TAET : Two-Stage Adversarial Equalization Training on Long-Tailed Distributions

Wang Yu-Hang[1], Junkang Guo[1], Aolei Liu[1], Kaihao Wang[1], Zaitong Wu[1], Zhenyu Liu[1], Wenfei Yin[1], ***Jian Liu[1] (Corresponding author)***

[1]School of Computer Science and Information Engineering, Hefei University of Technology

CVPR Nashville JUNE 11-15, 2025

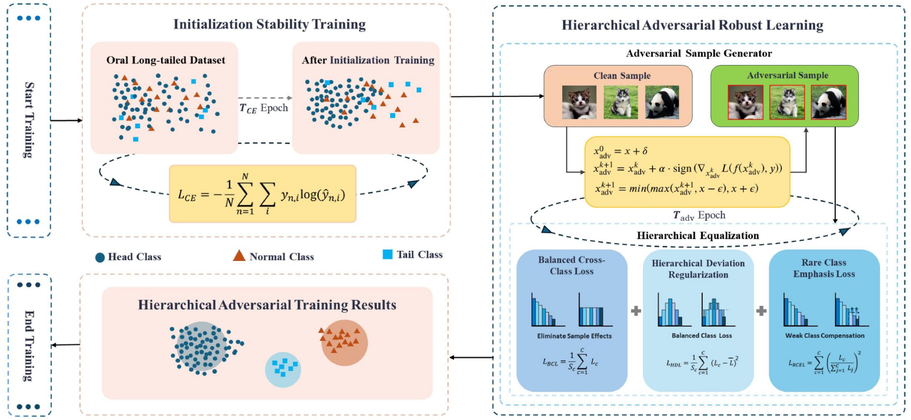
Motivation

Deep learning models in computer vision, despite their success, face a critical security challenge: vulnerability to adversarial attacks. While Adversarial Training (AT) is a leading defense, its effectiveness diminishes significantly when confronted with real-world long-tailed data distributions, where some classes are common ("head") and many others are rare ("tail"). This leads to several pressing issues:

- **Poor Robustness on Tail Classes:** Standard AT methods fail to adequately protect underrepresented tail classes, making models unreliable for many real-world scenarios.
- **Robustness Overfitting:** Models trained on long-tailed data may show improved accuracy but paradoxically become more vulnerable to adversarial attacks, especially on tail classes. Adversarial robustness does not scale with accuracy improvements.
- **Inadequate Evaluation:** Current metrics for adversarial robustness (often standard accuracy) do not accurately reflect performance across imbalanced class distributions, masking vulnerabilities in tail classes.
- **Training Inefficiency:** Existing approaches can be computationally expensive and memory-intensive when trying to achieve robust performance on long-tailed data.

There is an urgent need for novel adversarial training strategies and evaluation metrics specifically designed for long-tailed distributions to ensure reliable and secure deployment of computer vision models in practical, imbalanced scenarios.

TAET Framework



The TAET framework enhances model robustness through a two-stage process. It begins with an Initial Stabilization Module that uses cross-entropy loss to establish stable accuracy in early training. This stabilized model is then passed to the Hierarchical Adversarial Robust Learning (HARL) module, which further refines robustness using three key components: BCL, HDL, and RCEL. Adversarial perturbations are generated via a multi-step process and managed by normalization components within the framework.

A Novel Evaluation Metric: Balanced Robustness

$$\text{Balance Robustness} = \frac{1}{S_C}\sum_{i=1}^{C}\mathcal{R}_i^{x'} = \frac{1}{S_C}\sum_{i=1}^{C}\frac{\mathrm{TP}_i^{x'}}{\mathrm{TP}_i^{x'} + \mathrm{FN}_i^{x'}}$$

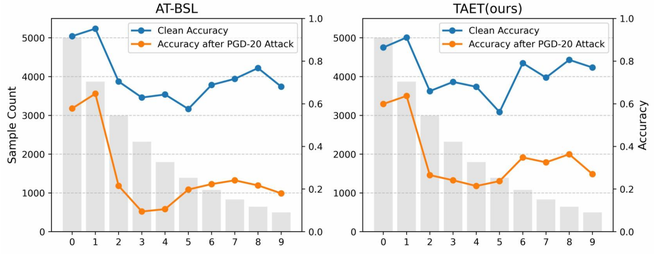
Balanced Robustness is a novel evaluation metric specifically designed to address the challenge of fairly assessing a model's resilience to adversarial attacks in datasets with long-tailed class distributions. Traditional accuracy metrics can be misleading in such scenarios due to class imbalance. Extending the concept of balanced accuracy (which measures average per-class accuracy), Balanced Robustness quantifies the model's average defense capability across all classes when subjected to adversarial examples. It is specifically calculated as the average of each class's recall (true positive rate) under adversarial conditions. Introduced as a pioneering metric for long-tail learning, Balanced Robustness aims to provide a more reliable and standardized tool for evaluating and advancing adversarial robustness, with significant potential in critical fields like medicine.

Performance on Challenging Tail Classes: CIFAR-10-LT

Method	Clean					Attacked (PGD-20)				
	Dog	Frog	Horse	Ship	Truck	Dog	Frog	Horse	Ship	Truck
AT-BSL[45]	58.63	73.48	70.48	78.29	66	23.02	26.97	21.7	21.7	19
RoBal[40]	**59.13**	74.73	68.52	78.96	65	**23.74**	25.78	20.79	19.5	20
TRADES[46]	53.96	65.11	63.25	47.28	47	11.87	13.2	16.27	3.87	4
AT[26]	51.07	62.32	63.85	64.34	54	14.74	17.67	19.27	17.05	10
ADT[12]	52.15	65.58	65.66	56.59	55	16.18	16.74	18.67	10.07	10
MART[38]	37.76	52.09	58.43	47.25	41	17.62	19.06	25.3	8.52	9
REAT[21]	58.63	73.02	**72.89**	78.29	69	16.18	13.48	18.67	20.15	17
LAS-AT[17]	50.72	55.34	56.02	48.83	34	19.06	13.02	20.48	12.4	8
HE[30]	57.31	67.9	65.66	51.16	54	14.74	20	18.07	10.85	12
GAIRAT[47]	48.56	58.13	56.62	48.06	43	15.82	17.2	22.28	6.2	11
TAET (our)	56.11	**79.06**	72.28	**80.6**	**77**	**23.74**	**34.83**	**32.53**	**36.34**	**27**

Table provides a detailed comparison of various methods, including our proposed approach, on the last five (tail) classes of the CIFAR-10-LT dataset with an imbalance ratio (IR) of 10, using the ResNet-18 architecture. The evaluation covers performance on both clean images and under adversarial attack conditions. The results, with the best performances highlighted in bold and second-best results underlined, clearly demonstrate the superior effectiveness of our method in these particularly challenging and underrepresented categories.

Long-Tail Adversarial Robustness: A Key Bottleneck



Observations reveal that when models process long-tailed data, the decline in adversarial robustness for tail classes becomes particularly pronounced. Specifically, as the sample volume per class decreases, its robustness correspondingly weakens, and the magnitude of this decline for tail classes is far greater than that for head classes. The performance of AT-BSL, a previously leading SOTA (State-of-the-Art) method in this domain, also attests to this prevalent challenge.

Our Code

We consider our work to be a significant contribution to addressing the challenges of long-tail robustness. Our code is open-source. Should you wish to use it, please scan the QR code on the left.

https://github.com/BuhuiOK/TAET-Two-Stage-Adversarial-Equalization-Training-on-Long-Tailed-Distributions