AWP | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | REAT | TAET | RoBal | BSL | Base | +UDR | +CFA | +DAFA | +RLT | +Ours | Base | +UDR | +CFA | +DAFA | +RLT | +Ours | Base | +UDR | +CFA | +DAFA | +RLT | +Ours |
| C10-LT | Nat.(all) | 65.5% | 65.7% | 72.6% | 65.2% | 57.9% | 60.0% | 57.5% | 65.8% | 57.4% | **74.4%** | 56.5% | 54.9% | 54.4% | 63.9% | 57.2% | 68.1% | 50.8% | 50.3% | 55.3% | 57.2% | 45.3% | 71.9% |
| | Nat.(tail) | 57.9% | 58.3% | 68.6% | 57.5% | 48.3% | 50.9% | 47.6% | 59.0% | 47.6% | **72.0%** | 46.5% | 44.4% | 43.8% | 57.2% | 47.5% | 64.9% | 39.3% | 38.5% | 44.9% | 48.1% | 32.3% | 69.9% |
| | PGD(all) | 33.0% | 32.6% | 36.1% | 33.2% | 26.4% | 29.9% | 29.8% | 30.5% | 31.0% | 34.4% | 31.3% | 30.4% | 30.7% | 31.9% | 31.9% | 36.5% | 29.9% | 31.2% | 33.0% | 33.4% | 31.4% | **38.3%** |
| | PGD(tail) | 20.9% | 20.8% | 29.1% | 21.1% | 13.1% | 17.0% | 16.4% | 21.5% | 19.0% | 31.1% | 18.1% | 16.8% | 16.7% | 22.5% | 19.5% | 32.2% | 15.9% | 17.6% | 19.4% | 23.6% | 17.7% | **36.6%** |
| | C&W(all) | 31.8% | 31.8% | 35.0% | 32.2% | 30.6% | 30.0% | 30.1% | 31.2% | 31.0% | 34.9% | 30.3% | 29.7% | 29.8% | 30.7% | 31.0% | 35.1% | 28.5% | 30.5% | 32.1% | 32.6% | 30.7% | **37.0%** |
| | C&W(tail) | 19.7% | 19.7% | 27.8% | 20.1% | 17.4% | 16.9% | 16.4% | 22.2% | 18.4% | 31.8% | 17.0% | 16.1% | 15.7% | 21.2% | 18.6% | 30.2% | 14.5% | 16.9% | 18.4% | 22.5% | 17.0% | **35.1%** |
| | AA(all) | 29.7% | 29.8% | **32.7%** | 29.8% | 28.0% | 27.6% | 27.9% | 28.5% | 28.5% | 30.4% | 29.1% | 28.4% | 28.6% | 29.1% | 29.6% | **32.7%** | 27.6% | 28.7% | 30.5% | 30.3% | 29.2% | 32.3% |
| | AA(tail) | 17.8% | 17.7% | 25.3% | 17.7% | 14.5% | 14.4% | 14.1% | 19.4% | 16.0% | 26.9% | 15.8% | 14.9% | 14.4% | 19.6% | 17.2% | 27.9% | 13.4% | 14.9% | 16.6% | 20.3% | 15.4% | **30.0%** |
| C100-LT | Nat.(all) | 46.5% | 45.9% | 45.1% | 46.8% | 43.1% | 42.2% | 43.2% | 44.3% | 46.7% | 50.5% | 43.3% | 41.0% | 43.9% | 44.4% | 45.5% | 50.3% | 45.2% | 46.0% | 46.9% | 44.3% | 45.8% | **55.1%** |
| | Nat.(tail) | 43.7% | 42.6% | 44.3% | 43.7% | 38.8% | 37.9% | 38.6% | 40.7% | 42.7% | 50.0% | 38.5% | 35.9% | 38.5% | 40.6% | 40.7% | 49.0% | 39.8% | 41.1% | 41.6% | 39.1% | 39.9% | **54.0%** |
| | PGD(all) | 21.0% | 19.4% | 20.2% | 20.8% | 15.5% | 15.6% | 17.8% | 16.9% | 18.1% | 18.3% | 20.4% | 18.7% | 20.9% | 19.9% | 20.9% | 21.4% | 21.8% | 21.6% | 21.8% | 23.1% | 23.2% | **23.7%** |
| | PGD(tail) | 19.7% | 18.0% | 20.1% | 19.4% | 13.8% | 13.7% | 15.8% | 15.7% | 16.5% | 17.9% | 17.8% | 16.2% | 17.8% | 18.5% | 18.9% | 21.3% | 18.6% | 18.7% | 19.0% | 20.6% | 20.3% | **21.8%** |
| | C&W(all) | 19.0% | 17.7% | 17.6% | 19.1% | 17.7% | 15.7% | 17.2% | 16.9% | 18.0% | 17.8% | 18.5% | 17.1% | 18.8% | 17.7% | 19.3% | 20.1% | 19.3% | 20.0% | 18.4% | 21.1% | 21.3% | **21.6%** |
| | C&W(tail) | 17.8% | 16.5% | 17.5% | 17.9% | 15.6% | 13.8% | 15.1% | 15.9% | 16.3% | 17.5% | 16.4% | 15.0% | 16.4% | 16.8% | 17.6% | 19.8% | 16.7% | 17.4% | 17.5% | 19.0% | 18.8% | **20.7%** |
| | AA(all) | 17.3% | 16.4% | 16.4% | 17.6% | 16.3% | 14.2% | 15.7% | 15.3% | 16.2% | 15.9% | 17.1% | 15.8% | 17.5% | 16.4% | 17.8% | 18.1% | 17.9% | 18.1% | 17.7% | **19.2%** | **19.2%** | 18.9% |
| | AA(tail) | 16.2% | 15.3% | 16.3% | 16.5% | 14.5% | 12.6% | 13.8% | 14.4% | 14.7% | 15.3% | 15.1% | 14.0% | 15.3% | 15.5% | 16.2% | 17.8% | 15.5% | 15.8% | 16.0% | 17.4% | 16.9% | **18.2%** |
| T-LT | Nat.(all) | 38.3% | 37.5% | 36.4% | 38.6% | 35.1% | 34.5% | 37.0% | 36.7% | 37.3% | 40.9% | 35.9% | 34.2% | 37.6% | 36.5% | 36.8% | 40.2% | 34.5% | 34.0% | 38.8% | 35.3% | 34.6% | **43.4%** |
| | Nat.(tail) | 36.6% | 34.7% | 35.3% | 36.8% | 30.8% | 29.9% | 32.2% | 32.6% | 32.5% | 40.5% | 30.7% | 29.0% | 32.2% | 32.4% | 31.7% | 39.0% | 28.6% | 28.4% | 33.1% | 29.9% | 28.2% | **42.3%** |
| | PGD(all) | 13.9% | 12.4% | 13.1% | 14.3% | 9.1% | 10.5% | 12.2% | 12.1% | 12.8% | 13.3% | 13.8% | 12.3% | 15.0% | 13.2% | 14.2% | 13.9% | 17.0% | 15.1% | 16.3% | 17.1% | 16.2% | **17.7%** |
| | PGD(tail) | 13.1% | 10.9% | 12.3% | 13.4% | 7.3% | 8.4% | 9.8% | 10.3% | 10.6% | 12.8% | 10.9% | 9.5% | 11.6% | 11.3% | 11.4% | 13.5% | 13.4% | 11.8% | 12.8% | 14.1% | 12.8% | **15.5%** |
| | C&W(all) | 11.4% | 10.6% | 10.7% | 11.8% | 12.4% | 11.4% | 11.9% | 12.5% | 12.9% | 12.4% | 11.7% | 10.1% | 11.8% | 10.9% | 11.7% | 11.3% | **14.9%** | 13.5% | 14.4% | 14.6% | 14.4% | 14.5% |
| | C&W(tail) | 10.6% | 9.3% | 10.1% | 10.8% | 9.6% | 9.0% | 9.9% | 10.6% | 10.1% | 11.4% | 9.4% | 7.9% | 9.2% | 9.4% | 9.4% | 10.7% | 11.7% | 10.4% | 11.4% | 11.7% | 10.8% | **13.2%** |
| | AA(all) | 10.3% | 9.6% | 9.9% | 10.7% | 11.2% | 10.1% | 10.6% | 11.2% | 11.4% | 11.1% | 10.9% | 9.2% | 11.3% | 10.1% | 10.6% | 10.1% | **13.5%** | 12.1% | 12.6% | 13.0% | 12.9% | 12.5% |
| | AA(tail) | 9.5% | 8.5% | 9.3% | 9.8% | 8.6% | 8.1% | 8.3% | 9.5% | 9.0% | 9.3% | 8.8% | 7.2% | 8.8% | 8.6% | 8.4% | 9.6% | 10.5% | 9.3% | 9.9% | 10.5% | 9.7% | **11.2%** |

Table 3: Comparative results on WideResNet.

| Dataset | Metric | Dedicated LT-AT | | PGD-AT | | | | | TRADES | | | | | AWP | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | RoBal | BSL | Base | +UDR | +DAFA | +RLT | +Ours | Base | +UDR | +DAFA | +RLT | +Ours | Base | +UDR | +DAFA | +RLT | +Ours |
| C10-LT | Nat.(all) | 72.5% | 67.7% | 58.9% | 59.4% | 65.1% | 60.7% | **77.4%** | 59.0% | 58.7% | 65.6% | 60.5% | 72.7% | 53.4% | 57.9% | 63.3% | 68.9% | 76.2% |
| | Nat.(tail) | 67.7% | 60.6% | 49.3% | 50.1% | 57.4% | 51.6% | 74.6% | 49.5% | 49.1% | 58.5% | 51.4% | 69.5% | 42.4% | 47.9% | 55.3% | 56.6% | **74.7%** |
| | PGD(all) | 34.6% | 30.6% | 26.9% | 28.1% | 29.6% | 31.5% | 34.6% | 32.4% | 31.4% | 33.5% | 33.6% | 39.9% | 31.2% | 32.7% | 35.9% | 38.7% | **40.0%** |
| | PGD(tail) | 25.8% | 19.1% | 13.1% | 15.0% | 18.7% | 19.1% | 29.7% | 19.1% | 18.1% | 22.9% | 21.0% | 35.1% | 17.0% | 19.5% | 25.9% | 27.1% | **38.0%** |
| | C&W(all) | 35.5% | 32.6% | 31.2% | 28.9% | 31.1% | 32.7% | 34.6% | 31.8% | 30.9% | 33.0% | 32.7% | **38.8%** | 30.2% | 32.4% | 35.4% | 37.1% | 38.6% |
| | C&W(tail) | 27.8% | 25.8% | 18.0% | 15.9% | 21.0% | 20.5% | 29.7% | 18.4% | 17.5% | 22.5% | 20.0% | 33.7% | 15.8% | 19.1% | 25.1% | 28.4% | **36.3%** |
| | AA(all) | 32.3% | 29.6% | 28.9% | 27.0% | 28.7% | 30.0% | 31.7% | 30.8% | 29.5% | 31.2% | 31.2% | **36.1%** | 29.2% | 29.9% | 33.0% | 35.0% | 35.7% |
| | AA(tail) | 24.5% | 22.6% | 15.5% | 13.8% | 18.6% | 17.8% | 26.6% | 17.3% | 16.2% | 20.7% | 18.4% | 30.9% | 14.8% | 16.4% | 22.8% | 24.7% | **32.0%** |
| C100-LT | Nat.(all) | 48.8% | 47.9% | 46.2% | 45.5% | 46.9% | 49.1% | 53.9% | 47.6% | 45.0% | 48.5% | 50.0% | 54.2% | 50.0% | 50.7% | 49.8% | 50.8% | **60.1%** |
| | Nat.(tail) | 47.4% | 45.1% | 41.7% | 40.9% | 43.2% | 44.6% | 52.5% | 42.2% | 39.9% | 44.8% | 45.3% | 52.6% | 44.6% | 45.7% | 45.1% | 45.2% | **59.6%** |
| | PGD(all) | 21.9% | 20.1% | 17.2% | 17.3% | 18.1% | 18.6% | 18.7% | 22.1% | 20.3% | 22.0% | 23.2% | 23.4% | 24.8% | 22.9% | 25.1% | **25.3%** | 24.4% |
| | PGD(tail) | 21.2% | 18.8% | 15.0% | 15.2% | 16.9% | 16.7% | 18.6% | 19.6% | 17.9% | 20.5% | 21.1% | 22.9% | 21.3% | 20.3% | 22.8% | 22.6% | **23.1%** |
| | C&W(all) | 19.4% | 19.2% | 19.5% | 17.5% | 18.5% | 18.9% | 19.5% | 20.8% | 19.0% | 20.4% | 21.6% | 23.1% | 22.1% | 21.8% | 23.5% | **23.8%** | 23.6% |
| | C&W(tail) | 19.1% | 18.3% | 17.4% | 15.6% | 17.7% | 17.3% | 19.4% | 18.4% | 16.8% | 19.2% | 19.6% | 22.6% | 19.1% | 19.3% | 21.3% | 21.2% | **23.0%** |
| | AA(all) | 18.0% | 17.3% | 17.4% | 15.8% | 16.8% | 16.9% | 17.2% | 19.4% | 17.5% | 19.0% | 20.2% | 20.5% | 20.7% | 20.0% | 21.4% | **21.5%** | 20.9% |
| | AA(tail) | 17.7% | 16.6% | 15.5% | 14.1% | 16.1% | 15.5% | 17.0% | 17.1% | 15.5% | 18.0% | 18.3% | 20.1% | 18.0% | 17.8% | 19.5% | 19.2% | **20.5%** |
| T-LT | Nat.(all) | 41.3% | 41.7% | 37.7% | 39.2% | 39.8% | 41.7% | 47.0% | 38.8% | 40.9% | 39.2% | 39.8% | 43.3% | 39.1% | 41.3% | 35.1% | 38.0% | **49.8%** |
| | Nat.(tail) | 40.1% | 36.9% | 33.5% | 34.5% | 35.9% | 36.9% | 46.5% | 33.6% | 35.9% | 34.9% | 34.8% | 42.0% | 32.9% | 35.7% | 29.7% | 31.3% | **48.3%** |
| | PGD(all) | 16.1% | 14.7% | 10.5% | 12.8% | 13.8% | 14.7% | 14.4% | 16.1% | 13.7% | 15.9% | 16.8% | 16.1% | 14.9% | 12.7% | 17.0% | **19.3%** | 18.9% |
| | PGD(tail) | 15.1% | 12.2% | 8.4% | 10.3% | 11.8% | 12.2% | 14.2% | 12.7% | 11.2% | 13.5% | 13.7% | 15.5% | 11.3% | 10.7% | 14.0% | 15.3% | **17.6%** |
| | C&W(all) | 13.0% | 15.1% | 14.4% | 13.3% | 14.1% | 15.0% | 14.7% | 13.8% | 12.2% | 13.3% | 14.2% | 15.1% | 12.5% | 10.5% | 16.5% | **17.1%** | **17.1%** |
| | C&W(tail) | 12.1% | 13.5% | 11.6% | 10.6% | 12.1% | 12.1% | 13.5% | 10.9% | 10.0% | 11.2% | 11.5% | 14.4% | 9.4% | 8.4% | 13.1% | 13.4% | **16.2%** |
| | AA(all) | 12.1% | 13.3% | 12.7% | 11.9% | 12.9% | 13.2% | 12.6% | 12.9% | 11.1% | 12.3% | 13.1% | 13.4% | 11.2% | 9.8% | 11.7% | **15.5%** | 14.5% |
| | AA(tail) | 11.3% | 12.0% | 10.2% | 9.5% | 11.1% | 11.0% | 11.7% | 10.2% | 9.1% | 10.3% | 10.5% | 12.5% | 8.3% | 7.6% | 8.9% | 12.2% | **13.5%** |

**Compared Methods.** We benchmark COMPASS against long-tailed robust training methods under three conventional AT baselines: PGD-AT (Madry et al., 2018), TRADES (Zhang et al., 2019), and AWP (Wu et al., 2020). The competing approaches fall into two categories—plug-in techniques, including UDR (Bui et al., 2022), CFA (Wei et al., 2023), DAFA (Lee et al., 2024), and RLT (Zhang et al., 2026), and dedicated long-tailed AT algorithms, namely RoBal (Wu et al., 2021), REAT (Li et al., 2023), BSL (Yue et al., 2024), and TAET (Yu-Hang et al., 2025).

**Evaluation Metrics.** We evaluate both natural and robust accuracy over all test classes (Nat.(all), Rob.(all)) as well as over the 80% tail classes that contain the fewest training samples (Nat.(tail), Rob.(tail)). Robustness is assessed under a 20-step $\ell_\infty$ PGD attack with perturbation size 8/255 and step length 2/255. To test whether the observed gains hold against more powerful adversaries, we additionally run C&W Carlini & Wagner (2017) and AutoAttack (Croce & Hein, 2020).

**Implementation Details.** We use ResNet18, PreActResNet-18 (He et al., 2016), and WideResNet (WRN-28-10) (Zagoruyko & Komodakis, 2016) as backbones. All models are trained for 110 epochs with SGD, using momentum $0.9$, weight decay $5 \times 10^{-4}$, and an initial learning rate of $0.1$. The perturbation radius is set to $\epsilon = 8/255$. During training, adversarial examples are generated by 10-step PGD. At test time, robust accuracy is evaluated by 20-step PGD with step size $2/255$.

For RSR, we set the minimum perturbation scale $s_{\min} = 0.3$ in Eq. (4), use $\lambda = 0.5$ in Eq. (5), and set both $q_\epsilon^{max}$ and $\tilde{q}_\beta^{\max}$ to 2. For TCR, we freeze the backbone after RSR training, refit the final classifier with class-balanced sampling, and select $\alpha$ from $\{0, 0.1, \ldots, 1.0\}$.

Table 4: Module-level ablation study of COMPASS on C10-LT with ResNet. The base model utilizes the prior-calibrated classification (PCC), Rel. denotes sample-level reliability, Rob. Gap denotes class-level robust-gap feedback, and Interp. denotes the interpolation step in TCR. The best result in each column is shown in bold, and the second-best result is underlined.

| Configuration | Components | | | | | | Natural | | PGD | | C&W | | AA | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | PCC | RSR | Rel. | Rob. Gap | TCR | Interp. | All | Tail | All | Tail | All | Tail | All | Tail |
| Base | ✓ | × | × | × | × | × | 63.76 | 58.05 | 32.13 | 22.99 | 32.68 | 22.27 | 30.71 | 20.32 |
| RSR only | × | ✓ | ✓ | ✓ | × | × | 61.06 | 52.89 | 31.58 | 21.39 | 30.46 | 20.12 | 28.56 | 18.38 |
| TCR only | × | × | × | × | ✓ | ✓ | <u>70.92</u> | **66.56** | <u>37.09</u> | 30.45 | 34.96 | 28.00 | 32.75 | 25.81 |
| Base + RSR | ✓ | ✓ | ✓ | ✓ | × | × | 68.07 | 62.25 | 35.38 | 27.52 | 34.43 | 26.26 | 31.89 | 23.73 |
| Base + RSR w/o Rel. | ✓ | ✓ | × | ✓ | × | × | 63.05 | 56.38 | 36.05 | 27.07 | 33.74 | 24.29 | 32.03 | 22.54 |
| Base + RSR w/o Rob. Gap | ✓ | ✓ | ✓ | × | × | × | **72.06** | <u>66.52</u> | 32.24 | 23.41 | 32.77 | 23.67 | 29.01 | 20.25 |
| Base + TCR | ✓ | × | × | × | ✓ | ✓ | 68.25 | 65.09 | 36.84 | 34.82 | 35.02 | 31.55 | 33.09 | 30.17 |
| COMPASS w/o Interp. | ✓ | ✓ | ✓ | ✓ | ✓ | × | 49.84 | 57.99 | 31.96 | **38.03** | 30.02 | **35.74** | 29.01 | **34.55** |
| COMPASS (Base + RSR + TCR) | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | 68.55 | 65.64 | **37.96** | <u>36.33</u> | **36.06** | <u>33.90</u> | **34.05** | <u>31.54</u> |

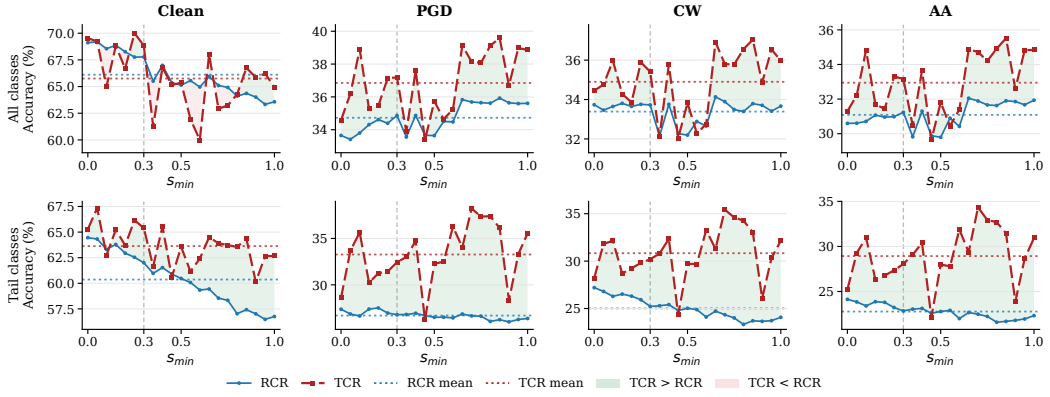



Figure 1: Sensitivity analysis of $s_{\min}$ in Eq. (4). We use $s_{\min} = 0.3$ because it corresponds to settings whose robust accuracy is close to the average performance.

### Comparative Results

Tables 1–3 present the main comparison on ResNet, PreActResNet, and WideResNet. **(1)** Under same base adversarial learner, Table 1 shows notable gains of COMPASS on tail-oriented metrics with ResNet while maintaining competitive clean accuracy. On C10-LT, PGD-AT+RLT obtains 55.9% Nat.(all), 45.8% Nat.(tail), and 18.5% PGD(tail), whereas PGD-AT+COMPASS reaches 73.8%, 71.4%, and 30.9%, respectively. Similar tail-oriented improvements are observed on PreActResNet and WideResNet. These results indicate that the tail-robustness gains are generally obtained without a severe loss in natural accuracy. **(2)** The effectiveness of COMPASS is not restricted to a specific robust objective. With WideResNet on C10-LT, TRADES+COMPASS achieves 39.9% PGD(all) and 35.1% PGD(tail), while AWP+COMPASS reaches 40.0% and 38.0%. On C100-LT, AWP+COMPASS obtains 60.1% Nat.(all), 59.6% Nat.(tail), and the highest PGD(tail) among AWP-based variants. **(3)** Compared with dedicated LT-AT methods, COMPASS remains competitive. For instance, on C10-LT, it outperforms representative baselines in both tail natural accuracy and tail PGD robustness across all backbones. **(4)** C&W and AutoAttack provide additional evaluations. On C10-LT with WideResNet, TRADES+COMPASS obtains 33.7% C&W(tail) and 30.9% AA(tail), while AWP+COMPASS reaches 36.3% and 32.0%, higher than comparative methods. On C100-LT, AWP+COMPASS also achieves the highest C&W(tail) and AA(tail), reaching 23.0% and 20.5%, respectively. These results indicate that the improvements of COMPASS are maintained under more powerful attack protocols.

### Ablation Study

This subsection uses the TRADES as the base learner.

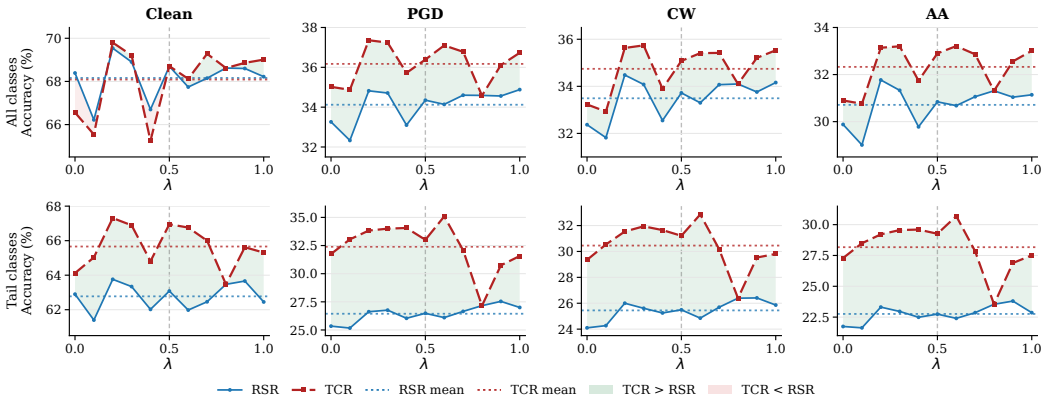



Figure 2: Sensitivity analysis of the class-level perturbation coefficient $\lambda$ in Eq. (5). We set $\lambda = 0.5$ because it corresponds to settings whose robust accuracy is close to the average performance.

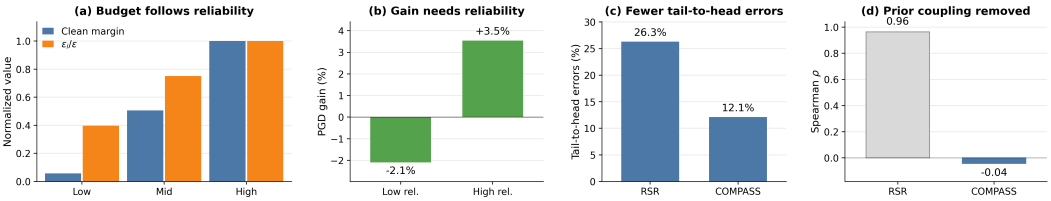



Figure 3: Mechanism-level ablation of COMPASS. Here, $\epsilon_i/\epsilon$ denotes the normalized perturbation budget assigned to sample $i$, PGD gain denotes the change in robust accuracy over the baseline, and $\rho$ denotes the Spearman rank correlation between classifier bias and log class frequency. (a) Low-reliability samples receive smaller perturbation budgets. (b) Reliable tail samples obtain larger robustness gains. The default uses no additional supervision for both high- and low-reliability samples. (c) Tail-to-head errors are substantially reduced after TCR. (d) The bias-frequency correlation is nearly eliminated after TCR.

**Module ablation.** Table 4 separates the effects of PCC, RSR, and TCR. Without PCC, RSR alone is weak on both clean accuracy and robustness. Adding RSR to PCC improves PGD(all)/PGD(tail) from 32.13%/22.99% to 35.38%/27.52%, with consistent gains under C&W and AutoAttack. Removing either sample reliability or robust-gap feedback weakens the clean-robust trade-off, confirming that RSR requires both reliable samples and class-level robust deficiency. TCR further improves tail robustness, but removing interpolation severely reduces Nat.(all), showing that direct replacement over-corrects the boundary.

**Hyperparameter ablation.** Figures 1 and 2 analyze the sensitivity of COMPASS to the main RSR controls. In our experiments, we set the default values to $s_{\min} = 0.3$ and $\lambda = 0.5$. These defaults are not selected as the best point; instead, they correspond to settings whose robust accuracy is close to the average performance, which avoids over-tuning the method to a specific hyperparameter value.

Figure 1 investigates the effect of the minimum perturbation scale $s_{\min}$, which controls the adversarial budget assigned to low-reliability samples. For small $s_{\min}$, uncertain or misclassified instances receive small perturbations, which helps preserve natural accuracy. As $s_{\min}$ increases, the protection is progressively relaxed, exposing these samples to stronger attacks. However, the resulting robustness does not improve monotonically, indicating that stronger perturbations are not always beneficial when the clean prediction is unstable.

Figure 2 studies the class-level perturbation coefficient $\lambda$. When $\lambda$ is small, the perturbation radius changes only slightly across classes. As a result, classes with delayed robust learning receive insufficient additional adversarial supervision, which limits the improvement in robust accuracy. As $\lambda$ increases, RSR applies stronger perturbations to these classes, which can enhance robustness.

Table 5: Computational overhead of COMPASS. 'Base' utilizes the prior-calibrated classification. Full COMPASS includes both RSR training and TCR recalibration.

| Method | C10-LT | | | C100-LT | | | T-LT | | |
|---|---|---|---|---|---|---|---|---|---|
| | Time | Rel. time | Peak mem. | Time | Rel. time | Peak mem. | Time | Rel. time | Peak mem. |
| Base | 23m 18s | 1.00× | 2100 MiB | 33m 52s | 1.00× | 2104 MiB | 4h 3m | 1.00× | 6663 MiB |
| Base+RSR | 23m 25s | 1.00× | 2104 MiB | 33m 47s | 1.00× | 2474 MiB | 4h 3m | 1.00× | 6096 MiB |
| TCR extra | 1m 38s | 0.07× | 1936 MiB | 2m 10s | 0.06× | 1956 MiB | 13m 11s | 0.05× | 6150 MiB |
| Full COMPASS | 25m 3s | 1.08× | 2104 MiB | 35m 57s | 1.06× | 2474 MiB | 4h 16m | 1.05× | 6150 MiB |

Table 6: Effect of the TCR interpolation coefficient $\alpha$ on C10-LT with ResNet. The selected coefficient is marked by $^*$ and determined by the harmonic score over Nat.(all), PGD(all), Nat.(tail), and PGD(tail) on the training-derived selection split.

| $\alpha$ | All classes | | | | Tail classes | | | | $\alpha$ | All classes | | | | Tail classes | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Nat. | PGD | C&W | AA | Nat. | PGD | C&W | AA | | Nat. | PGD | C&W | AA | Nat. | PGD | C&W | AA |
| 0.1 | 68.3 | 35.9 | 35.0 | 32.4 | 62.9 | 28.7 | 27.2 | 24.8 | 0.6 | 65.1 | 37.6 | 35.3 | 33.4 | 65.4 | 37.2 | 34.5 | 32.6 |
| 0.2 | 68.5 | 36.9 | 35.6 | 33.0 | 63.8 | 30.8 | 28.9 | 26.4 | 0.7 | 62.7 | 36.9 | 34.5 | 33.0 | 64.6 | 38.2 | 35.0 | 33.6 |
| 0.3 | 68.5 | 37.7 | **36.1** | **33.8** | 64.8 | 32.8 | 30.6 | 28.3 | 0.8 | 59.4 | 36.0 | 33.7 | 32.3 | 63.0 | 38.5 | 35.8 | 34.4 |
| 0.4 | 68.1 | 37.8 | 36.0 | **33.8** | 65.4 | 34.3 | 31.8 | 29.9 | 0.9 | 55.3 | 34.3 | 32.2 | 30.9 | 60.7 | **38.6** | **36.1** | **34.8** |
| 0.5$^*$ | 67.7 | **38.0** | 35.6 | 33.7 | **65.7** | 36.4 | 33.1 | 31.5 | 1.0 | 50.2 | 32.3 | 30.2 | 29.3 | 58.1 | 38.3 | 35.7 | 34.7 |

Notably, the clipping in Eq. (5) limits the coefficient range, preventing catastrophic overfitting and saturating performance for large $\lambda$.

**Mechanism-level ablation.** Figure 3 verifies the intended behavior of RSR and TCR. **(1)** Low-reliability samples have smaller clean margins and receive smaller perturbation budgets, showing that RSR suppresses strong perturbations on clean-unstable samples. **(2)** The effect of RSR on tail samples is reliability-dependent. Low-reliability tail samples show a negative PGD gain, suggesting that adversarial calibration is ineffective when the clean decision is still unstable. By contrast, high-reliability tail samples improve by 3.5%, showing that stronger adversarial supervision is beneficial when the sample already has a stable clean decision. This contrast supports sample-level reliability conditioning as a necessary component of robust supervision calibration. **(3)** TCR markedly reduces head-class attraction: the fraction of tail errors predicted as head classes drops from 26.3% to 12.1%. **(4)** Additionally, at the parameter level, the Spearman correlation (Spearman, 1987) between classifier bias and log class frequency decreases from 0.96 to -0.04, indicating that TCR removes the frequency dependence of the classifier bias term.

**Computational overhead.** Table 5 shows that COMPASS introduces only modest computational overhead. Adding RSR to Base keeps the training time essentially unchanged across C10-LT, C100-LT, and T-LT, since its additional operations are limited to reliability estimation and class-wise feedback. TCR is also lightweight, taking only $0.07\times$, $0.06\times$, and $0.05\times$ of the base training time on the three datasets. Furthermore, Full COMPASS keeps peak memory comparable to the base pipeline.

**Effect of TCR interpolation.** Table 6 examines the role of $\alpha$ in TCR. A small $\alpha$ preserves the original adversarial classifier, favoring natural accuracy over tail robustness. Increasing $\alpha$ substantially enhances tail robustness, e.g., PGD(tail) rises from 28.7% to 38.6% at $\alpha = 0.9$, but this comes at the cost of natural accuracy, which drops from 68.3% to 55.3% at $\alpha = 0.9$ and further to 50.2% at $\alpha = 1.0$. The chosen $\alpha = 0.5$ realizes a balance.

## 5 Conclusion

In this paper, we present COMPASS, a two-stage framework for long-tailed adversarial training. RSR follows a reliability-aware principle and couples class-level robust feedback with sample-level clean reliability during representation learning. Namely, under-robust classes warrant stronger adversarial supervision only when their clean prediction structure is trustworthy. TCR then corrects the residual classifier bias using class-balanced sampled training data, while preserving the adversarially learned classifier. Together, the two stages improve tail-class robustness while maintaining competitive overall clean and robust accuracy. *This study uses public benchmarks and involves no human participants or personal data. We adhere to the ICLR Code of Ethics.*

## REFERENCES


Anh Tuan Bui, Trung Le, Quan Hung Tran, He Zhao, and Dinh Phung. A unified wasserstein distributional robustness framework for adversarial training. In *International Conference on Learning Representations*, 2022.

Kaidi Cao, Colin Wei, Adrien Gaidon, Nikos Arechiga, and Tengyu Ma. Learning imbalanced datasets with label-distribution-aware margin loss. In *Advances in neural information processing systems*, volume 32, 2019.

Nicholas Carlini and David Wagner. Towards evaluating the robustness of neural networks. In *2017 ieee symposium on security and privacy (sp)*, pp. 39–57. Ieee, 2017.

Seungju Cho, Hongsin Lee, and Changick Kim. Long-tailed adversarial training with self-distillation. In *International Conference on Learning Representations*, volume 2025, pp. 35954–35972, 2025.

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

Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In *International Conference on Learning Representations*, 2015.

Sven Gowal, Sylvestre-Alvise Rebuffi, Olivia Wiles, Florian Stimberg, Dan Calian, and Timothy Mann. Improving robustness using generated data. In *Advances in Neural Information Processing Systems*, 2021.

Naoya Hasegawa and Issei Sato. Exploring weight balancing on long-tailed recognition problem. In *International Conference on Learning Representations*, volume 2024, pp. 40705–40732, 2024.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Identity mappings in deep residual networks. In *European Conference on Computer Vision*, pp. 630–645, 2016.

Charles Herrmann, Kyle Sargent, Lu Jiang, Ramin Zabih, Huiwen Chang, Ce Liu, Dilip Krishnan, and Deqing Sun. Pyramid adversarial training improves vit performance. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13419–13429, 2022.

Andong Hua, Jindong Gu, Zhiyu Xue, Nicholas Carlini, Eric Wong, and Yao Qin. Initialization matters for adversarial transfer learning. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 24831–24840. IEEE, 2024.

Samyak Jain and Tanima Dutta. Towards understanding and improving adversarial robustness of vision transformers. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 24736–24745. IEEE, 2024.

Xiaojun Jia, Yong Zhang, Baoyuan Wu, Ke Ma, Jue Wang, and Xiaochun Cao. Las-at: Adversarial training with learnable attack strategy. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13398–13408, 2022.

Bingyi Kang, Saining Xie, Marcus Rohrbach, Zhicheng Yan, Albert Gordo, Jiashi Feng, and Yannis Kalantidis. Decoupling representation and classifier for long-tailed recognition. *arXiv preprint arXiv:1910.09217*, 2019.

Alex Krizhevsky. Learning multiple layers of features from tiny images. 2009.

Yann Le, Xuan Yang, et al. Tiny imagenet visual recognition challenge. *CS 231N*, 7(7):3, 2015.

Hyungyu Lee, Saehyung Lee, Hyemi Jang, Junsung Park, Ho Bae, and Sungroh Yoon. Dafa: Distance-aware fair adversarial training. In *International Conference on Learning Representations*, volume 2024, pp. 28780–28809, 2024.

Binghui Li and Yuanzhi Li. Adversarial training can provably improve robustness: Theoretical analysis of feature learning process under structured data. In *International Conference on Learning Representations*, 2025.

Guanlin Li, Guowen Xu, and Tianwei Zhang. Alleviating the effect of data imbalance on adversarial training, 2023.

Tao Li, Yingwen Wu, Sizhe Chen, Kun Fang, and Xiaolin Huang. Subspace adversarial training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13409–13418, 2022.

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 *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 24776–24785. IEEE, 2024.

Max Losch, Mohamed Omran, David Stutz, Mario Fritz, and Bernt Schiele. On adversarial training without perturbing all examples. In *International Conference on Learning Representations*, volume 2024, pp. 47995–48015, 2024.

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations*, 2018.

Xiaofeng Mao, Gege Qi, Yuefeng Chen, Xiaodan Li, Ranjie Duan, Shaokai Ye, Yuan He, and Hui Xue. Towards robust vision transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12042–12051, 2022.

Aditya Krishna Menon, Sadeep Jayasumana, Ankit Singh Rawat, Himanshu Jain, Andreas Veit, and Sanjiv Kumar. Long-tail learning via logit adjustment. *arXiv preprint arXiv:2007.07314*, 2020.

Tianyu Pang, Xiao Yang, Yinpeng Dong, Kun Xu, Jun Zhu, and Hang Su. Bag of tricks for adversarial training. In *International Conference on Learning Representations*, 2021.

Rahul Rade and Seyed-Mohsen Moosavi-Dezfooli. Helper-based adversarial training: Reducing excessive margin to achieve a better accuracy vs. robustness trade-off. In *ICML 2021 Workshop on Adversarial Machine Learning*, 2021.

Sylvestre-Alvise Rebuffi, Sven Gowal, Dan A. Calian, Florian Stimberg, Olivia Wiles, and Timothy A. Mann. Fixing data augmentation to improve adversarial robustness. In *Advances in Neural Information Processing Systems*, 2021.

Jiawei Ren, Cunjun Yu, Xiao Ma, Haiyu Zhao, Shuai Yi, et al. Balanced meta-softmax for long-tailed visual recognition. In *Advances in Neural Information Processing Systems*, pp. 4175–4186, 2020.

Charles Spearman. The proof and measurement of association between two things. *The American journal of psychology*, 100(3/4):441–471, 1987.

Siyu Sun, Han Lu, Jiangtong Li, Yichen Xie, Tianjiao Li, Xiaokang Yang, Liqing Zhang, and Junchi Yan. Rethinking classifier re-training in long-tailed recognition: Label over-smooth can balance. In *International Conference on Learning Representations*, volume 2025, pp. 26681–26697, 2025.

Satoshi Suzuki, Shin'ya Yamaguchi, Shoichiro Takeda, Sekitoshi Kanai, Naoki Makishima, Atsushi Ando, and Ryo Masumura. Adversarial finetuning with latent representation constraint to mitigate accuracy-robustness tradeoff. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4390–4401, 2023.

Runzhi Tian and Yongyi Mao. Algorithmic stability based generalization bounds for adversarial training. In *International Conference on Learning Representations*, volume 2025, pp. 16610–16631, 2025.

Theodoros Tsiligkaridis and Jay Roberts. Understanding and increasing efficiency of frank-wolfe adversarial training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 50–59, 2022.

Binwu Wang, Pengkun Wang, Wei Xu, Xu Wang, Yudong Zhang, Kun Wang, and Yang Wang. Kill two birds with one stone: Rethinking data augmentation for deep long-tailed learning. In *International Conference on Learning Representations*, volume 2024, pp. 19402–19425, 2024a.

Zeyu Wang, Tianyu Pang, Chao Du, Min Lin, and Weiwei Liu. Better diffusion models further improve adversarial training. In *International Conference on Machine Learning*, 2023.

Zeyu Wang, Xianhang Li, Hongru Zhu, and Cihang Xie. Revisiting adversarial training at scale. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 24675–24685. IEEE, 2024b.

Futa Waseda, Ching-Chun Chang, and Isao Echizen. Rethinking invariance regularization in adversarial training to improve robustness-accuracy trade-off. In *International Conference on Learning Representations*, volume 2025, pp. 45053–45076, 2025.

Zeming Wei, Yifei Wang, Yiwen Guo, and Yisen Wang. Cfa: Class-wise calibrated fair adversarial training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8193–8201, 2023.

Dongxian Wu, Shu-Tao Xia, and Yisen Wang. Adversarial weight perturbation helps robust generalization. In *Advances in Neural Information Processing Systems*, pp. 2958–2969, 2020.

Tong Wu, Ziwei Liu, Qingqiu Huang, Yu Wang, and Dahua Lin. Adversarial robustness under long-tailed distribution. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8655–8664. IEEE, 2021.

Yu-Yu Wu, Hung-Jui Wang, and Shang-Tse Chen. Annealing self-distillation rectification improves adversarial training. In *International Conference on Learning Representations*, volume 2024, pp. 8516–8537, 2024.

Guanmeng Xian, Ning Yang, and Philip S Yu. Manifold-constrained adversarial training for long-tailed robustness via geometric alignment. *arXiv preprint arXiv:2605.02183*, 2026.

Shenglin Yin, Kelu Yao, Sheng Shi, Yangzhou Du, and Zhen Xiao. Again: Adversarial training with attribution span enlargement and hybrid feature fusion. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 20544–20553. IEEE, 2023.

Wang Yu-Hang, Junkang Guo, Aolei Liu, Kaihao Wang, Zaitong Wu, Zhenyu Liu, Wenfei Yin, and Jian Liu. Taet: Two-stage adversarial equalization training on long-tailed distributions. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15476–15485. IEEE, 2025.

Xinli Yue, Ningping Mou, Qian Wang, and Lingchen Zhao. Revisiting adversarial training under long-tailed distributions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 24492–24501, 2024.

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

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

Lilin Zhang, Yimo Guo, Yue Li, Jiancheng Shi, and Xianggen Liu. Taming the long tail: Rebalancing adversarial training via adaptive perturbation. In *2026 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2026.

Mengnan Zhao, Lihe Zhang, Jingwen Ye, Huchuan Lu, Baocai Yin, and Xinchao Wang. Adversarial training: A survey, 2024a.

Mengnan Zhao, Lihe Zhang, Bo Wang, Tianhang Zheng, Hong Zhong, and Geyong Min. Mitigating error amplification in fast adversarial training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2026.

Qihao Zhao, Yalun Dai, Hao Li, Wei Hu, Fan Zhang, and Jun Liu. Ltgc: Long-tail recognition via leveraging llms-driven generated content. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19510–19520. IEEE, 2024b.

Zhipeng Zhou, Liu Liu, Peilin Zhao, and Wei Gong. Pareto deep long-tailed recognition: A conflict-averse solution. In *International Conference on Learning Representations*, volume 2024, pp. 54500–54526, 2024.

Kaijie Zhu, Xixu Hu, Jindong Wang, Xing Xie, and Ge Yang. Improving generalization of adversarial training via robust critical fine-tuning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4424–4434, 2023.