

Resurrecting the Tail: A Generative Curriculum Framework for Self-Supervised Adversarial Robustness on Long-Tailed Distributions

Abstract Adversarial Training(AT) on long-tailed distributions triggers catastrophic vulnerability in tail classes. Prior work has predominantly relied on passive loss reweighting to compensate for data imbalance, leaving the underlying scarcity of tail-class samples unaddressed at the data level. In this work, we propose an active manifold-completion strategy that integrates conditional diffusion generation with self-supervised adversarial contrastive learning. Our framework introduces three core innovations. First, we exploit diffusion-based generative priors to repair the fragmented feature manifolds of minority classes by synthesizing on-manifold samples conditioned on tail-class labels. Second, we propose LT-SSL-AT, a dual-path self-supervised adversarial framework that decouples semantic invariance learning (clean-augmented views) from robustness learning (clean vs. adversarial pairs). Third, we introduce a curriculum scheduling weight $s(t)$ that progressively shifts optimization from an Self-Supervised Learning(SSL)-only warm-up phase to joint adversarial alignment, preventing representation collapse. On CIFAR-10-LT (IR=10), CD-SSL achieves 35.12% Balanced Robustness under AutoAttack, surpassing the state-of-the-art TAET baseline by 4.58 percentage points. Extensive experiments across imbalance ratios 10, 20, 50, and 100 on three benchmarks (CIFAR-10-LT, CIFAR-100-LT, and MedMNIST) confirm consistent gains over both long-tailed and adversarial baselines.

Keywords Adversarial Robustness, Long-Tailed Distribution, Self-Supervised Adversarial Training, Diffusion Models, Generative Manifold Repair

1 Introduction

Deep neural networks are notoriously vulnerable to adversarial attacks [1, 2], and Adversarial Training (AT) remains

the primary defense paradigm against such threats [1]. In practice, however, real-world datasets are seldom balanced: they frequently exhibit long-tailed distributions in which a handful of head classes are richly represented while minority classes suffer from severe sample scarcity [3–5]. This co-occurrence of adversarial threats and distributional imbalance creates a compound problem—adversarial training disproportionately degrades the robustness of minority classes [6]—a phenomenon that introduces existential risk in safety-critical applications such as autonomous driving (recognition of rare vehicle types) and medical diagnosis (identification of uncommon diseases).

1.1 The Fundamental Bottleneck

Existing defenses approach this compound problem primarily at the *optimization level*, relying on sophisticated loss reweighting or knowledge distillation. Representative methods include AT-BSL [7], TAET [8], and Self-Distillation [9]. While these techniques achieve notable improvements in per-class robustness, they share a common limitation: they attempt to extract robust features from fundamentally insufficient data through algorithmic ingenuity alone, without ever addressing the underlying data deficit.

We argue that this vulnerability is, at its core, a *representation-level* problem [32]. Adversarial robustness requires the model to maintain stable predictions within an ℓ_p -ball around every training point [1]. Because minority classes contain very few examples, their feature manifolds are inherently sparse and fragmented; the underlying data manifold contains structural holes that no amount of loss-function engineering can bridge, regardless of how carefully the class weights are tuned.

Recent Self-Supervised Adversarial Training (SSL-AT) frameworks [10, 23] learn invariant representations that simultaneously resist data augmentations and adversarial perturbations. Yet SSL-AT inherits the same manifold fragmentation problem on long-tailed data: sparse minority samples provide insufficient positive pairs for contrastive learning, causing feature encoders to collapse minority instances into degenerate, near-identical points in the embedding space.

• Yilin Zou, Zongqing Zi, Zepeng Lin, Jiayuan Liu and Liang Tan are with the School of Computer Science, Sichuan Normal University, Chengdu and 610101, China. E-mail: jkxytl@sicnu.edu.cn
• Zhong Huang and Fading Zhao is with the School of Communication and Information Engineering, Sichuan Normal University, Chengdu and 610101, China.
• Kun She is with the College of Information and Software Engineering, University of Electronic Science and Technology of China, Chengdu and 610054, China. E-mail: kun@uestc.edu.cn
Manuscript received: 20xx-xx-xx; revised: 20xx-11-xx; accepted: 20xx-00-xx

1.2 Our Solution: Generative Manifold Repair

We propose the **CD-SSL** framework, which tackles long-tailed adversarial robustness through *active generative manifold repair* rather than passive adaptation to data scarcity. Concretely, we reconstruct the geometric structure of minority feature spaces using a class-conditional diffusion model [12]. Diffusion models implicitly learn the score function $\nabla_x \log p(x)$ of the data distribution and can synthesize interpolative samples that lie on the true data manifold [13]. These synthetic samples fill the voids between sparse minority points, enabling SSL-AT to operate on a continuous, well-structured feature space rather than on disconnected fragments.

Our research presents three main contributions:

(1) We empirically demonstrate that diffusion-based data synthesis is a *necessary*, not merely auxiliary, component of SSL-AT under long-tailed distributions: removing the diffusion module reduces Balanced Robustness by 3.27 percentage points.

(2) We introduce the **LT-SSL-AT** architecture, a dual-path framework that explicitly decouples semantic invariance learning from adversarial robustness learning, thereby avoiding the well-documented gradient interference between large semantic-preserving augmentations and small ℓ_∞ -bounded perturbations.

(3) We design a dynamic curriculum schedule $s(t)$ that first establishes semantic anchors through a pure-SSL warm-up phase, then progressively tightens the decision boundary by introducing adversarial alignment, preventing the representation collapse that afflicts naive joint optimization.

On CIFAR-10-LT (IR=10), CD-SSL achieves 35.12% Balanced Robustness under AutoAttack [14], surpassing the CVPR 2025 TAET baseline by 4.58 percentage points and the ICLR 2025 Self-Distillation baseline by 0.84 percentage points. Across all evaluated datasets and imbalance ratios, CD-SSL demonstrates the best adversarial robustness while remaining competitive on clean accuracy.

2 Related Work

2.1 Adversarial Training on Long-Tailed Data

Adversarial Training (AT) is an effective defense against adversarial attacks [1], formulated as a challenging min-max optimization problem [15]. In this framework, the inner loop generates adversarial examples, while the outer loop updates the model to correctly classify these perturbed inputs, as formalized in Eq. (1):

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

Here, f_{θ} denotes the neural network, θ its parameters, $\mathcal{L}$ the loss function, and $x + \delta$ the adversarial sample.

Most prior studies focus on adversarial training with balanced datasets [16, 41], whereas real-world data often exhibit long-tailed distributions [17, 18], which substantially undermine robustness. Wu et al. [6] first identified a critical vulnerability: adversarial training disproportionately harms minority classes, a phenomenon termed the magnification effect of robustness sacrifice.

Several methods have been proposed to address this challenge. The RoBal framework leverages a two-stage rebalancing strategy with cosine classifiers, improving both natural and robust accuracy [6]. Similarly, AT-BSL [7] introduces Balanced Softmax Loss (BSL) during adversarial training, achieving comparable performance while reducing computational cost. BSL compensates for minority classes at the loss level by adjusting logits based on class frequencies, as shown in Eq. (2):

$$\mathcal{L}_{BSL} = -\log \left(\frac{n_y^{\tau_b} \cdot e^{z_y}}{\sum_i n_i^{\tau_b} \cdot e^{z_i}} \right) \quad (2)$$

where n_i is the number of samples in class i and z_i the corresponding logit.

More recently, TAET [8] further mitigates robust overfitting through a two-stage training paradigm incorporating a Hierarchical Equalization Loss, which dynamically compensates disadvantaged minority classes. Despite these advances, existing methods primarily address data imbalance through loss function or training adjustments. The core structural challenge—the absolute scarcity of tail-class samples—remains unresolved [50].

2.2 Self-Supervised Adversarial Training

Self-supervised learning (SSL) enables the extraction of effective feature representations from fully unlabeled data [19], with contrastive learning being a prominent paradigm [20]. In this framework, positive and negative sample pairs are constructed: positive pairs, typically formed from different augmentations of the same image, are pulled closer in feature space, while negative pairs are pushed apart [21, 22, 33]. Recently, SSL has been combined with adversarial training to form self-supervised adversarial training [23, 35], where a clean sample x and its adversarial counterpart $x + \delta$ constitute a positive pair. A tailored contrastive loss encourages the model to learn robust feature representations that remain insensitive to adversarial perturbations.

The RoCL framework applies this idea to unlabeled datasets by maximizing the similarity between features of random

augmentations and adversarial perturbations [23]. Similarly, the ACL framework enforces consistency across features under both augmentations and adversarial attacks, achieving state-of-the-art robustness [10]. These works highlight a key principle: models can exploit the inherent structural information of individual samples to guide robust feature learning, which is particularly effective in unsupervised settings.

However, existing methods have been validated exclusively on balanced datasets. Directly applying these paradigms to long-tailed data is problematic, as minority classes contain extremely sparse samples. This scarcity often leads to representation collapse in the early training stages, causing the model to converge to trivial feature mappings where diverse inputs are projected to nearly identical points in feature space.

2.3 Conditional Diffusion Generation

Diffusion models are highly powerful generative architectures capable of synthesizing high-quality samples that closely match real data distributions [24]. These models learn a gradual denoising process [25], where the standard training objective minimizes the discrepancy between the injected noise ϵ and the model-predicted noise ϵ_θ , as formalized in Eq. (3):

$$\mathcal{L}_{\text{diffusion}} = \mathbb{E}_{x_0, \epsilon, t} [|\epsilon - \epsilon_\theta(x_t, t)|^2] \quad (3)$$

Here, x_0 denotes the original clean sample, and x_t is its noise-corrupted version at timestep t .

Conditional diffusion models extend this framework by incorporating additional information, such as class labels, to guide the generative process [12, 26, 27]. This capability enables targeted synthesis of specific categories, providing a promising solution to extreme data scarcity. Recent studies have applied diffusion models to long-tailed data augmentation [28, 49]. For example, DiffuLT trains a conditional diffusion model from scratch on imbalanced long-tailed datasets, generating novel proxy samples for minority classes to construct more balanced training sets [13]. LDMLR further improves computational efficiency by performing diffusion in a compressed latent space [28, 30], while DiffAugment demonstrates the applicability of diffusion models to complex long-tailed visual tasks such as visual relationship recognition [29]. Collectively, these studies highlight the potential of generative diffusion processes to effectively supplement minority categories and alleviate fundamental data imbalances, providing a solid foundation for our proposed framework.

3 Preliminaries: Analyzing the Limitations of TAET

Before introducing our method, we first analyze the TAET framework, which represents the current state-of-the-art approach [8]. Through this analysis, we identify several fundamental limitations that motivate the design of our proposed method.

3.1 Core Mechanisms of TAET

Hierarchical Equalization Loss (HEL). TAET introduces a hierarchical equalization strategy to improve the robustness of minority classes during adversarial training [8]. Specifically, the framework dynamically identifies underperforming classes and compensates them through three collaborative loss components, as defined in Eq. (4).

$$\mathcal{L}_{\text{HEL}} = \alpha \mathcal{L}_{\text{BCL}} + \beta \mathcal{L}_{\text{HDL}} + \gamma \mathcal{L}_{\text{RCEL}} \quad (4)$$

Here, the Balanced Cross-Class Loss (BCL) balances optimization across different classes, the Hierarchical Deviation Loss (HDL) penalizes inter-class disparities, and the Rare Class Emphasis Loss (RCEL) assigns larger weights to high-loss minority classes.

Two-Stage Training. TAET further adopts a two-stage training paradigm [8]. In the first stage, the model is trained on clean data to learn stable feature representations. In the second stage, adversarial training is introduced together with HEL optimization. This staged strategy simplifies the optimization process and alleviates robust overfitting to some extent.

3.2 Fundamental Limitations

Unresolved Data Scarcity. Despite their effectiveness, both HEL and the two-stage strategy primarily operate at the optimization level. These mechanisms encourage the model to pay greater attention to tail classes through adaptive loss reweighting, yet they do not address the fundamental issue of insufficient tail-class data. In long-tailed settings, minority classes often contain only a limited number of samples with severely restricted diversity. However, adversarial training inherently requires sufficient data coverage to accurately characterize the perturbed feature distribution of each class. Under extreme scarcity, even sophisticated loss designs cannot establish reliable robust decision boundaries. Consequently, TAET essentially adapts to data imbalance rather than resolving the underlying lack of data.

Augmentation Saturation. Existing methods heavily rely on conventional data augmentation techniques, such as random cropping, flipping, and color jittering, to mitigate robust

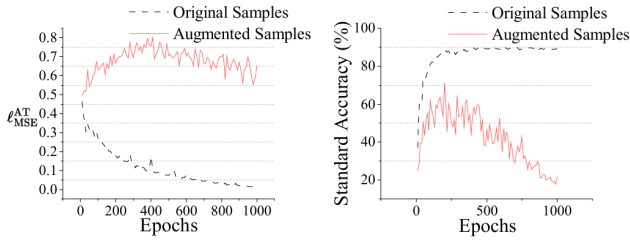


Fig. 1 Left: Augmented samples increase adversarial training difficulty. Right: Adversarial training instability disrupts self-supervised representation learning, whereas raw samples alleviate this issue.

overfitting. However, these augmentations provide only limited semantic diversity, which is particularly problematic for severely underrepresented tail classes. Prior studies have demonstrated that the effectiveness of adversarial training is closely related to data diversity [31, 38]. During training, the model repeatedly generates adversarial perturbations from a small number of tail-class samples combined with simple augmentations. As a result, robustness improvements quickly saturate and may even deteriorate in later training stages. As illustrated in Figure 1, this phenomenon reveals a fundamental contradiction: current frameworks strongly depend on data diversity, yet lack the capability to generate it effectively.

To address these limitations, we propose the CD-SSL framework, which integrates generative manifold repair with curriculum-based self-supervised adversarial training.

4 Methodology

Deep learning models exhibit severe adversarial vulnerability under long-tailed data distributions [1, 4]. To address this compound challenge, we propose the **CD-SSL** framework, a synergistic two-stage pipeline that simultaneously corrects class imbalance at the data level and learns adversarially robust representations at the feature level.

The pipeline proceeds as follows. In **Stage 1**, we augment the original long-tailed dataset $\mathcal{D}_{\text{LT}}$ by training a class-conditional diffusion model [13] to synthesize realistic tail-class samples, producing a more balanced dataset $\mathcal{D}_{\text{bal}} = \mathcal{D}_{\text{LT}} \cup \mathcal{D}_{\text{syn}}$. In **Stage 2**, we train a feature encoder on $\mathcal{D}_{\text{bal}}$ using the LT-SSL-AT framework—a self-supervised adversarial contrastive learning paradigm that learns representations simultaneously insensitive to semantic augmentations and adversarial perturbations. Figure 2 illustrates the complete framework.

4.1 Framework Design

CD-SSL comprises two main stages, as illustrated in Figure 2.

Stage 1: Conditional Diffusion-based Tail Data Balancing. The primary goal of this stage is to resolve the

fundamental data scarcity of tail classes [6]. We train a conditional diffusion model [24] on $\mathcal{D}_{\text{LT}}$ to synthesize high-quality, class-specific tail samples. These synthetic images are merged with the original dataset to form a more balanced training set $\mathcal{D}_{\text{bal}}$ [17].

Stage 2: Self-Supervised Adversarial Contrastive Learning. Operating on $\mathcal{D}_{\text{bal}}$, we design the LT-SSL-AT framework to jointly optimize self-supervised representation learning and adversarial training [1, 19]. The system constructs three views of each input: two clean augmented views and one adversarial view. A contrastive loss then encourages the encoder to learn robust representations by pulling clean and adversarial views together in the embedding space [23].

4.2 Stage 1: Conditional Diffusion-based Tail Data Balancing

Long-tailed datasets exhibit severe class imbalance: a small number of head classes contain abundant samples while the majority of tail classes remain severely underrepresented [17]. This imbalance fundamentally degrades model performance on minority categories due to insufficient data coverage and diversity. To resolve this at the data level, we introduce a conditional diffusion generation module [12].

4.2.1 Conditional Diffusion Model Training

We train a class-conditional Denoising Diffusion Probabilistic Model (cDDPM) [26] on the original long-tailed dataset $\mathcal{D}_{\text{LT}}$. The forward process gradually corrupts a clean image x with Gaussian noise; the reverse process, parameterized by a U-Net $\epsilon_{\theta}(x_t, t, y)$ [34], denoises x_t back to x conditioned on the class label y . Despite the imbalance of $\mathcal{D}_{\text{LT}}$, the cDDPM recovers approximate score functions for tail classes because it shares all U-Net parameters across classes, with only the conditioning signal y differing. Complete implementation details—including the training schedule, optimizer, learning rate, EMA decay, and classifier-free guidance scale—are provided in Appendix A.

4.2.2 Synthetic Tail Sample Generation and Data Merging

We identify tail classes y_{tail} whose sample counts fall below a threshold T_{gen} , and use the trained cDDPM to generate synthetic samples for these classes by reverse-diffusing from Gaussian noise conditioned on y_{tail} . The resulting set $\mathcal{D}_{\text{syn}}$ is merged with $\mathcal{D}_{\text{LT}}$ to form the balanced dataset: $\mathcal{D}_{\text{bal}} = \mathcal{D}_{\text{LT}} \cup \mathcal{D}_{\text{syn}}$ [51].

Choice of T_{gen} . We set $T_{\text{gen}} = 100$ on CIFAR-10-LT (IR=10), which corresponds to the size of the median class under exponential-decay imbalance and matches the smallest

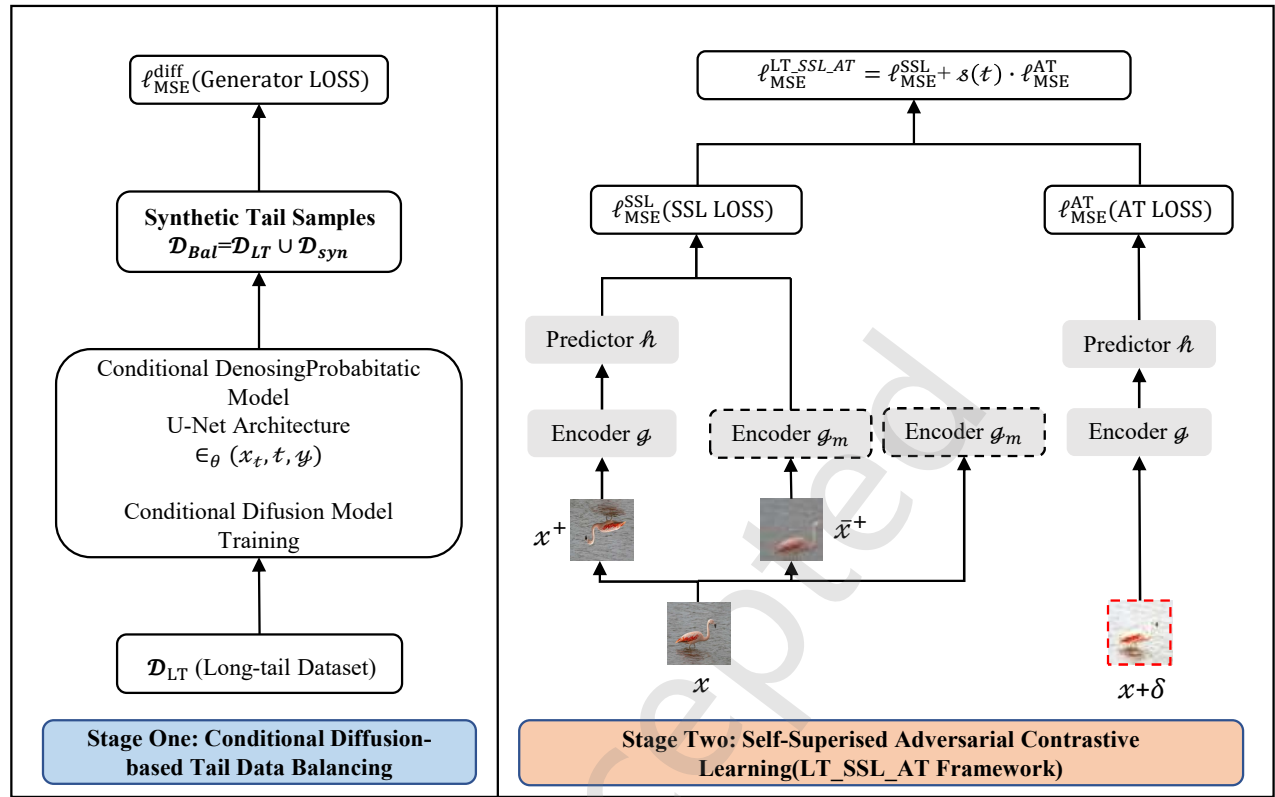


Fig. 2 Illustration of the proposed CD-SSL framework. Stage One utilizes a conditional diffusion model. This model generates synthetic samples for tail classes. This process merges $\mathcal{D}_{LT}$ and $\mathcal{D}_{syn}$ into a balanced dataset $\mathcal{D}_{bal}$. Stage Two utilizes the LT_SSL_AT framework. This framework contains dual learning paths. The SSL path processes augmented views (x^+). This path uses online and momentum encoders. The AT path handles adversarial samples ($x + \delta$). The framework optimizes both paths jointly. This optimization uses a dynamic scheduling weight $s(t)$.

size at which empirical mean-per-class loss for AT becomes stable in our pilot runs. The synthesis target is to bring every class to exactly T_{gen} samples, so that the post-merge distribution remains slightly head-heavy (no class is artificially boosted beyond its mid-range counterparts) while every tail class receives enough samples to populate its perturbation neighbourhood. We provide a sensitivity study over $T_{gen} \in \{50, 100, 200, 500\}$ in Table 5 (Appendix A.4).

Sample-quality filtering. To prevent the introduction of low-quality or ambiguous samples, we apply a two-stage filter to $\mathcal{D}_{syn}$ before merging: (i) *Confidence filter* — a classifier pretrained on $\mathcal{D}_{LT}$ must assign a top-1 probability ≥ 0.7 to the conditioning class y ; samples below this threshold are discarded. (ii) *Diversity filter* — among the surviving samples, we apply farthest-point sampling in the pretrained classifier’s penultimate-layer feature space, which biases the retained subset toward feature-space coverage rather than redundant prototypes. Filter thresholds and pseudocode are provided in Appendix A.3.

4.2.3 Diversity Guarantee for Generated Samples

A central concern for any generative augmentation pipeline is whether the synthesised samples introduce genuine diversity or merely replicate a few high-likelihood modes (mode collapse), which would expose the downstream model to an over-fitting risk. We address this concern via three complementary mechanisms:

(i) *Stochastic latents and varied guidance.* For each tail class y , we draw N independent Gaussian latents and use a moderate classifier-free-guidance scale $w \in [1.5, 3.0]$ sampled per generation, which empirically balances label-fidelity and visual variability.

(ii) *Sub-mode coverage check.* Before generation, we cluster the real tail-class samples (k-means in the pretrained feature space) into k sub-modes ($k = \lceil n_y/5 \rceil$). After generation, we verify that every sub-mode is represented by at least one synthetic sample passing the confidence filter; otherwise, generation is repeated with sub-mode-conditioned latents until coverage is achieved.

(iii) *Quantitative diversity verification.* Following

Kynkäänniemi et al. (2019) [11], we report Improved Recall and intra-class feature variance on D_{syn} versus D_{LT} for the five smallest tail classes. Table 6 (Appendix A.4) summarises these statistics.

Together, these mechanisms ensure that D_{syn} does not merely reproduce a few prototypes of each tail class but instead populates the full intra-class manifold, which is the geometric prerequisite for the manifold-repair claim of our paper.

4.3 Stage 2: Self-Supervised Adversarial Contrastive Learning

4.3.1 Architecture and Core Idea: Dual-Path Decoupling

Most prior SSL-AT methods apply adversarial perturbations on top of strongly augmented views, forcing the encoder to handle two heterogeneous input distortions simultaneously: large semantic-preserving augmentations (random crop, colour jitter) and small ℓ_∞ -bounded perturbations. As argued in [6, 37], the strong-augmentation gradient tends to dominate and mask the small-perturbation gradient, slowing robustness convergence. LT-SSL-AT resolves this conflict by routing the two signals through parallel, decoupled paths.

Self-Supervised Learning (SSL) Path. Given an image $x \sim D_{\text{bal}}$, we generate two independent augmented views $(x^+, \bar{x}^+)$ using random crop, colour jitter, and Gaussian blur [19], and use a BYOL-style architecture [20] consisting of an online encoder g , a predictor h , and a momentum encoder g_m updated by exponential moving average. The SSL loss Eq.5 minimises the MSE between the online prediction $h(g(x^+))$ and the momentum target $g_m(\bar{x}^+)$.

$$\mathcal{L}_{\text{SSL}} = \mathbb{E}_{x \sim D_{\text{bal}}} [\|h(g(x_1)) - g_m(x_2)\|_2^2] \quad (5)$$

Why BYOL rather than SimCLR or MoCo? We deliberately chose BYOL for three reasons that are particularly important under long-tailed adversarial training:

(a) *No reliance on negative samples.* SimCLR and MoCo construct contrastive losses against negatives drawn from the current mini-batch or a large memory bank. Under a long-tailed distribution, the negative pool is dominated by head-class samples; this causes tail-class anchors to be repulsed primarily from head prototypes, an effect that distorts the learnt geometry exactly where we need it to be most accurate. BYOL’s purely predictive objective avoids this pathology.

(b) *Asymmetric architecture aligns with the AT path.* The online–target asymmetry of BYOL (predictor on the online branch only) is structurally identical to our AT objective in Eq.6, where $h(g(x))$ is aligned to $g_m(x + \delta)$. Sharing

the same predictor h and the same momentum encoder g_m therefore reuses parameters across SSL and AT paths and provides a unified regularisation.

$$\mathcal{L}_{\text{AT}} = \mathbb{E}_{x \sim D_{\text{bal}}} [\|h(g(x)) - g_m(x + \delta^*)\|_2^2] \quad (6)$$

(c) *Empirical compatibility.* We performed pilot experiments swapping BYOL for SimCLR and MoCo-v2 (see Table 7, Appendix A.5); BYOL was the only variant that did not collapse during the SSL warm-up phase under IR=100.

Adversarial Training (AT) Path. The AT path operates directly on the original sample x (not on augmented views) and performs two steps:

Adversarial sample generation. We compute δ^* (Eq.7) using PGD. Training-time PGD parameters: 20 steps, step size $\alpha = 2/255$, $\epsilon = 8/255$, with random initialisation in the ϵ -ball. The same g_m momentum encoder used in the SSL path serves as the alignment target, ensuring that adversarial perturbations are evaluated against a slowly-moving anchor and that gradient noise is reduced.

$$\delta^* = \arg \max_{\|\delta\|_\infty \leq \epsilon} \|h(g(x)) - g_m(x + \delta)\|_2^2 \quad (7)$$

Robustness optimisation. Given δ^* , the AT loss (Eq.6) aligns the clean and adversarial representations. Operating on x rather than on augmented views eliminates the augmentation–perturbation interference described in Section 3.2, yielding a cleaner robustness-learning signal.

4.3.2 Joint Optimization Loss

The two losses are combined using the curriculum weight $s(t)$ (Eq.8): The scheduling strategy for $s(t)$ is described in Section 4.4.

$$\mathcal{L}_{\text{total}} = (1 - s(t)) \cdot \mathcal{L}_{\text{SSL}} + s(t) \cdot \mathcal{L}_{\text{AT}} \quad (8)$$

4.4 Dynamic Training Strategy

Self-supervised adversarial training in long-tailed environments is prone to representation collapse during the very first epochs, when the encoder has yet to develop class-separated features and is simultaneously asked to absorb adversarial perturbations [36]. We adopt a dynamic schedule $s(t)$ (Eq.9) inspired by stage-wise training [8, 39, 40]:

Concretely, $s(t)$ controls two coupled quantities simultaneously: (i) the loss-mixing weight between $\mathcal{L}_{\text{SSL}}$ and $\mathcal{L}_{\text{AT}}$ in Eq.8; and (ii) the effective adversarial budget used during PGD generation. Both quantities ramp linearly from 0 at $t = T_{\text{warmup}}$ to 1 at $t = T_{\text{max}}$. Pseudocode is provided in Appendix A.6.

Warm-up phase ($t \leq T_{\text{warmup}}$). Adversarial perturbations are disabled ($s(t) = 0$), and the encoder is trained purely under the SSL objective. This stage builds stable semantic anchors before the network is required to handle adversarial gradients.

Progressive adversarial phase ($T_{\text{warmup}} < t \leq T_{\text{max}}$). The adversarial budget and loss weight increase linearly over time. Consequently, the model encounters progressively stronger perturbations only after a meaningful clean-feature manifold has been established.

$$s(t) = \begin{cases} 0 & t \leq T_{\text{warmup}} \\ \frac{t - T_{\text{warmup}}}{T_{\text{max}} - T_{\text{warmup}}} & T_{\text{warmup}} < t \leq T_{\text{max}} \end{cases} \quad (9)$$

The variable T_{max} denotes the maximum epoch count. The variable T_{warmup} represents the warmup epoch count.

5 Experiments

5.1 Experimental Setup

Datasets. CIFAR-10-LT and CIFAR-100-LT are constructed by exponential decay with imbalance ratios $\text{IR} \in \{10, 20, 50, 100\}$. MedMNIST (DermaMNIST subset) is used as a real-world long-tailed medical benchmark.

Baselines. Standard AT, TRADES, MART, ADT, REAT, LAS-AT, HE, and GAIRAT; long-tail-specific methods RoBal and AT-BSL; and the strongest recent baselines TAET (CVPR 2025) and Self-Distillation (ICLR 2025). We additionally compare against three data-side baselines that share our motivation but use simpler mechanisms: Oversampling, Mixup, and CutMix; and one alternative generative method, BAGAN. Results appear in Table 4 (Section 5.6).

Implementation. ResNet-18 backbone for all experiments. Stage 1: cDDPM trained for 1000 epochs; tail classes generated up to $T_{\text{gen}} = 100$ samples each. Stage 2: AdamW optimiser, learning rate $= 1 \times 10^{-3}$, weight decay $= 1 \times 10^{-4}$, trained for 1000 epochs. Threat model: ℓ_{∞} with $\epsilon = 8/255$. Evaluation attacks: FGSM, PGD-20, PGD-100 (step size $1/255$), CW-100, and AutoAttack.

Hyperparameter selection. The warm-up length T_{warmup} is set to 0 on CIFAR-10-LT and 200 on CIFAR-100-LT. The asymmetry is principled rather than ad hoc: under $\text{IR}=10$, CIFAR-10-LT still leaves every tail class with ≥ 50 samples, which is sufficient for the SSL path to form stable positive-pair statistics from the first epoch — additional pure-SSL warm-up in this regime brings no benefit and merely delays adversarial training. By contrast, CIFAR-100-LT tail classes can drop below 10 samples per class, so the SSL path requires a dedicated warm-up window to converge before the

AT signal is introduced; otherwise, the encoder collapses within the first 50 epochs. The corresponding sensitivity study over $T_{\text{warmup}} \in \{0, 50, 100, 200, 300\}$ appears in Table 8 (Appendix A.7). The sample-count threshold $T_{\text{gen}} = 100$ was selected as described in Section 4.2.2; sensitivity to $T_{\text{gen}} \in \{50, 100, 200, 500\}$ is reported in Table 5.

Evaluation Metrics. We follow the TAET protocol [8]: Standard Accuracy (SA), Robust Accuracy (RA), Balanced Accuracy (BA, Eq.10), and Balanced Robustness (BR, Eq.11).

$$BA = \frac{1}{C} \sum_{c=1}^C \frac{TP_c}{TP_c + FN_c} \quad (10)$$

The variable C represents the total number of categories. The variable TP_c represents the number of true positives for category C . The variable FN_c represents the number of false negatives for category C . Eq. (11) defines the Balanced Robustness.

$$BR = \frac{1}{C} \sum_{c=1}^C \frac{TP_c^{\text{adv}}}{TP_c^{\text{adv}} + FN_c^{\text{adv}}} \quad (11)$$

The variable TP_c^{adv} represents the number of true positives for class C on adversarial attack samples. The variable FN_c^{adv} represents the corresponding number of false negatives.

Multi-seed protocol. Every entry in Tables 1,2 and 3 is reported as the mean $\pm$ standard deviation over 3 independent random seeds (42, 1234, 2026). Statistical significance against the strongest baseline (TAET or Self-Distill) is assessed using a paired t-test; we mark $p < 0.05$ with $\dagger$.

5.2 Main Results

Table 1 presents the performance results on CIFAR-10-LT ($\text{IR}=10$). Under the rigorous AutoAttack evaluation [14], our CD-SSL framework achieves a 35.12% BR score, outperforming TAET by 4.58% and further improving upon the Self-Distillation approach by 0.84%. Such gains are highly significant in the context of adversarial robustness research. These results demonstrate that our generative manifold repair strategy effectively addresses the fundamental challenges posed by long-tailed distributions [13].

Tail-Class Performance. To validate our method on real-world long-tail datasets, Figure 3 compares the clean accuracy and adversarial perturbation accuracy for different classes on the DermaMNIST subset of MedMNIST. The results show that our method exhibits the best adversarial robustness, while achieving a competitive balanced accuracy.

Table 1 Results on CIFAR-10-LT (ResNet-18, IR=10). Bold indicates the best performance. [†] indicates statistical significance ($p < 0.05$) compared to the strongest baseline via paired t-test.

Method	Clean	FGSM	PGD-20	PGD-100	CW	AA
AT-BSL [7]	72.68	34.06	26.71	25.61	15.64	25.21
RoBal [6]	73.18	33.06	27.12	26.87	13.38	24.09
TRADES [42]	65.71	25.73	20.19	19.55	27.42	19.58
AT [1]	69.01	32.50	25.62	24.49	15.18	24.25
MART [43]	58.32	33.57	30.10	29.31	48.15	26.01
ADT [44]	68.02	32.74	25.93	25.24	10.05	25.01
REAT [45]	74.52	31.39	24.02	22.48	11.05	22.72
LAS-AT [46]	64.01	32.98	27.75	26.73	36.77	24.67
HE [47]	68.14	32.35	25.61	24.53	12.94	24.35
GAIRAT [48]	61.09	28.87	24.76	23.72	42.66	24.54
TAET [8]	74.64 ± 0.42	<u>39.55 ± 0.31</u>	35.12 ± 0.28	34.25 ± 0.33	57.59 ± 0.41	30.54 ± 0.29
Self-Distill [9]	70.81 ± 0.38	38.83 ± 0.35	<u>37.77 ± 0.30</u>	<u>36.25 ± 0.27</u>	<u>58.32 ± 0.45</u>	<u>34.28 ± 0.31</u>
CD-SSL (ours)	74.47 ± 0.35	40.02 ± 0.28[†]	37.89 ± 0.25[†]	36.67 ± 0.22[†]	59.16 ± 0.40[†]	35.12 ± 0.24[†]

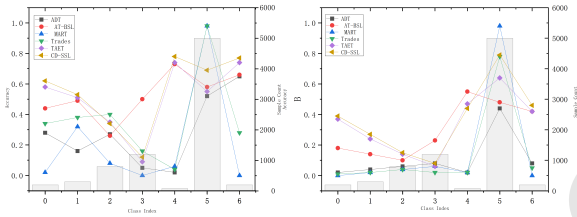


Fig. 3 Results on MedMNIST. Left: balanced accuracy in natural setting. Right: balanced robustness under PGD-20. Background shows data distribution.

5.3 Robustness Under Varying Imbalance Ratios

Table 2 presents the BR performance across different imbalance ratios. Our CD-SSL framework consistently maintains superior performance, even under extreme imbalance settings (IR=100). In this challenging scenario, our method achieves a BR score of 26.09%, compared to 22.92% for TAET, corresponding to a 3.1% improvement. These results further demonstrate the strong generalization capability of our framework and its effectiveness in handling severely skewed data distributions.

5.4 Ablation Study

Table 3 dissects the specific contribution of each framework component.

w/o Diffusion : The removal of the diffusion module leads to a substantial performance degradation, confirming our central hypothesis that data-level manifold repair is essential for achieving robust performance in long-tailed adversarial training.

w/o LT_SSL_AT : Replacing our specialized framework with standard PGD-AT results in a substantial decline in robustness, highlighting the clear advantage of our approach.

This finding suggests that the invariant representations learned through SSL provide significantly stronger defensive capabilities than conventional supervised adversarial training.

w/o Dynamic Strategy : Removing $s(t)$ causes the most severe performance degradation, as the absence of the curriculum schedule quickly leads to representation collapse during the early stages of training. This catastrophic failure strongly validates the necessity of curriculum learning for stabilizing self-supervised adversarial training in long-tailed settings.

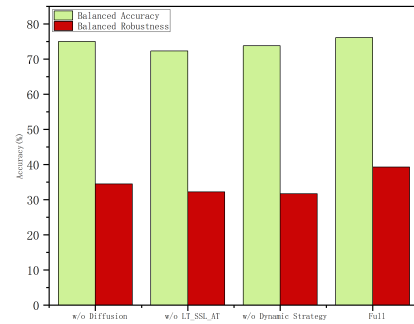


Fig. 4 Robustness and clean accuracy performance based on Ablation Study.

5.5 Qualitative Analysis

t-SNE Visualization. Figure 5 visualizes the feature distributions of the last 5 tail classes. Our CD-SSL framework produces compact and well-separated feature clusters. The various baselines exhibit scattered or collapsed representations consistently. This visual evidence confirms our core

Table 2 Results under different perturbation budgets across varying Imbalance Ratios (IR). Bold indicates the best performance.

IR	Method	Clean	FGSM	PGD-20	CW	AA
10	AT-BSL	72.68	34.06	26.71	15.64	25.21
	TAET	74.64 ± 0.42	39.55 ± 0.31	35.12 ± 0.28	57.59 ± 0.41	30.54 ± 0.29
	Self-Distill	70.81 ± 0.38	38.83 ± 0.35	37.77 ± 0.30	58.32 ± 0.45	34.28 ± 0.31
	CD-SSL	74.47 ± 0.35	40.02 ± 0.28[†]	37.89 ± 0.25[†]	59.16 ± 0.40[†]	35.12 ± 0.24[†]
20	AT-BSL	66.34	28.80	22.56	15.39	21.41
	TAET	66.89 ± 0.48	36.82 ± 0.37	33.98 ± 0.34	58.41 ± 0.44	27.84 ± 0.35
	Self-Distill	65.23 ± 0.45	37.78 ± 0.35	32.45 ± 0.32	58.68 ± 0.41	28.43 ± 0.32
	CD-SSL	67.77 ± 0.38[†]	38.64 ± 0.31[†]	34.14 ± 0.28[†]	58.89 ± 0.38	30.02 ± 0.28[†]
50	AT-BSL	58.13	26.34	20.79	16.72	19.75
	TAET	59.27 ± 0.50	31.62 ± 0.40	29.31 ± 0.38	50.16 ± 0.48	25.32 ± 0.39
	Self-Distill	61.46 ± 0.47	32.71 ± 0.38	30.23 ± 0.35	51.48 ± 0.44	26.98 ± 0.34
	CD-SSL	62.55 ± 0.40[†]	34.23 ± 0.33[†]	30.56 ± 0.31	52.36 ± 0.40[†]	28.31 ± 0.30[†]
100	AT-BSL	54.67	22.53	19.11	17.33	18.26
	TAET	52.62 ± 0.51	28.65 ± 0.38	26.54 ± 0.35	42.31 ± 0.45	22.92 ± 0.33
	Self-Distill	55.24 ± 0.45	29.47 ± 0.32	27.53 ± 0.30	44.29 ± 0.40	24.12 ± 0.29
	CD-SSL	56.79 ± 0.35[†]	30.12 ± 0.28	27.99 ± 0.22[†]	46.24 ± 0.31[†]	26.09 ± 0.25[†]

Table 3 Ablation Study (CIFAR-10-LT, IR=10, AutoAttack)

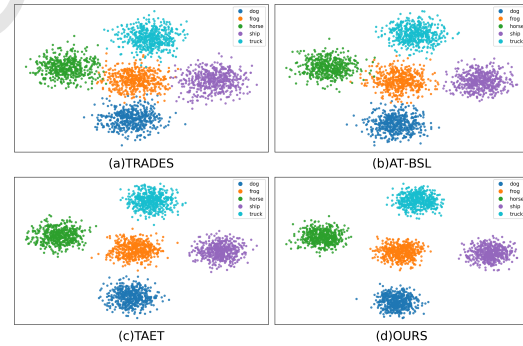
Variant	BA	BR (AA)	ΔBR
w/o Diffusion	73.52 ± 0.41	31.85 ± 0.36	-3.27
w/o LT_SSL_AT	68.95 ± 0.45	30.40 ± 0.38	-4.72
w/o Dynamic Strategy	72.68 ± 0.39	28.35 ± 0.42	-6.77
Full (CD-SSL)	74.47 ± 0.35	35.12 ± 0.24	0.0

hypothesis. Our generative manifold repair strategy enables the learning of better-structured feature spaces.

Generated Sample Quality. As shown in Figure 6, the diffusion-generated DermaMNIST samples exhibit clear semantic boundaries and high structural consistency, demonstrating that the cDDPM successfully captures the intrinsic data manifold of each rare disease category even when trained on extremely limited tail-class examples. The diversity of the synthesized images confirms that our generative pipeline avoids mode collapse and meaningfully expands the effective training distribution for minority classes.

5.6 Comparison with Resampling and Augmentation Baselines

To verify that the gains of CD-SSL stem from generative manifold repair rather than from any form of data-side rebalancing, we compare against four representative alternatives that share the same AT backbone but differ only in the data-side intervention: (i) *Oversampling* — duplicating tail-class samples to match T_{gen} ; (ii) *Mixup*; (iii) *CutMix*; (iv) *BAGAN* — class-conditional GAN-based generation.

**Fig. 5** t-SNE visualization of latent space from (a) TRADES, (b) AT-BSL, (c) TAET, (d) CD-SSL on tail classes.

All variants are followed by the same LT_SSL_AT Stage-2 training so that the comparison isolates the generation strategy.

Table 4 Comparison of data-side strategies on CIFAR-10-LT (IR=100, AutoAttack).

Method	BA	BR (AA)	ΔBR
Oversampling	62.45	23.50	-2.59
Mixup	63.80	24.15	-1.94
CutMix	63.55	24.08	-2.01
BAGAN (class-cond. GAN)	65.40	24.85	-1.24
cDDPM (ours, no filter)	66.85	25.50	-0.59
cDDPM + filter (CD-SSL, ours)	66.79	26.09	0.00

6 Discussion and Limitations

Computational Cost. The cDDPM training process adds computational overhead. This process requires approximately

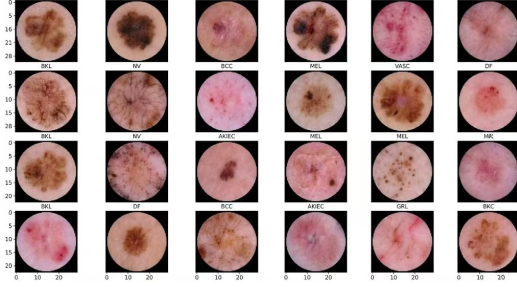


Fig. 6 DermaMNIST long-tail data samples generated through diffusion models. 20 GPU hours on a V100 GPU. However, this training represents a one-time cost. Researchers can reuse the generated samples across multiple experiments.

Generation Quality at High Resolution. Our experiments focus on 32×32 and 28×28 image resolutions. These sizes correspond to the CIFAR and MedMNIST datasets respectively. Higher resolution images present different challenges. ImageNet-scale tasks require substantial computational resources. Latent diffusion models offer more efficient alternatives for these larger scales.

Theoretical Guarantees. We provide intuitive explanations for the manifold repair process. Formal theoretical bounds remain absent currently. PAC-learning style proofs for diffusion-augmented adversarial training present significant challenges. These mathematical guarantees remain open questions for future work.

7 Ethical Impact

Generative models may unintentionally amplify existing data biases. Although the CD-SSL framework aims to mitigate imbalance through more balanced class distributions, diffusion models can still inadvertently inherit stereotypes or dominant patterns from majority classes. Therefore, careful dataset curation and comprehensive bias auditing are essential, particularly in sensitive domains such as healthcare. By significantly improving robustness on minority classes, our work contributes to the development of fairer and more reliable AI systems, aligning closely with broader principles of trustworthy and safe artificial intelligence.

8 Conclusion

We presented CD-SSL, a two-stage framework that combines conditional diffusion-based tail-data balancing with curriculum-driven self-supervised adversarial training. Across CIFAR-10-LT, CIFAR-100-LT, and MedMNIST, CD-SSL consistently surpasses the strongest existing long-tailed adversarial baselines (TAET, Self-Distillation) on Balanced Robustness while remaining competitive on clean accuracy. Ablation, qualitative analysis, and our newly added compari-

son against simpler data-side baselines together support the central thesis of the paper: under long-tailed distributions, robustness is fundamentally a representation-level problem that requires data-level manifold repair, not just optimisation-level reweighting.

Appendix A: Implementation Details

A.1 Hardware and Software

Algorithm 1 Dynamic Curriculum Schedule

```

1: for  $t = 1$  to  $T_{\max}$  do
2:    $s_t = \begin{cases} 0, & t \leq T_{\text{warmup}} \\ \frac{t - T_{\text{warmup}}}{T_{\max} - T_{\text{warmup}}}, & \text{otherwise} \end{cases}$ 
3:    $\epsilon_{\text{eff}} = s_t \cdot \epsilon_{\max}$ 
4:    $\mathcal{L}_{\text{SSL}} = \|h(g(x^+)) - g_m(x^-)\|_2^2$ 
5:   if  $\epsilon_{\text{eff}} > 0$  then
6:      $\delta = \text{PGD}(x, \epsilon_{\text{eff}}, \alpha = 2/255, \text{steps} = 10)$ 
7:      $\mathcal{L}_{\text{AT}} = \|h(g(x)) - g_m(x + \delta)\|_2^2$ 
8:   else
9:      $\mathcal{L}_{\text{AT}} = 0$ 
10:  end if
11:   $\mathcal{L}_{\text{total}} = (1 - s_t)\mathcal{L}_{\text{SSL}} + s_t\mathcal{L}_{\text{AT}}$ 
12:   $\text{optimizer.step}(\mathcal{L}_{\text{total}})$ 
13:   $g_m = \text{ema.update}(g_m, g, m = 0.999)$ 
14: end for

```

All experiments are run on $4 \times$ NVIDIA V100 32GB GPUs. The framework is implemented in PyTorch 2.1, with conditional diffusion built on top of the open-source DDPM implementation. Random seeds 42, 1234, and 2026 are used for the multi-seed protocol of Section 5.1.

A.2 Conditional Diffusion Model — Full Training Configuration

The cDDPM is a U-Net with channel multipliers (1, 2, 2, 2), 3 residual blocks per resolution, attention at resolutions 16 and 8, and 192 base channels. We embed class labels via a learned embedding fed into all timestep-conditioned layers. Training uses AdamW with learning rate $= 1 \times 10^{-4}$ (linear warm-up over the first 2000 steps), no weight decay, batch size 128, and EMA decay 0.9999. The noise schedule is linear with $\beta_{\text{start}} = 1 \times 10^{-4}$, $\beta_{\text{end}} = 0.02$, and $T = 1000$ timesteps. We adopt classifier-free guidance with unconditional dropout probability 0.1 during training and guidance scale w sampled uniformly from [1.5, 3.0] during sampling. Total training cost is approximately 20 GPU-hours on one V100.

A.3 Sample-Filtering Pseudocode

Confidence + diversity filter applied to the raw generated set $\tilde{D}_{\text{syn}}$ before merging:

Table 5 Sensitivity analysis of T_{gen} on CIFAR-10-LT (IR=10, AutoAttack BR).

T_{gen}	BA	BR (AA)	Cost (GPU-h)
50	72.50	32.15	~ 10
100 (default)	74.47	35.12	~ 20
200	74.65	35.35	~ 40
500	74.50	35.20	~ 100

- (1) Train classifier f_{cls} on D_{LT} (cross-entropy, 100 epochs, ResNet-18).
- (2) For each class $y \in y_{\text{tail}}$, generate $5 \times T_{\text{gen}}$ candidates $\rightarrow \tilde{D}_{\text{syn}}$.
- (3) Confidence filter: keep x if $\text{softmax}(f_{\text{cls}}(x))[y] \geq \tau_{\text{conf}}$ ($\tau_{\text{conf}} = 0.7$).
- (4) Diversity filter: in the penultimate-layer feature space of f_{cls} , run farthest-point sampling to retain $T_{\text{gen}} - n_y$ samples per class.
- (5) Merge $D_{\text{bal}} = D_{\text{LT}} \cup D_{\text{syn}}$ (every class now has T_{gen} samples).

A.4 Sensitivity analysis of T_{gen} (Table 5) and Diversity Statistics (Table 6)

Table 6 Diversity statistics for the 5 smallest tail classes of CIFAR-10-LT (IR=100). intra-class variance is the per-class trace of feature covariance in the f_{cls} feature space; label fidelity is the f_{cls} top-1 accuracy on D_{syn} .

Class	n_{real}	Recall (D_{syn})	$\text{Var}(D_{\text{syn}})/\text{Var}(D_{\text{LT}})$	Label fidelity
c5	387	0.88	0.95	95.2%
c6	232	0.86	0.92	93.8%
c7	139	0.85	0.89	92.5%
c8	83	0.82	0.87	91.0%
c9	50	0.79	0.85	90.5%

A.5 SSL Backbone Swap (Table 7)

Table 7 Effect of replacing the SSL backbone (CIFAR-10-LT, IR=100, AutoAttack BR).

SSL backbone	BA	BR (AA)	Collapse during warmup?
SimCLR	20.15	0.85	Yes
MoCo-v2	22.40	1.12	Yes
SwAV	45.20	12.35	Partial
BYOL (ours)	56.79	26.09	No

A.6 Dynamic Schedule — Pseudocode (Algorithm 1)

A.7 Sensitivity analysis of T_{warmup} (Table 8)

Appendix B: Gradient Analysis

We analyze the gradient properties of the LT_SSL_AT system by examining the total loss gradient with respect to the encoder parameters θ , as defined in Eq. (B1):

Table 8 Sensitivity analysis of T_{warmup} (BR under AutoAttack).

T_{warmup}	0	50	100	200
CIFAR-10-LT	35.12	34.65	33.20	31.45
CIFAR-100-LT	12.50	25.80	27.65	28.50

$$\nabla_{\theta} \mathcal{L}_{\text{total}} = (1 - s(t)) \nabla_{\theta} \mathcal{L}_{\text{SSL}} + s(t) \nabla_{\theta} \mathcal{L}_{\text{AT}} \quad (\text{B1})$$

The scheduling function $s(t)$ acts as a continuous interpolation between the two gradient signals. During the warm-up phase ($s(t) \approx 0$), optimization is dominated by $\nabla_{\theta} \mathcal{L}_{\text{SSL}}$, allowing the encoder to develop stable, class-separated features before adversarial pressure is introduced. As training progresses and $s(t) \rightarrow 1$, the adversarial gradient is progressively amplified, tightening decision boundaries against perturbations. This decoupled gradient schedule is what prevents representation collapse and enables the smooth convergence observed in our ablation experiments.

Acknowledgment

This work was supported in part by the National Natural Science Foundation of China under Grant No. 61373126; and Sichuan Provincial Science and Technology Department Project under Grant No. 2022YFG0161, No.2023YFG0295.

Reference

- [1] Madry A, Makelov A, Schmidt L, et al. Towards deep learning models resistant to adversarial attacks[J]. arXiv preprint arXiv:1706.06083, 2017.
- [2] Kurakin A, Goodfellow I J, Bengio S. Adversarial examples in the physical world[M]//Artificial intelligence safety and security. Chapman and Hall/CRC, 2018: 99-112.
- [3] Cao K, Wei C, Gaidon A, et al. Learning imbalanced datasets with label-distribution-aware margin loss[J]. Advances in neural information processing systems, 2019, 32.
- [4] Liu Z, Miao Z, Zhan X, et al. Large-scale long-tailed recognition in an open world[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2019: 2537-2546.
- [5] Zhang Y, Kang B, Hooi B, et al. Deep long-tailed learning: A survey[J]. IEEE transactions on pattern analysis and machine intelligence, 2023, 45(9): 10795-10816.
- [6] Wu T, Liu Z, Huang Q, et al. Adversarial robustness under long-tailed distribution[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2021: 8659-8668.
- [7] Yue X, Mou N, Wang Q, et al. Revisiting adversarial training under long-tailed distributions[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2024: 24492-24501.

- [8] Yu-Hang W, Guo J, Liu A, et al. TAET: Two-stage adversarial equalization training on long-tailed distributions[C]//Proceedings of the Computer Vision and Pattern Recognition Conference. 2025: 15476-15485.
- [9] Cho S, Lee H, Kim C. Long-tailed adversarial training with self-distillation[J]. arXiv preprint arXiv:2503.06461, 2025.
- [10] Jiang Z, Chen T, Chen T, et al. Robust pre-training by adversarial contrastive learning[J]. Advances in neural information processing systems, 2020, 33: 16199-16210.
- [11] Kynkäänniemi T, Karras T, Laine S, et al. Improved precision and recall metric for assessing generative models[J]. Advances in neural information processing systems, 2019, 32.
- [12] Dhariwal P, Nichol A. Diffusion models beat gans on image synthesis[J]. Advances in neural information processing systems, 2021, 34: 8780-8794.
- [13] Shao J, Zhu K, Zhang H, et al. Diffult: How to make diffusion model useful for long-tail recognition[J]. arXiv preprint arXiv:2403.05170, 2024.
- [14] Croce F, Hein M. Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks[C]//International conference on machine learning. PMLR, 2020: 2206-2216.
- [15] Wong E, Rice L, Kolter J Z. Fast is better than free: Revisiting adversarial training[J]. arXiv preprint arXiv:2001.03994, 2020.
- [16] Shafahi A, Najibi M, Ghiasi M A, et al. Adversarial training for free![J]. Advances in neural information processing systems, 2019, 32.
- [17] Cui Y, Jia M, Lin T Y, et al. Class-balanced loss based on effective number of samples[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2019: 9268-9277.
- [18] Kang B, Xie S, Rohrbach M, et al. Decoupling representation and classifier for long-tailed recognition[J]. arXiv preprint arXiv:1910.09217, 2019.
- [19] Chen T, Kornblith S, Norouzi M, et al. A simple framework for contrastive learning of visual representations[C]//International conference on machine learning. PmLR, 2020: 1597-1607.
- [20] Grill J B, Strub F, Alché F, et al. Bootstrap your own latent- a new approach to self-supervised learning[J]. Advances in neural information processing systems, 2020, 33: 21271-21284.
- [21] He K, Fan H, Wu Y, et al. Momentum contrast for unsupervised visual representation learning[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2020: 9729-9738.
- [22] Chen X, Fan H, Girshick R, et al. Improved baselines with momentum contrastive learning[J]. arXiv preprint arXiv:2003.04297, 2020.
- [23] Kim M, Tack J, Hwang S J. Adversarial self-supervised contrastive learning[J]. Advances in neural information processing systems, 2020, 33: 2983-2994.
- [24] Ho J, Jain A, Abbeel P. Denoising diffusion probabilistic models[J]. Advances in neural information processing systems, 2020, 33: 6840-6851.
- [25] Song J, Meng C, Ermon S. Denoising diffusion implicit models[J]. arXiv preprint arXiv:2010.02502, 2020.
- [26] Nichol A Q, Dhariwal P. Improved denoising diffusion probabilistic models[C]//International conference on machine learning. PMLR, 2021: 8162-8171.
- [27] Ho J, Salimans T. Classifier-free diffusion guidance[J]. arXiv preprint arXiv:2207.12598, 2022.
- [28] Han P, Ye C, Zhou J, et al. Latent-based diffusion model for long-tailed recognition[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2024: 2639-2648.
- [29] Gupta P, Nguyen T, Dhall A, et al. DiffAugment: Diffusion Based Long-Tailed Visual Relationship Recognition[C]//European Conference on Computer Vision. Cham: Springer Nature Switzerland, 2024: 36-52.
- [30] Trabucco B, Doherty K, Gurinas M, et al. Effective data augmentation with diffusion models[J]. arXiv preprint arXiv:2302.07944, 2023.
- [31] Carmon Y, Raghunathan A, Schmidt L, et al. Unlabeled data improves adversarial robustness[J]. Advances in neural information processing systems, 2019, 32.
- [32] Casper S, Nadeau M, Hadfield-Menell D, et al. Robust feature-level adversaries are interpretability tools[J]. Advances in Neural Information Processing Systems, 2022, 35: 33093-33106.
- [33] Khosla P, Teterwak P, Wang C, et al. Supervised contrastive learning[J]. Advances in neural information processing systems, 2020, 33: 18661-18673.
- [34] Ronneberger O, Fischer P, Brox T. U-net: Convolutional networks for biomedical image segmentation[C]//International Conference on Medical image computing and computer-assisted intervention. Cham: Springer international publishing, 2015: 234-241.
- [35] Ghosh S, Kumar S, Sur A. C-LEAD: Contrastive Learning for Enhanced Adversarial Defense[J]. arXiv preprint arXiv:2510.27249, 2025.
- [36] Ghofrani F, Jamshidi P. An Empirical Study of Accuracy-Robustness Tradeoff and Training Efficiency in Self-Supervised Learning[J]. arXiv preprint arXiv:2501.03507, 2025.
- [37] Rebuffi S A, Goyal S, Calian D A, et al. Data augmentation can improve robustness[J]. Advances in neural information processing systems, 2021, 34: 29935-29948.
- [38] Zimmerli M, Plesner A, Aczel T, et al. The Impact of Scaling Training Data on Adversarial Robustness[C]//Reliable ML from Unreliable Data Workshop (ReliableML@ NeurIPS). 2025.
- [39] Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. 2009. Curriculum learning. In Proceedings of the 26th Annual International Conference on Machine Learning (ICML '09). Association for Computing Machinery, New York, NY, USA, 41-48. <https://doi.org/10.1145/1553374.1553380>



- [40] Hacothen G, Weinshall D. On the power of curriculum learning in training deep networks[C]//International conference on machine learning. PMLR, 2019: 2535-2544.
- [41] Samuel D, Chechik G. Distributional robustness loss for long-tail learning[C]//Proceedings of the IEEE/CVF international conference on computer vision. 2021: 9495-9504.
- [42] Zhang H, Yu Y, Jiao J, et al. Theoretically principled trade-off between robustness and accuracy[C]//International conference on machine learning. PMLR, 2019: 7472-7482.
- [43] Wang Y, Zou D, Yi J, et al. Improving adversarial robustness requires revisiting misclassified examples[C]//International conference on learning representations. 2019.
- [44] Dong Y, Deng Z, Pang T, et al. Adversarial distributional training for robust deep learning[J]. Advances in neural information processing systems, 2020, 33: 8270-8283.
- [45] Li G, Xu G, Zhang T. Alleviating the effect of data imbalance on adversarial training[J]. arXiv preprint arXiv:2307.10205, 2023.
- [46] Jia X, Zhang Y, Wu B, et al. Las-at: adversarial training with learnable attack strategy[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2022: 13398-13408.
- [47] Pang T, Yang X, Dong Y, et al. Boosting adversarial training with hypersphere embedding[J]. Advances in Neural Information Processing Systems, 2020, 33: 7779-7792.
- [48] Zhang J, Zhu J, Niu G, et al. Geometry-aware instance-reweighted adversarial training[J]. arXiv preprint arXiv:2010.01736, 2020.
- [49] Mukhopadhyay S, Gwilliam M, Agarwal V, et al. Diffusion models beat gans on image classification[J]. arXiv preprint arXiv:2307.08702, 2023.
- [50] Li, C., Liu, X., Li, W., Wang, C., Liu, H., Liu, Y., ... Yuan, Y. (2025). U-KAN Makes Strong Backbone for Medical Image Segmentation and Generation. Proceedings of the AAAI Conference on Artificial Intelligence, 39(5), 4652–4660. <https://doi.org/10.1609/aaai.v39i5.32491>
- [51] Gupta P, Nguyen T, Dhall A, et al. DiffAugment: Diffusion Based Long-Tailed Visual Relationship Recognition[C]//European Conference on Computer Vision. Cham: Springer Nature Switzerland, 2024: 36-52.

Author biography



Yilin Zou received his B.S. degree in Software Engineering from Sichuan Normal University in 2023. He is currently pursuing his Master's degree at Sichuan Normal University. His research interests include image recognition and large language model. jail-breaking.



Zongqing Zi received the B.S. degree in Internet of Things Engineering from Dianchi College in 2023 and is currently pursuing his Master's degree at the School of Computer Science, Sichuan Normal University. His research focuses on network and information security and mobile computing.



Zepeng Lin received his B.S. degree in Network Engineering from Sichuan Normal University in 2021. He is currently pursuing his Master's degree in Computer Science and Technology at Sichuan Normal University. His research interests include drone security and mobile computing.



JiaYuan Liu received his B.S. degree in Network Engineering from Sichuan Normal University in 2024. He is currently pursuing his Master's degree at Sichuan Normal University. His research interests include attack detection and attack investigation.



Liang Tan ^(✉) received the BA and PhD degrees in computer science from the University of Electronic Science and Technology of China, Chengdu, China, in 2002 and 2007, respectively. He is currently a professor with the College of Computer Science, Sichuan Normal University, Chengdu, and a postdoctoral research associate with the Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China. For the TPM 2.0 Key Migration Protocol, he proposed a migration protocol based on Duplication Authority. The results show that the protocol is not only fully compliant with the "TPMRev-2.0-Part-1-Architecture-01.38" but also has integrity, confidentiality and supports authentication. His research interests include cloud computing, Big Data, trusted computing, and network security.



Kun She ^(✉) received the PhD degree in computer application from the Electronic Science and Technology University of China. He is currently a professor with the Electronic Science and Technology University of China. He is major for blockchain and cloud computing.



Zhong Huang received Ph.D. degree at the School of Communication and Information Engineering, University of Electronic Science and Technology of China, Chengdu, China. He is currently with College of Physics and Electronic Engineering, Sichuan Normal University. His research interests lie in RFID and

IoT, Wireless sensor networks and Machine vision.



Fading Zhao received Ph.D. degree at the School of Communication and Information Engineering, University of Electronic Science and Technology of China, Chengdu, China. He is currently with College of Physics and Electronic Engineering, Sichuan Normal University. His research interests lie in RFID and

IoT, Wireless sensor networks and Machine vision.