

Manifold-Constrained Adversarial Training for Long-Tailed Robustness via Geometric Alignment

Guanmeng Xian¹, Ning Yang^{1*}, Philip S. Yu²

¹Sichuan University, Chengdu, China

²University of Illinois at Chicago, USA

xianguanmeng@stu.scu.edu.cn, yangning@scu.edu.cn, psyu@uic.edu

Abstract

Adversarial training is effective on balanced datasets, but its robustness degrades under long-tailed class distributions, where tail classes suffer high robust error and unstable decision boundaries. We propose *Manifold-Constrained Adversarial Training (MCAT)*, a unified framework that enforces the semantic validity of adversarial examples by penalizing deviations from class-conditional manifolds in feature space, while promoting balanced geometric separation across classes via an ETF-inspired regularization. We provide theoretical results that link geometric separation to lower bounds on adversarially robust margins, and show that manifold-constrained adversarial risk upper-bounds robust risk on high-density semantic regions. Extensive experiments on standard long-tailed benchmarks demonstrate consistent improvements in overall, balanced, and tail-class adversarial robustness. The codes and appendix are available on <https://github.com/yneversky/MCAT>.

1 Introduction

Deep neural networks have achieved remarkable success in visual recognition tasks, yet their vulnerability to adversarial perturbations remains a fundamental concern. Among existing defenses, adversarial training, formulated as a min-max optimization problem, is widely regarded as one of the most effective and principled approaches. However, the evaluation of adversarial robustness has largely focused on balanced benchmarks, whereas real-world data are often characterized by long-tailed class distributions [Wu *et al.*, 2021; Zhang *et al.*, 2023; Zhang *et al.*, 2025]. Under such imbalance, tail classes not only suffer from degraded clean accuracy, but also exhibit disproportionately weaker adversarial robustness, raising serious concerns about the reliability and fairness of robust models in practice.

Motivated by this gap, recent studies have begun to investigate adversarial robustness under long-tailed distributions. RoBal [Wu *et al.*, 2021] pioneers this line of work by introducing margin rebalancing combined with classifier adjust-

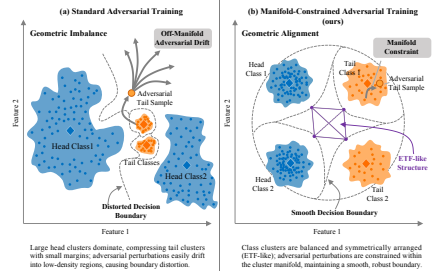


Figure 1: Adversarial training under long-tailed data in feature space. **Left:** Standard adversarial training leads to geometric imbalance and off-manifold adversarial drift, resulting in unstable and spurious decision boundaries for tail classes. **Right:** MCAT alleviates both issues by enforcing balanced class geometry and constraining adversarial perturbations to semantic manifolds.

ment. Subsequent methods further improve tail robustness through loss reweighting, multi-stage training strategies, or class-aware regularization [Ren *et al.*, 2020; Li *et al.*, 2021; Liu *et al.*, 2022; Zhang and Feng, 2024; Zhang *et al.*, 2022; Ahn *et al.*, 2023; Du *et al.*, 2023; Xu *et al.*, 2021; Li *et al.*, 2023; Yu-Hang *et al.*, 2025; Yue *et al.*, 2024; Gupta *et al.*, 2025]. Despite the progress achieved, most existing approaches primarily operate at the level of loss design or optimization heuristics, and do not explicitly regulate the geometry of learned representations or the semantic validity of adversarial examples.

In this paper, we attribute the failure of adversarial training under long-tailed distributions to two closely coupled mechanisms: *imbalance-induced geometric misalignment* and *off-manifold adversarial drift* (Figure 1). First, head-class-dominated optimization distorts the representation geometry, compressing inter-class margins associated with tail classes and rendering their decision boundaries fragile and unstable. Second, due to the scarcity of tail samples, unconstrained adversarial optimization is prone to exploit low-density and semantically unsupported regions of the feature space, diverting robustness away from the true data support. Together, these effects result in severe robust-margin collapse and unreliable predictions for tail classes.

Several complementary approaches based on knowledge

*Corresponding author

transfer, such as long-tailed adversarial self-distillation [Cho *et al.*, 2025], attempt to alleviate data scarcity at the decision level. While effective to some extent, these methods remain largely orthogonal to the representation-space issues discussed above, as they neither explicitly correct geometric misalignment nor constrain adversarial examples to lie within semantically meaningful regions.

We argue that achieving adversarial robustness under long-tailed distributions fundamentally requires addressing both the geometry of the learned decision space and the location of adversarial examples. To this end, we propose **Manifold-Constrained Adversarial Training (MCAT)**, a unified framework grounded in a twofold geometric principle. First, MCAT constrains adversarial perturbations to remain close to class-conditional semantic manifolds in feature space, ensuring that robustness is learned within high-density and semantically valid regions. Notably, although tail classes are sparsely sampled in pixel space, their representation-space structure is substantially more regular and lower-dimensional, which makes such manifold constraints feasible even with limited tail data. Second, MCAT promotes balanced inter-class geometry by aligning classifier weight vectors toward a simplex *Equiangular Tight Frame (ETF)* structure [Papayan *et al.*, 2020], thereby restoring margin-balanced decision boundaries. Our theoretical analysis shows that manifold constraints effectively control robust risk within the semantic support, while geometric alignment induces provable lower bounds on adversarially robust margins. By jointly enforcing semantic validity and geometric alignment, MCAT stabilizes adversarial decision boundaries and substantially improves robustness for both head and tail classes.

Our main contributions are summarized as follows:

- We identify two fundamental mechanisms underlying the degradation of adversarial robustness under long-tailed distributions: geometric misalignment and off-manifold adversarial drift.
- We propose **MCAT**, a unified adversarial training framework that integrates manifold-constrained perturbations with ETF-inspired geometric alignment.
- We provide theoretical guarantees that link geometric separation to adversarially robust margins and show that manifold constraints control robust risk on semantic support.
- Extensive experiments demonstrate consistent improvements in overall, balanced, and tail-class adversarial robustness.

2 Preliminaries

Let $\mathcal{D} = \{(x, y)\}$ denote a long-tailed dataset with class prior π_y , where $x \in \mathbb{R}^d$ and $y \in \{1, \dots, C\}$. Let $f_\Theta : \mathbb{R}^d \rightarrow \mathbb{R}^C$ be a classifier parameterized by Θ , and let $\phi_\Theta(x) \in \mathbb{R}^m$ corresponds to the output of the feature extractor before the final linear classification layer of f . The final linear classifier is parameterized by weights $W \in \mathbb{R}^{C \times m}$, whose y -th row w_y corresponds to class y . We denote by $s_\Theta(x) = f_\Theta(x)$ the logit vector, where $s_k(x)$ is the logit associated with class k . Let $\ell(\cdot, \cdot)$ denote a standard classification loss, such as cross-entropy. Expectations $\mathbb{E}_{(x,y) \sim \mathcal{D}}[\cdot]$ are taken with respect to

the empirical training distribution induced by \mathcal{D} . We consider an ℓ_∞ threat model: $\mathcal{B}_\epsilon(x) = \{x' \mid \|x' - x\|_\infty \leq \epsilon\}$. The robust risk is defined as

$$R_{\text{robust}}(\Theta) = \mathbb{E}_{(x,y) \sim \mathcal{D}} \left[\max_{x' \in \mathcal{B}_\epsilon(x)} \ell(f_\Theta(x'), y) \right]. \quad (1)$$

3 Method

3.1 Overview

As illustrated in Figure 2, under long-tailed distributions, standard adversarial training tends to shape decision boundaries in low-density regions while being dominated by head-class geometry. This leads to unstable decision boundaries and severely reduced margins for tail classes. MCAT addresses these issues by jointly enforcing the *semantic validity* of adversarial examples and a *balanced geometry* of the decision space.

Concretely, MCAT consists of two complementary components. First, adversarial perturbations are constrained to remain close to *class-conditional semantic manifolds* in the feature space (Section 3.2), thereby guiding adversarial optimization toward high-density and semantically meaningful regions. Second, the classifier weight vectors are regularized toward a simplex *Equiangular Tight Frame (ETF)* structure (Section 3.3), which encourages uniform angular separation between classes. These two components are combined into a unified min-max training objective (Section 3.4) and optimized using a manifold-aware PGD procedure (Section 3.5).

3.2 Class-Conditional Semantic Manifolds in Feature Space

We assume that features of each class y concentrate around a low-dimensional semantic support in representation space, denoted as \mathcal{M}_y . Rather than explicitly recovering \mathcal{M}_y , we employ a class-conditional generator G_y as a proxy to characterize off-manifold deviation.

Let $z \sim \mathcal{N}(0, I)$ be a latent code. Each generator $G_y : \mathbb{R}^k \rightarrow \mathbb{R}^m$ is a lightweight MLP mapping latent codes to the feature space, $z \sim \mathcal{N}(0, I)$, $\tilde{\phi}_y = G_y(z)$. The generators $\{G_y\}$ are pretrained using features extracted by a classifier f_Θ by minimizing

$$\min_{G_y} \mathbb{E}_{x \sim \mathcal{D}_y, z \sim \mathcal{N}(0, I)} \|G_y(z) - \phi_\Theta(x)\|_2^2. \quad (2)$$

Learning G_y directly in representation space substantially reduces the intrinsic complexity of tail classes compared to pixel-space generation, making class-conditional manifold approximation feasible even with limited samples. After pretraining, all generators are frozen throughout adversarial training. Although ϕ_Θ continues to evolve during robust optimization, its class-conditional structure changes gradually, allowing G_y to act as a stable semantic reference that regularizes adversarial drift rather than enforcing exact reconstruction.

We measure off-manifold deviation of an embedding $u = \phi_\Theta(x)$ by

$$d_{\mathcal{M}_y}(u) = \min_z \|u - G_y(z)\|_2^2, \quad (3)$$

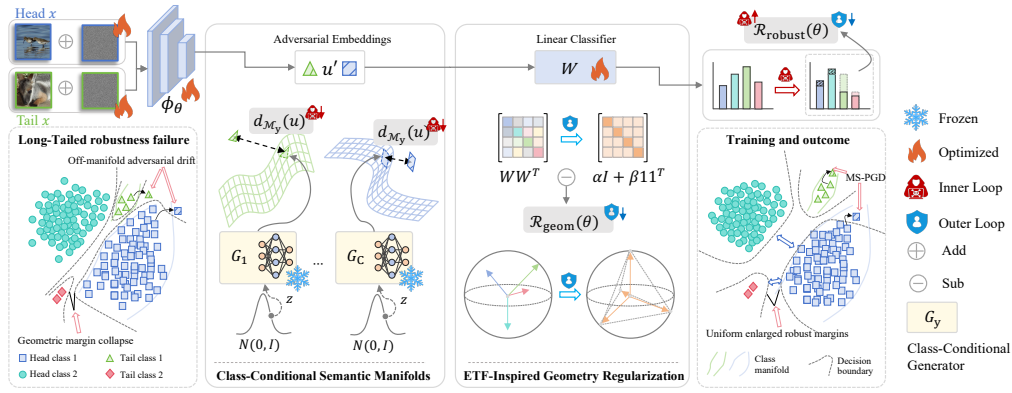


Figure 2: Overview of MCAT for long-tailed adversarial robustness. **Left:** Under long-tailed data, standard adversarial training exhibits (i) *off-manifold adversarial drift* and (ii) *geometric margin collapse* for tail classes. **Middle:** MCAT couples two mechanisms: a *class-conditional manifold distance penalty* in feature space and an *ETF-inspired geometric alignment* of classifier weights. **Right:** Manifold-Constrained PGD (MS-PGD) learns robust decision boundaries near high-density semantic support while preserving enlarged (and more uniform) robust margins across classes.

where the inner minimization is approximated by T_z steps of gradient descent on z , warm-started from a per-sample cache. We report a sensitivity analysis with respect to T_z in Appendix Table 6, and further verify the validity of frozen generators by tracking reconstruction error over training epochs, which remains stable across classes, including tail classes (Appendix Figure 10a).

3.3 ETF-Inspired Geometry Regularization

To counteract imbalance-induced geometric compression, we regularize classifier weights W toward a simplex Equiangular Tight Frame (ETF) structure by penalizing deviations of the Gram matrix:

$$\mathcal{R}_{geom}(\Theta) = \|W^\top W - \alpha I - \beta \mathbf{1}\mathbf{1}^\top\|_F^2, \quad (4)$$

where α and β are scalar parameters, I is the identity matrix, and $\mathbf{1}$ is the all-ones vector.

This regularizer promotes approximately equal-norm and equiangular classifier weights, thereby enlarging and stabilizing the minimum inter-class angle θ_{\min} . By Theorem 1, a larger θ_{\min} directly implies a larger certifiable robust margin. Since the simplex ETF maximizes the minimum pairwise angle among class vectors, it represents an optimal geometry for robustness under fixed dimensionality. Under long-tailed adversarial training, head-dominated optimization distorts balanced geometry, which our regularization counteracts to prevent tail-class margin collapse.

3.4 Unified Objective

We combine manifold constraints (Eq. (3)) and geometric regularization (Eq. (4)) into a single objective:

$$R_{MCAT}(\Theta) = \min_{\theta} \mathbb{E}_{(x,y) \sim \mathcal{D}} \left[\max_{\|\delta\|_{\infty} \leq \epsilon} (\ell(f_{\Theta}(x + \delta), y) - \lambda d_{\mathcal{M}_y}(\phi_{\Theta}(x + \delta))) \right] + \beta \mathcal{R}_{geom}(\Theta). \quad (5)$$

where λ controls semantic consistency and β controls geometric balancing.

Algorithm 1 MCAT Training

Require: Dataset \mathcal{D} , classifier f_{Θ} , generators $\{G_y\}$, steps T , budget ϵ , step size η , weights λ, β

- 1: **for** each iteration **do**
- 2: Sample mini-batch $\{(x_i, y_i)\}_{i=1}^n$
- 3: **for** each sample i **do**
- 4: Initialize $\delta_{i,0} \sim [-\epsilon, \epsilon]$
- 5: **for** $t = 0$ to $T - 1$ **do**
- 6: Update $\delta_{i,t+1}$ according to Equation (6)
- 7: **end for**
- 8: $x_i^{adv} \leftarrow x_i + \delta_{i,T}$
- 9: **end for**
- 10: Update θ according to Equation (5)
- 11: **end for**

3.5 Manifold-Constrained Inner Maximization

The inner maximization is solved using Manifold Supported PGD (MS-PGD):

$$\delta_{t+1} = \Pi_{\|\delta\|_{\infty} \leq \epsilon} \left(\delta_t + \eta \nabla_x [\ell(f_{\Theta}(x + \delta_t), y) - \lambda d_{\mathcal{M}_y}(\phi_{\Theta}(x + \delta_t))] \right), \quad (6)$$

where $\Pi_{\|\delta\|_{\infty} \leq \epsilon}$ denotes projection by clipping. MS-PGD preserves the standard ℓ_{∞} threat model while biasing adversarial search toward semantically supported regions.

3.6 Training Algorithm

Algorithm 1 summarizes the MCAT training procedure.

4 Theoretical Analysis

We provide two complementary results that connect MCAT to (i) margin-based robustness induced by balanced representation geometry, and (ii) robust-risk control through suppressing off-manifold adversarial drift. Proofs are deferred to the appendix A

4.1 Theorem 1: Geometric Separation Implies Robust Margin Lower Bound

We assume $\|\phi_\Theta(x)\|_2 = 1$ for all x , and that the feature map ϕ_Θ is L -Lipschitz under ℓ_∞ perturbations, i.e., $\|\phi_\Theta(x + \delta) - \phi_\Theta(x)\|_2 \leq L\epsilon$ for all $\|\delta\|_\infty \leq \epsilon$. Let w_y denote the classifier weight vector for class y , and define the minimum inter-class angle as

$$\theta_{min} = \min_{i \neq j} \arccos\left(\frac{w_i^\top w_j}{\|w_i\|_2 \|w_j\|_2}\right).$$

Theorem 1 (Robust Margin from Geometric Separation). *If $\epsilon < \sin(\theta_{min}/2)/L$, then the predicted label of x remains invariant to all perturbations in $\mathcal{B}_\epsilon(x)$.*

Corollary 1 (Sample-wise Robust Radius). *Let $s_y(x) = w_y^\top \phi_\Theta(x)$ denote the logit of the true class y , and define the logit margin $\gamma(x) = s_y(x) - \max_{k \neq y} s_k(x)$. Then the sample-wise robust radius satisfies $r(x) \geq \frac{\gamma(x)}{2L}$.*

Remark 1. Theorem 1 shows that adversarial robustness is governed by the minimum inter-class angle θ_{min} . In particular, ETF maximizes θ_{min} among normalized class weights, and thus attains the largest robust margin implied by the theorem. Theorem 1 establishes a sufficient robustness condition governed by the minimum inter-class angle, while Corollary 1 yields a sample-dependent lower bound via the logit margin. Together, they illustrate why geometry matters in long-tailed robust training: when head-dominated updates compress tail margins, the guaranteed robust radius for tail samples can quickly vanish.

4.2 Theorem 2: Manifold Constraint and Robust Risk Control

Theorem 2 (Manifold-Constrained Training Controls Robust Risk). *Assume that, for each class y , the data distribution is supported on a semantic manifold \mathcal{M}_y , and that regions far from \mathcal{M}_y have negligible probability mass. Then, for the MCAT objective $R_{MCAT}(\theta)$,*

$$R_{robust}(\Theta) \leq R_{MCAT}(\Theta) + O(\lambda^{-1}).$$

Remark 2. Theorem 2 shows that robust risk is controlled by the MCAT objective up to a residual term $O(\lambda^{-1})$, which arises from off-manifold adversarial drift under finite λ . While such drift cannot be fully eliminated, it is the *uncontrolled* drift—common in standard adversarial training—that is especially harmful for tail classes due to sparse sampling and weak manifold support. The manifold penalty in MCAT significantly suppresses this drift and confines it to a limited range, preventing large semantic deviations and spurious decision boundaries in low-density regions, without restricting the adversary’s ℓ_∞ budget.

5 Experiments

5.1 Goals and Research Questions

We evaluate **MCAT** on standard long-tailed adversarial robustness benchmarks to answer the following research questions.

- **RQ1 (Overall robustness):** Does MCAT improve adversarial robustness on long-tailed data compared with standard adversarial training and long-tailed robust baselines?
- **RQ2 (Tail and balanced robustness):** Does MCAT improve robustness for tail classes and class-balanced metrics without sacrificing head-class performance?
- **RQ3 (Component contribution and sensitivity):** How do the individual components of MCAT (manifold constraint and geometric alignment) and their associated hyperparameters contribute to robustness under long-tailed distributions?
- **RQ4 (Mechanism verification and theory consistency):** Can we empirically verify the two failure mechanisms in Figure 1 (*geometry compression* and *off-manifold adversarial drift*), and observe empirical trends consistent with our theoretical analysis?

5.2 Experimental Settings

Datasets and Long-Tailed Construction

Benchmarks. We conduct experiments on CIFAR-10-LT, CIFAR-100-LT, and Tiny-ImageNet-LT following prior long-tailed robustness protocols [Wu *et al.*, 2021; Yue *et al.*, 2024; Cho *et al.*, 2025].

Imbalance ratio (IR). Given a dataset with C classes, we construct a long-tailed training set by exponentially decaying the number of samples per class. Let n_{max} be the maximum class size (head) and n_{min} the minimum class size (tail), then $IR = n_{max}/n_{min}$. We evaluate multiple imbalance levels, e.g., $IR \in \{10, 20, 50, 100\}$, and report the default setting for each benchmark consistent with prior work [Wu *et al.*, 2021; Yue *et al.*, 2024; Cho *et al.*, 2025].

Baselines

Standard adversarial training baselines. We consider commonly used adversarial training methods including TRADES [Zhang *et al.*, 2019], MART [Wang *et al.*, 2020], AWP [Wu *et al.*, 2020], LAST-AT [Jia *et al.*, 2022] as widely adopted in recent studies [Yue *et al.*, 2024].

Long-tailed adversarial training baselines. We compare to representative long-tailed robustness methods such as RoBal [Wu *et al.*, 2021], REAT [Li *et al.*, 2023], TAET [Yu-Hang *et al.*, 2025], and long-tailed adversarial self-distillation [Cho *et al.*, 2025]. We follow the official implementations or reproduce their reported settings under a unified protocol whenever possible.

MCAT ablations. To isolate the effects of each component, we evaluate: (i) **Base AT** (or the strongest common baseline), (ii) **Base AT + manifold constraint only**, (iii) **Base AT + geometric alignment only**, (iv) **MCAT (full)** (manifold constraint + geometric alignment + MS-PGD).

Architectures and Training Details

Backbones. For CIFAR-10/100-LT, we use ResNet-18 as the default backbone and optionally include WideResNet-34-10 for stronger capacity comparisons, following prior long-tailed robustness evaluations [Cho *et al.*, 2025; Yue *et al.*, 2024]. For Tiny-ImageNet-LT, we use a standard residual backbone (e.g., PreActResNet-18) consistent with previous protocols [Cho *et al.*, 2025].

Adversarial training. Unless otherwise specified, we consider the ℓ_∞ threat model with perturbation budget ϵ (e.g., $\epsilon = 8/255$ on CIFAR). For training, we use a multi-step PGD inner maximization (e.g., $T = 10$ steps) with step size η and random initialization in $[-\epsilon, \epsilon]$, following standard practice. For MCAT, the inner maximization is replaced by MS-PGD (Section 3.5) with the manifold penalty weight λ .

Generators for class-conditional manifolds. We train the class-conditional generators $\{G_y\}$ on clean features (Section 3.2) with gradients stopped to θ and keep $\{G_y\}$ fixed during robust training. We report generator architecture, latent dimension, and training iterations in Appendix B.

The hyperparameter settings of MCAT, the baselines, and the training process are provided by Table 8 in Appendix D.

Evaluation Protocol

Attacks. We evaluate robustness under a suite of increasingly strong white-box attacks: FGSM, multi-step PGD (e.g., PGD-20 and optionally PGD-100), and AutoAttack (AA). We keep ϵ consistent with training and use standard step sizes and iterations as in prior work [Wu *et al.*, 2021; Yue *et al.*, 2024; Yu-Hang *et al.*, 2025].

Model selection and robust overfitting. To mitigate robust overfitting effects, we report both: (i) **best checkpoint** selected by validation PGD robustness, and (ii) **last checkpoint** at the final epoch, following robust training evaluation conventions [Yue *et al.*, 2024; Cho *et al.*, 2025].

Metrics

Standard accuracy and robustness. We report clean accuracy (**Clean Acc**) and robust accuracy (**Robust Acc**) under each attack (PGD-20/AA).

Tail and group-wise robustness. We report head/tail robust accuracy under PGD-20 and AA. We also report **tail-only** robustness (e.g., Tail-PGD, Tail-AA) to directly quantify tail reliability.

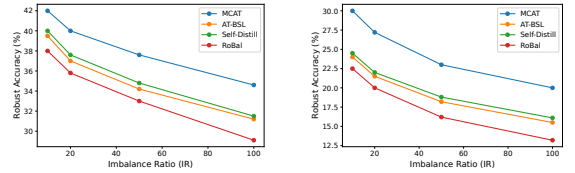
Balanced metrics. To measure fairness under class imbalance, we report: **Balanced Accuracy (BA)** and **Balanced Robustness (BR)** [Yu-Hang *et al.*, 2025], defined as the average per-class accuracy under clean and adversarial evaluation, respectively. Concretely, letting \mathcal{A}_c denote accuracy on class c , we compute $BA = \frac{1}{C} \sum_{c=1}^C \mathcal{A}_c^{\text{clean}}$, $BR = \frac{1}{C} \sum_{c=1}^C \mathcal{A}_c^{\text{adv}}$.

Reporting. We report mean and standard deviation over multiple runs with different random seeds.

5.3 RQ1: Overall Adversarial Robustness

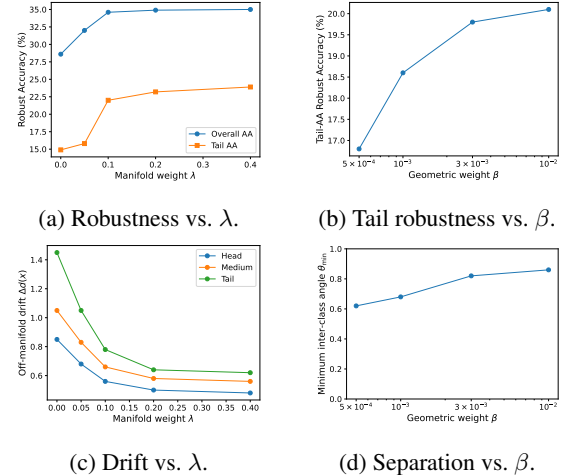
We evaluate **overall adversarial robustness** under long-tailed distributions. Table 1 reports clean accuracy and robust accuracy under PGD-20 and AutoAttack (AA) on CIFAR-10-LT, CIFAR-100-LT, and Tiny-ImageNet-LT with imbalance ratio $IR = 100$.

Across all benchmarks, MCAT consistently achieves the strongest adversarial robustness. In particular, MCAT substantially improves AutoAttack robustness over standard adversarial training baselines (PGD-AT, TRADES, MART, AWP) and recent long-tailed robust methods, with larger gains on CIFAR-100-LT and Tiny-ImageNet-LT. Notably, these improvements are achieved without sacrificing clean accuracy, indicating a favorable robustness–accuracy trade-off under severe imbalance.



(a) Overall AA robustness. (b) Tail-class AA robustness.

Figure 3: Adversarial robustness under increasing imbalance severity on CIFAR-100-LT. **Left:** overall robust accuracy under AutoAttack (AA). **Right:** tail-class robust accuracy under AutoAttack (Tail-AA).



(a) Robustness vs. λ . (b) Tail robustness vs. β . (c) Drift vs. λ . (d) Separation vs. β .

Figure 4: Sensitivity analysis of MCAT hyperparameters on CIFAR-100-LT ($IR=100$). Increasing λ suppresses off-manifold adversarial drift and improves robustness, while increasing β enlarges inter-class angular separation and enhances tail robustness.

5.4 RQ2: Tail and Balanced Robustness

Robustness under increasing imbalance severity. We evaluate robustness under increasing imbalance severity by varying the imbalance ratio on CIFAR-100-LT. Figure 3 summarizes the results on CIFAR-100-LT under AutoAttack.

As imbalance becomes more severe, all methods experience performance degradation. However, MCAT degrades substantially more gracefully. In particular, MCAT maintains higher overall AA robustness and preserves substantially stronger Tail-AA robustness even under severe imbalance. Complete numerical results are reported in Tables 4 and 5 in Appendix C.

Class-balanced and tail-class robustness. Overall robustness metrics can obscure failures on tail classes. To assess robustness fairness, Table 2 reports balanced accuracy (BA), balanced robustness (BR), and tail-class robustness on CIFAR-100-LT ($IR=100$).

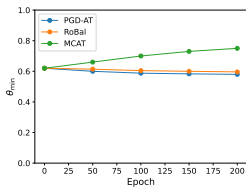
MCAT achieves the highest BA and BR among all compared methods and yields substantial gains in tail robustness under both PGD-20 and AA, indicating that its improvements

Method	CIFAR-10-LT (IR=100)			CIFAR-100-LT (IR=100)			Tiny-ImageNet-LT (IR=100)		
	Clean	PGD-20	AA	Clean	PGD-20	AA	Clean	PGD-20	AA
PGD-AT	83.20±0.30	48.10±0.40	45.20±0.45	55.30±0.35	27.40±0.50	24.60±0.55	46.80±0.40	18.90±0.55	16.80±0.60
TRADES	83.60±0.25	49.80±0.35	46.90±0.40	56.10±0.30	28.90±0.45	25.90±0.50	47.50±0.35	19.80±0.50	17.60±0.55
MART	83.40±0.28	50.20±0.38	47.30±0.42	55.90±0.32	29.20±0.48	26.20±0.52	47.20±0.38	20.10±0.52	18.00±0.58
AWP	84.10±0.22	51.60±0.34	48.70±0.38	56.80±0.28	30.50±0.44	27.30±0.48	48.30±0.32	21.40±0.48	19.20±0.52
RoBal	84.30±0.24	52.90±0.36	50.10±0.40	58.20±0.30	32.10±0.46	29.10±0.50	49.50±0.35	22.80±0.50	20.40±0.55
REAT	84.80±0.22	54.10±0.34	51.30±0.38	59.10±0.28	33.40±0.44	30.40±0.48	50.30±0.32	23.90±0.48	21.60±0.52
TAET	85.10±0.20	54.90±0.33	52.10±0.37	59.80±0.26	34.10±0.42	31.10±0.46	50.90±0.30	24.60±0.46	22.30±0.50
Self-Distill	85.30±0.21	55.30±0.34	52.60±0.38	60.10±0.27	34.60±0.43	31.50±0.47	51.20±0.31	25.10±0.47	22.80±0.51
AT-BSL	85.00±0.22	55.00±0.35	52.30±0.39	59.90±0.28	34.30±0.44	31.20±0.48	51.00±0.32	24.80±0.48	22.50±0.52
MCAT (ours)	86.20±0.18	57.40±0.30	55.10±0.34	62.30±0.24	37.10±0.40	34.60±0.44	53.80±0.28	28.90±0.44	26.40±0.48

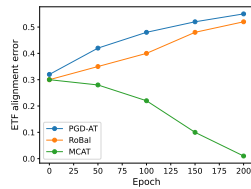
Table 1: Overall robustness under long-tailed distributions with imbalance ratio IR=100. We report clean accuracy and robust accuracy under PGD-20 and AutoAttack (AA). Results are reported as mean±std over three random seeds.

Method	BA ↑	BR (AA) ↑	Tail-PGD ↑	Tail-AA ↑
PGD-AT	39.80±0.45	15.60±0.50	12.90±0.55	11.20±0.60
TRADES	40.60±0.42	17.30±0.48	13.80±0.52	12.10±0.58
MART	40.10±0.44	17.80±0.49	14.20±0.54	12.40±0.59
AWP	41.20±0.40	18.70±0.46	15.00±0.50	13.10±0.55
RoBal	44.80±0.38	20.60±0.44	16.30±0.48	13.20±0.52
REAT	46.10±0.36	21.90±0.42	17.70±0.46	14.80±0.50
TAET	46.80±0.35	22.60±0.41	18.40±0.45	15.60±0.49
Self-Distill	47.20±0.34	23.10±0.40	19.00±0.44	16.10±0.48
AT-BSL	46.90±0.35	22.80±0.41	18.60±0.45	15.50±0.49
MCAT (ours)	51.80±0.30	27.40±0.36	22.30±0.40	20.00±0.44

Table 2: Balanced and tail robustness on CIFAR-100-LT (IR=100). We report balanced accuracy (BA) and balanced robustness (BR), defined as average per-class accuracy under clean evaluation and AutoAttack (AA), respectively, together with tail-class robust accuracy under PGD-20 (Tail-PGD) and AutoAttack (Tail-AA).



(a) Minimum inter-class angle.



(b) ETF alignment error.

Figure 5: Geometry under long-tailed adversarial training on CIFAR-100-LT (IR=100). **Left:** minimum inter-class angle θ_{\min} . **Right:** deviation from a margin-balanced ETF geometry. MCAT preserves larger angular margins and more balanced geometry.

are not driven solely by head classes but instead mitigate imbalance-induced bias. Results on Tiny-ImageNet-LT are deferred to Table 7 in Appendix C, where MCAT again shows the best results.

5.5 RQ3: Component Contribution and Sensitivity

Component ablations. Table 3 reports ablation results on CIFAR-100-LT. Adding the manifold constraint alone yields pronounced improvements in tail robustness, highlighting the importance of suppressing off-manifold adversarial drift. Geometric alignment alone substantially improves BA and BR, reflecting its role in alleviating imbalance-induced geometric bias. Combining both components yields the strongest and most consistent gains across all metrics.

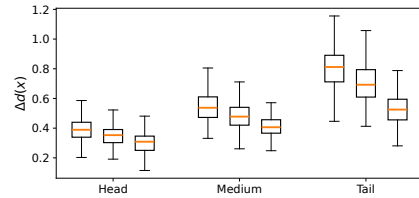


Figure 6: Off-manifold adversarial drift on CIFAR-100-LT (IR=100). Distributions of $\Delta d(x)$ for head, medium, and tail classes. For each class group, box plots from left to right correspond to PGD-AT, RoBal, and MCAT, respectively. MCAT suppresses drift, especially for tail classes.

Effect of λ (manifold constraint). Figures 4a and 4c show that λ controls a clear robustness–validity trade-off. With $\lambda = 0$, adversarial examples drift far off the class manifold and tail robustness drops. Increasing λ consistently suppresses drift and improves both overall and tail robustness, with gains saturating beyond a moderate range.

Effect of β (geometric alignment). Figures 4b and 4d show that increasing β enlarges the minimum inter-class angle θ_{\min} and improves Tail-AA robustness, with smooth and saturating trends consistent with the margin–robustness relationship in Theorem 1. This indicates that geometric alignment acts as a stable inductive bias rather than brittle tuning. As shown in Appendix Fig. 10b, moderate β improves tail robustness without degrading head-class performance, while overly large β leads to over-regularization.

Method	Clean \uparrow	PGD-20 \uparrow	AA \uparrow	BA \uparrow	BR (AA) \uparrow	Tail-AA \uparrow
Base AT	56.10 \pm 0.32	28.90 \pm 0.48	25.90 \pm 0.50	40.60 \pm 0.42	17.30 \pm 0.48	12.10 \pm 0.58
+ Manifold constraint ($\lambda > 0$)	56.30 \pm 0.30	30.40 \pm 0.46	27.60 \pm 0.48	42.10 \pm 0.40	19.40 \pm 0.45	15.80 \pm 0.52
+ Geometric alignment ($\beta > 0$)	56.50 \pm 0.29	31.20 \pm 0.44	28.60 \pm 0.46	45.30 \pm 0.38	21.80 \pm 0.42	14.90 \pm 0.50
MCAT (full)	62.30\pm0.24	37.10\pm0.40	34.60\pm0.44	51.80\pm0.30	27.40\pm0.36	20.00\pm0.44

Table 3: Ablation study of MCAT components on **CIFAR-100-LT (IR=100)**. We report clean accuracy, robust accuracy under PGD-20 and AutoAttack (AA), balanced accuracy (BA), balanced robustness (BR), and tail-class robustness under AutoAttack (Tail-AA). Results are reported as mean \pm std over three random seeds.

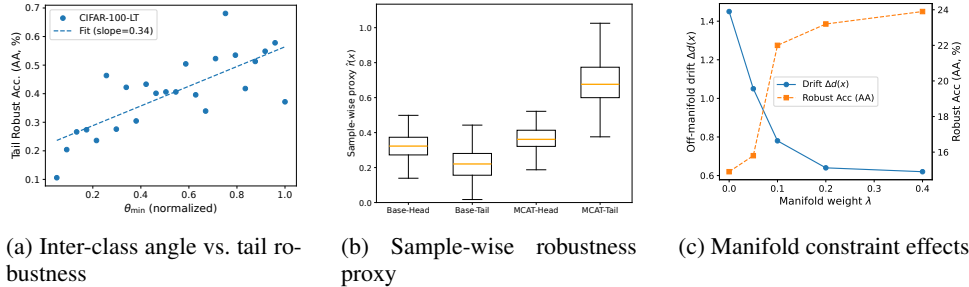


Figure 7: Theory-aligned empirical evidence on **CIFAR-100-LT (IR=100)**. (a) Larger minimum inter-class angle θ_{\min} correlates with stronger tail robustness. (b) MCAT shifts the tail-class distribution of the sample-wise robustness proxy $\hat{r}(x)$ toward larger values. (c) Increasing the manifold constraint weight λ jointly suppresses off-manifold drift and improves robust accuracy.

5.6 RQ4: Mechanism Verification and Theory Consistency

Imbalance-induced geometric bias and Off-manifold adversarial drift. Figure 5 reports geometry diagnostics. Baseline methods exhibit reduced inter-class angular separation and increased deviation from a margin-balanced ETF geometry. In contrast, MCAT preserves larger angular margins and maintains more balanced decision geometry. Figure 6 shows distributions of off-manifold drift. Standard adversarial training induces pronounced drift, especially for tail classes, whereas MCAT substantially suppresses drift across all class groups. In addition to quantitative diagnostics, we provide a qualitative case study by Figure 8 in Appendix C. Compared to Base AT, MCAT yields noticeably tighter and better-separated tail-class embeddings while preserving compact head-class structure, offering intuitive evidence of improved geometric balance.

Theory-aligned empirical evidence. Figure 7 provides theory-consistent observations: (i) larger minimum inter-class angles correlate with stronger tail robustness, (ii) MCAT shifts tail-class distributions of the sample-wise robustness proxy toward larger values, and (iii) increasing λ jointly suppresses drift and improves robustness. These results align with Theorems 1, 2, and Corollary 1. Additional per-class results are deferred to Figure 9 in Appendix C, where MCAT again shows the best results.

6 Related Work

General Long-Tailed Learning. Long-tailed learning in the standard setting has been widely studied through data augmentation, training paradigms, and representation rebalancing. Recent work leverages generative models for tail data

synthesis [Zhao *et al.*, 2024a; Shao *et al.*, 2024], controllable expert-based training [Zhao *et al.*, 2024b], and feature-space analyses that attribute tail failures to geometric distortion and representation collapse [Yi *et al.*, 2025; Sun *et al.*, 2025; Zhou *et al.*, 2024].

Long-Tailed Adversarial Robustness. Prior studies show that adversarial training disproportionately harms tail classes under imbalance. Existing solutions rely on margin or sampling rebalancing [Wu *et al.*, 2021; Liu *et al.*, 2022], staged or reweighted optimization [Li *et al.*, 2023; Yu-Hang *et al.*, 2025], and robustness distillation [Cho *et al.*, 2025], but do not explicitly regulate feature geometry or semantic validity of adversarial examples.

Geometry and Manifold Structure. Neural collapse and simplex ETF analyses highlight the role of balanced geometry in robustness [Papayan *et al.*, 2020; Cao *et al.*, 2025; Kothapalli, 2022; Zhu *et al.*, 2022], while manifold-based studies link adversarial vulnerability to off-manifold perturbations [Li *et al.*, 2025; Satou *et al.*, 2025; Zhang *et al.*, 2024]. Our work integrates these perspectives by jointly enforcing geometric balance and class-conditional manifold constraints for long-tailed adversarial training.

7 Conclusion

We proposed MCAT, a unified framework for long-tailed adversarial robustness that combines manifold-constrained adversarial training with ETF-inspired geometry regularization. We provided theoretical results connecting balanced geometry to robust margins and showing the benefit of constraining adversarial drift away from semantic low-density regions. Experiments on standard long-tailed benchmarks validate improved balanced robustness and tail performance under standard adversarial attacks.

References

- [Ahn *et al.*, 2023] Sumyeong Ahn, Jongwoo Ko, and Se-Young Yun. Cuda: Curriculum of data augmentation for long-tailed recognition. In *The Eleventh International Conference on Learning Representations*, 2023.
- [Cao *et al.*, 2025] Yang Cao, Yanbo Chen, and Weiwei Liu. Prevalence of simplex compression in adversarially robust neural networks. In *Proceedings of the National Academy of Sciences*, 2025.
- [Cho *et al.*, 2025] Seungju Cho, Hongsin Lee, and Chang-ick Kim. Long-tailed adversarial training with self-distillation. In *The Thirteenth International Conference on Learning Representations*, 2025.
- [Du *et al.*, 2023] Fei Du, Peng Yang, Qi Jia, Fengtao Nan, Xiaoting Chen, and Yun Yang. Global and local mixture consistency cumulative learning for long-tailed visual recognitions. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 15814–15823, 2023.
- [Gupta *et al.*, 2025] Sunny Gupta, Nikita Jangid, Shounak Das, and Amit Sethi. Fedtail: Federated long-tailed domain generalization with sharpness-guided gradient matching. *arXiv preprint arXiv:2506.08518*, 2025.
- [Jia *et al.*, 2022] 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*, pages 13398–13408, 2022.
- [Kothapalli, 2022] Vignesh Kothapalli. Neural collapse: A review on modelling principles and generalization. *arXiv preprint arXiv:2206.04041*, 2022.
- [Li *et al.*, 2021] Xiangxian Li, Haokai Ma, Lei Meng, and Xiangxu Meng. Comparative study of adversarial training methods for long-tailed classification. In *Proceedings of the 1st International Workshop on Adversarial Learning for Multimedia*, pages 1–7, 2021.
- [Li *et al.*, 2023] Guanlin Li, Guowen Xu, and Tianwei Zhang. Alleviating the effect of data imbalance on adversarial training. In *Advances in Neural Information Processing Systems*, 2023.
- [Li *et al.*, 2025] Zhiting Li, Shibai Yin, Tai-Xiang Jiang, Yexun Hu, Jia-Mian Wu, Guowei Yang, and Guisong Liu. Enhancing the adversarial robustness via manifold projection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 451–459, 2025.
- [Liu *et al.*, 2022] Bo Liu, Haoxiang Li, Hao Kang, Gang Hua, and Nuno Vasconcelos. Breadcrumbs: Adversarial class-balanced sampling for long-tailed recognition. In *European conference on computer vision*, pages 637–653, 2022.
- [Papayan *et al.*, 2020] Vardan Papayan, X. Y. Han, and David L. Donoho. Prevalence of neural collapse during the terminal phase of deep learning training. *Proceedings of the National Academy of Sciences*, 2020.
- [Ren *et al.*, 2020] Jiawei Ren, Cunjun Yu, Xiao Ma, Haiyu Zhao, Shuai Yi, et al. Balanced meta-softmax for long-tailed visual recognition. *Advances in neural information processing systems*, 33:4175–4186, 2020.
- [Satou *et al.*, 2025] Hana Satou, Alan Mitkiy, Emma Collins, and Finn Kingston. Geometrically regularized transfer learning with on-manifold and off-manifold perturbation. *arXiv preprint arXiv:2505.15191*, 2025.
- [Shao *et al.*, 2024] Jie Shao, Ke Zhu, Hanxiao Zhang, and Jianxin Wu. Diffult: Diffusion for long-tailed recognition without external knowledge. In *Advances in Neural Information Processing Systems*, 2024.
- [Sun *et al.*, 2025] 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*, 2025.
- [Wang *et al.*, 2020] 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*, 2020.
- [Wu *et al.*, 2020] Dongxian Wu, Shu-Tao Xia, and Yisen Wang. Adversarial weight perturbation helps robust generalization. *Advances in neural information processing systems*, 33:2958–2969, 2020.
- [Wu *et al.*, 2021] Tong Wu, Ziwei Liu, Qingqiu Huang, Yu Wang, and Dahua Lin. Adversarial robustness under long-tailed distribution. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8659–8668, 2021.
- [Xu *et al.*, 2021] Zhengzhuo Xu, Zenghao Chai, and Chun Yuan. Towards calibrated model for long-tailed visual recognition from prior perspective. *Advances in Neural Information Processing Systems*, 34:7139–7152, 2021.
- [Yi *et al.*, 2025] Lingjie Yi, Michael Yao, Weimin Lyu, Haibin Ling, Raphael Douady, and Chao Chen. Geometry of long-tailed representation learning: Rebalancing features for skewed distributions. In *International Conference on Learning Representations*, 2025.
- [Yu-Hang *et al.*, 2025] 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 *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 15476–15485, 2025.
- [Yue *et al.*, 2024] 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*, pages 24492–24501, 2024.
- [Zhang and Feng, 2024] Jinghao Zhang and Zhenhua Feng. Robust long-tailed image classification via adversarial feature re-calibration. In *Proceedings of the 19th International Joint Conference on Computer Vision, Imaging and*

Computer Graphics Theory and Applications, volume 2, pages 213–220, 2024.

- [Zhang *et al.*, 2019] 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 *Proceedings of the International Conference on Machine Learning*, 2019.
- [Zhang *et al.*, 2022] Jie Zhang, Lei Zhang, Gang Li, and Chao Wu. Adversarial examples for good: Adversarial examples guided imbalanced learning. In *2022 IEEE International Conference on Image Processing (ICIP)*, pages 136–140. IEEE, 2022.
- [Zhang *et al.*, 2023] Yifan Zhang, Bingyi Kang, Bryan Hooi, Shuicheng Yan, and Jiashi Feng. Deep long-tailed learning: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 45(9):10795–10816, 2023.
- [Zhang *et al.*, 2024] Wenjia Zhang, Yikai Zhang, Xiaoling Hu, Yi Yao, Mayank Goswami, Chao Chen, and Dimitris Metaxas. Manifold-driven decomposition for adversarial robustness. *Frontiers in Computer Science*, 5:1274695, 2024.
- [Zhang *et al.*, 2025] Chongsheng Zhang, George Almpandis, Gaojuan Fan, Binqun Deng, Yanbo Zhang, Ji Liu, Aouaidjia Kamel, Paolo Soda, and João Gama. A systematic review on long-tailed learning. *IEEE Transactions on Neural Networks and Learning Systems*, 2025.
- [Zhao *et al.*, 2024a] Qihao Zhao, Yalun Dai, Hao Li, Wei Hu, Fan Zhang, and Jun Liu. Ltgc: Long-tail recognition via leveraging llm-driven generated content. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.
- [Zhao *et al.*, 2024b] Zhe Zhao, Haibin Wen, Zikang Wang, Pengkun Wang, Fanfu Wang, Song Lai, Qingfu Zhang, and Yang Wang. Breaking long-tailed learning bottlenecks: A controllable paradigm with hypernetwork-generated diverse experts. In *Advances in Neural Information Processing Systems*, 2024.
- [Zhou *et al.*, 2024] Zihao Zhou, Siyuan Fang, Zijiang Zhou, Tong Wei, Yuanyu Wan, and Minling Zhang. Continuous contrastive learning for long-tailed semi-supervised recognition. In *Advances in Neural Information Processing Systems*, 2024.
- [Zhu *et al.*, 2022] Jianggang Zhu, Zheng Wang, Jingjing Chen, Yi-Ping Phoebe Chen, and Yu-Gang Jiang. Balanced contrastive learning for long-tailed visual recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6908–6917, 2022.

A Appendix: Proofs

A.1 Notation

We consider a classifier of the form

$$f_{\Theta}(x) = W\phi_{\Theta}(x),$$

where $W \in \mathbb{R}^{C \times m}$ is the linear classifier and $\phi_{\Theta}(x) \in \mathbb{R}^m$ is the feature representation. Let w_k denote the k -th row of W , and define the logit score for class k as

$$s_k(x) = w_k^{\top} \phi_{\Theta}(x).$$

The ℓ_{∞} adversarial ball is denoted by

$$\mathcal{B}_{\epsilon}(x) = \{x' \mid \|x' - x\|_{\infty} \leq \epsilon\}.$$

The robust risk is

$$R_{robust}(\theta) = \mathbb{E}_{(x,y) \sim \mathcal{D}} \left[\max_{x' \in \mathcal{B}_{\epsilon}(x)} \ell(f_{\Theta}(x'), y) \right],$$

where $\ell(\cdot, \cdot)$ denotes the classification loss.

A.2 Proof of Theorem 1

Assumptions. We assume normalized features and classifier weights, i.e.,

$$\|\phi_{\Theta}(x)\|_2 = 1, \quad \|w_k\|_2 = 1 \quad \forall k,$$

which can be enforced without loss of generality. We further assume that ϕ_{Θ} is L -Lipschitz under ℓ_{∞} perturbations, as stated in the theorem.

Proof. Fix a sample (x, y) such that

$$y = \arg \max_k s_k(x).$$

Define the logit margin

$$\gamma(x) = s_y(x) - \max_{k \neq y} s_k(x).$$

Step 1: Margin lower bound from ETF geometry. Since both w_k and $\phi_{\Theta}(x)$ are unit-norm, we have

$$s_k(x) = \cos(\angle(w_k, \phi_{\Theta}(x))).$$

Let

$$k^* = \arg \max_{k \neq y} s_k(x).$$

Under the approximate ETF assumption, the angle between w_y and w_{k^*} is at least θ_{min} .

The configuration minimizing the margin occurs when $\phi_{\Theta}(x)$ lies in the two-dimensional subspace spanned by w_y and w_{k^*} . Elementary geometric arguments then yield

$$\gamma(x) \geq \sin(\theta_{min}/2).$$

Step 2: Stability under adversarial perturbations. Let $x' = x + \delta$ with $\|\delta\|_{\infty} \leq \epsilon$. By the L -Lipschitz assumption,

$$\|\phi_{\Theta}(x') - \phi_{\Theta}(x)\|_2 \leq L\epsilon.$$

For any class k , we have

$$|s_k(x') - s_k(x)| = |w_k^{\top}(\phi_{\Theta}(x') - \phi_{\Theta}(x))| \leq L\epsilon.$$

Therefore, the margin at x' satisfies

$$\gamma(x') \geq \gamma(x) - 2L\epsilon.$$

Step 3: Robustness condition. Combining the two bounds,

$$\gamma(x') \geq \sin(\theta_{min}/2) - 2L\epsilon.$$

Thus, if

$$\epsilon < \frac{\sin(\theta_{min}/2)}{L},$$

the margin remains strictly positive and the predicted label is invariant to all perturbations in $\mathcal{B}_{\epsilon}(x)$. This proves the theorem. \square

A.3 Proof of Corollary 1

Proof. Fix a sample (x, y) and define the logit score $s_k(x) = w_k^{\top} \phi_{\Theta}(x)$. Let

$$\gamma(x) = s_y(x) - \max_{k \neq y} s_k(x)$$

denote the logit margin at x .

Consider any perturbation $x' = x + \delta$ with $\|\delta\|_{\infty} \leq r$. By the L -Lipschitz assumption on ϕ_{Θ} ,

$$\|\phi_{\Theta}(x') - \phi_{\Theta}(x)\|_2 \leq Lr.$$

Assuming $\|w_k\|_2 = 1$ for all k , we have for each class k ,

$$|s_k(x') - s_k(x)| = |w_k^{\top}(\phi_{\Theta}(x') - \phi_{\Theta}(x))| \leq Lr.$$

Let $k^* = \arg \max_{k \neq y} s_k(x)$. Then

$$s_y(x') - s_{k^*}(x') \geq (s_y(x) - Lr) - (s_{k^*}(x) + Lr) = \gamma(x) - 2Lr.$$

Therefore, if $r \leq \gamma(x)/(2L)$, the right-hand side remains non-negative, implying

$$s_y(x') \geq s_{k^*}(x') \geq s_k(x') \quad \forall k \neq y,$$

and the predicted label is unchanged within the ℓ_{∞} ball of radius r . Thus the sample-wise robust radius satisfies

$$r(x) \geq \frac{\gamma(x)}{2L}. \quad \square$$

A.4 Proof of Theorem 2

Assumptions. We assume that the per-sample loss is bounded,

$$0 \leq \ell(f_{\Theta}(x), y) \leq \ell_{max},$$

and that for each class y , the data distribution is supported on a semantic manifold \mathcal{M}_y in feature space, while regions far from \mathcal{M}_y carry negligible probability mass.

Proof. We formalize the intuition described in the main text. Under long-tailed distributions, adversarial optimization may place excessive emphasis on perturbations whose features drift far away from the semantic manifold \mathcal{M}_y of a tail class. Although such off-manifold perturbations can induce high classification loss, they lie in low-density regions that are weakly supported by the data distribution and therefore do not meaningfully contribute to robustness on the semantic support.

Fix a sample (x, y) and consider the inner maximization in the robust risk. Let

$$x^* = \arg \max_{x' \in \mathcal{B}_\epsilon(x)} \ell(f_\Theta(x'), y).$$

Step 1: Decomposition of robust risk. To make the above distinction precise, we decompose the robust risk into on-manifold and off-manifold contributions:

$$R_{robust}(\Theta) = R_{on}(\Theta) + R_{off}(\Theta),$$

where R_{on} corresponds to adversarial examples whose features remain within a neighborhood of the class manifold \mathcal{M}_y , and R_{off} corresponds to adversarial examples whose features drift far away from \mathcal{M}_y .

By the manifold support assumption, off-manifold regions contribute negligible probability mass. Therefore, there exists a constant $\rho \ll 1$ such that

$$R_{off}(\Theta) \leq \rho \ell_{max}.$$

Step 2: Control of on-manifold risk via manifold-constrained objective. Recall the MCAT objective without the geometric regularizer:

$$R_{MCAT}(\Theta) = \mathbb{E}_{(x,y) \sim \mathcal{D}} \left[\max_{\|\delta\|_\infty \leq \epsilon} (\ell(f_\Theta(x + \delta), y) + \lambda d_{\mathcal{M}_y}(\phi_\Theta(x + \delta))) \right].$$

For adversarial examples whose features lie within a bounded neighborhood of \mathcal{M}_y , the manifold deviation term is uniformly bounded:

$$d_{\mathcal{M}_y}(\phi_\Theta(x')) \leq C.$$

As a result, for such on-manifold adversarial points,

$$\begin{aligned} \ell(f_\Theta(x'), y) &\leq \ell(f_\Theta(x'), y) + \lambda d_{\mathcal{M}_y}(\phi_\Theta(x')) \\ &\leq R_{MCAT}(\Theta) + \frac{C}{\lambda}. \end{aligned}$$

Step 3: Combining the bounds. Taking expectation over (x, y) and combining the on-manifold and off-manifold contributions yields

$$R_{robust}(\Theta) \leq R_{MCAT}(\Theta) + \frac{C}{\lambda} + \rho \ell_{max}.$$

Since ρ is negligible under the manifold support assumption, we conclude that

$$R_{robust}(\Theta) \leq R_{MCAT}(\Theta) + O(\lambda^{-1}),$$

which completes the proof. \square

B Generator Architecture

Each class-conditional generator G_y is implemented as a lightweight multilayer perceptron operating in the classifier feature space. Given a latent code $z \in \mathbb{R}^{d_z}$ sampled from a standard Gaussian, G_y outputs a feature vector in \mathbb{R}^{d_f} , where d_f is the dimension of the penultimate-layer features of the backbone.

Concretely, G_y consists of three fully connected layers with widths $d_z \rightarrow 1024 \rightarrow 1024 \rightarrow d_f$. ReLU activations are applied after the first two layers, and the output layer is linear. No batch normalization or dropout is used. All generators share the same architecture but are trained independently for each class. Unless otherwise specified, we set $d_z = 128$ and $d_f = 512$ for CIFAR-based experiments.

C More Experimental Results

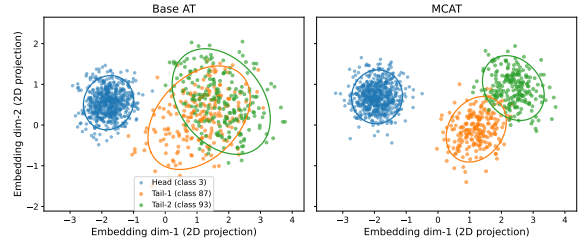


Figure 8: Case study on CIFAR-100-LT (IR=100) showing 2D embedding projections of one head and two tail classes. Compared to Base AT, MCAT yields tighter and better-separated tail clusters while maintaining compact head representations.

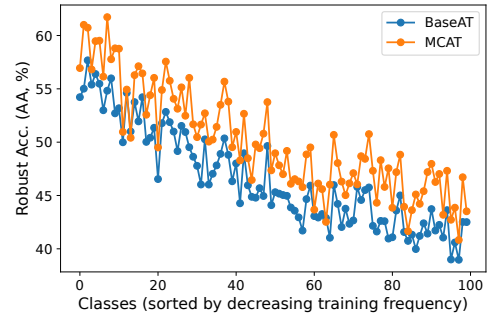
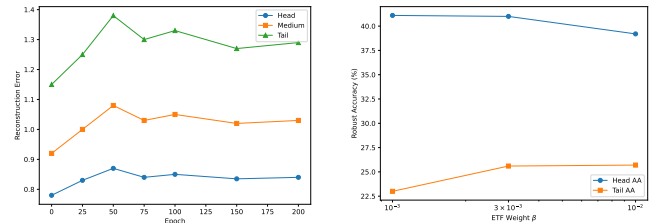


Figure 9: Robust accuracy over all classes under AutoAttack (AA) on CIFAR-100-LT (IR=100), plotted against the sorted class frequency rank.



(a) Reconstruction error $\|\phi_\Theta(x) - G_y(z^*)\|_2$ over training epochs on CIFAR-100-LT (IR=100). The error exhibits mild non-monotonic fluctuations due to feature adaptation, but remains stable overall and does not diverge for tail classes. (b) Effect of ETF regularization weight β on robust accuracy for head and tail classes. Moderate β improves tail robustness without degrading head performance, while excessively large β leads to over-regularization.

Figure 10: Manifold stability and geometric alignment trade-off on CIFAR-100-LT (IR=100). Left: reconstruction error of frozen class-conditional generators remains stable throughout adversarial training. Right: ETF-inspired geometric alignment improves tail robustness without sacrificing head-class performance under moderate regularization.

Method	IR=10	IR=20	IR=50	IR=100
RoBal	31.80	29.10	26.40	25.90
Self-Distill	33.40	31.20	29.10	28.10
AT-BSL	32.90	30.80	28.70	27.30
MCAT (ours)	34.60	33.10	32.00	31.50

Table 4: AutoAttack robustness (%) on CIFAR-100-LT under varying imbalance ratios. MCAT exhibits consistently higher robustness and degrades more gracefully as imbalance severity increases.

Method	IR=10	IR=20	IR=50	IR=100
RoBal	19.60	17.40	14.80	13.20
Self-Distill	21.30	19.50	17.40	16.10
AT-BSL	20.80	18.90	16.90	15.50
MCAT (ours)	26.20	23.60	22.10	21.05

Table 5: Tail-class AutoAttack robustness (%) on CIFAR-100-LT under varying imbalance ratios. MCAT consistently improves tail robustness and degrades more gracefully as imbalance severity increases.

T_z	Overall AA (%)	Tail AA (%)	Relative Train Time
0	32.4	18.7	1.00
1	34.9	21.5	1.07
3	36.8	24.3	1.14
5	37.5	25.6	1.18
8	37.6	25.7	1.26

Table 6: Sensitivity to the number of latent optimization steps T_z on CIFAR-100-LT (IR=100).

Method	BA \uparrow	BR (AA) \uparrow	Tail-PGD \uparrow	Tail-AA \uparrow
PGD-AT	31.20 \pm 0.60	11.40 \pm 0.65	9.30 \pm 0.70	8.10 \pm 0.75
TRADES	32.10 \pm 0.58	12.80 \pm 0.62	10.20 \pm 0.68	8.90 \pm 0.73
MART	31.80 \pm 0.59	13.20 \pm 0.63	10.60 \pm 0.69	9.20 \pm 0.74
AWP	32.90 \pm 0.55	14.10 \pm 0.60	11.30 \pm 0.65	10.10 \pm 0.70
RoBal	35.60 \pm 0.52	15.90 \pm 0.58	12.80 \pm 0.62	11.20 \pm 0.67
REAT	36.80 \pm 0.50	17.10 \pm 0.56	13.90 \pm 0.60	12.30 \pm 0.65
TAET	37.40 \pm 0.48	17.80 \pm 0.55	14.60 \pm 0.59	12.90 \pm 0.64
Self-Distill	37.90 \pm 0.49	18.20 \pm 0.54	15.00 \pm 0.58	13.40 \pm 0.63
AT-BSL	37.60 \pm 0.50	17.90 \pm 0.55	14.70 \pm 0.59	13.10 \pm 0.64
MCAT (ours)	42.30\pm0.45	22.60\pm0.50	18.90\pm0.54	16.80\pm0.58

Table 7: Balanced and tail robustness on **Tiny-ImageNet-LT (IR=100)**. We report balanced accuracy (BA) and balanced robustness (BR), defined as average per-class accuracy under clean evaluation and AutoAttack (AA), respectively, together with tail-class robust accuracy under PGD-20 (Tail-PGD) and AutoAttack (Tail-AA). Results are reported as mean \pm std over three random seeds.

D Hyperparameter Settings

Category	Hyperparameter	Value
Training	Optimizer	SGD with momentum
	Momentum	0.9
	Weight decay	5×10^{-4}
	Batch size	128
	Training epochs	200
	Learning rate schedule	Cosine decay
Adversarial Setup	Threat model	ℓ_∞
	Perturbation budget ϵ	8/255
	Step size α	2/255
	PGD steps	10 (train), 20 (eval)
Backbone	Architecture	ResNet-18
	Initialization	He initialization
	Normalization	BatchNorm
Long-tailed Setting	Imbalance type	Exponential
	Imbalance ratio (IR)	{10, 50, 100}
	Sampling strategy	Class-uniform
	Evaluation metric	Overall / Many / Medium / Few
MCAT (Ours)	Manifold penalty weight λ_{man}	0.1
	Geometric alignment weight β	3×10^{-3}
	Equivalent λ_{geom} in implementation	0.01
	Manifold distance metric	ℓ_2 in feature space
	Manifold update frequency	Every iteration
	ETF target dimension	Equal to number of classes
Baselines	AT / TRADES	Official recommended settings
	RoBal	Margin reweighting as in [Wu <i>et al.</i> , 2021]
	REAT	Loss reweighting as in [Li <i>et al.</i> , 2023]
	TAET	Two-stage schedule as in [Yuhang <i>et al.</i> , 2025]
	Distillation-based	Teacher trained on balanced AT

Table 8: Hyperparameter settings for MCAT and baseline adversarial training methods. β denotes the ETF-inspired geometric alignment weight used in Eq. (3). Unless otherwise specified, all methods share the same backbone architecture and training protocol.