SAT: Semantic-Guided Adversarial Training on Long-Tailed Distributions

Huaxing Feng[1], Yuanbo Li[1], Cong Hu[1]*, Xiao-Jun Wu[1]
[1]School of Artificial Intelligence and Computer Science, Jiangnan University, Wuxi, China
6233114007@stu.jiangnan.edu.cn, liyuanbo12138@163.com,
conghu@jiangnan.edu.cn, wu_xiaojun@jiangnan.edu.cn

***Abstract*—Adversarial robustness remains a significant challenge in deploying deep neural networks (DNNs) for real-world applications. While adversarial training is widely acknowledged as a promising defense strategy, most existing studies primarily focus on balanced datasets, neglecting the fact that real-world data often exhibit a long-tailed distribution, which introduces substantial challenges to robustness. Existing adversarial training methods based on examples or loss balancing strategies, although able to alleviate class bias in long-tailed distributions to some extent, rely on data-driven balancing constraints that struggle to learn tail class features with sufficient discriminative information. To address this problem, we propose a novel framework based on adversarial training, termed Semantic-guided Adversarial Training (SAT). SAT utilizes class-level representations extracted by a text encoder as fixed semantic embeddings to guide model training. Specifically, we employ a vision–language model, CLIP, to obtain textual semantic representations for each class, exploit the semantic consistency of the text encoder, and introduce an alignment loss to mitigate the model's bias toward head classes, thereby improving adversarial robustness for tail classes. Furthermore, to more effectively evaluate the class-balanced performance of our method under long-tailed distributions, we introduce a more comprehensive evaluation metric, termed Balanced Robustness. Extensive experiments on multiple datasets demonstrate that SAT significantly enhances adversarial robustness under long-tailed distributions compared to state-of-the-art methods. In particular, on the CIFAR-10 dataset, our method achieves more than a 5% absolute improvement over existing state-of-the-art approaches.**

***Index Terms*—Adversarial Robustness, Adversarial Training, Long-Tailed Distribution, Semantic Embedding**

I. Introduction

Deep Neural Networks (DNNs) [1]–[3] have demonstrated outstanding performance across various tasks. However, they are highly vulnerable to adversarial examples [4], [5], where small perturbations in the input can lead to incorrect predictions, limiting their reliability in real-world applications. Adversarial training, which integrates adversarial examples into the training process, has shown significant effectiveness in improving model robustness. However, real-world data often exhibit long-tailed distributions [6], [7], where a small number of head classes have abundant samples, while the majority of tail classes are severely underrepresented. Under such conditions, models tend to be biased towards head classes, resulting in degraded natural accuracy and adversarial robustness for tail classes. Studies have shown that conventional adversarial training methods do not effectively improve robustness for tail classes under long-tailed distributions, and may even exacerbate performance imbalance across categories.

To address the issue of adversarial robustness under long-tailed distributions, prior works have explored loss reweighting and stage-wise training strategies [8], [9]. The TAET method [10] proposes a two-stage training pipeline, where standard training is performed initially, followed by the addition of hierarchical equalization loss to balance the robustness of tail classes. Although this method can somewhat alleviate class imbalance, tail classes still face limitations due to data-driven balancing constraints, leading to insufficient discriminative information for tail classes during training. This, in turn, limits effective feature learning and further restricts improvements in robustness.

Motivated by this issue, we propose the Semantic-guided Adversarial Training (SAT) framework to enhance adversarial robustness under long-tailed distributions. This method is driven by prior knowledge and does not rely on data-driven balancing constraints. By introducing semantic priors from large-scale vision-language models like CLIP, we construct fixed semantic embeddings for each class. These embeddings maintain cross-class consistency and strong discriminative power, independent of sample quantity. During adversarial training, we design a semantic alignment mechanism to guide the model's visual features to align with the corresponding text embeddings, reducing bias towards head classes and significantly improving both the natural accuracy and adversarial robustness of tail classes. We also emphasize class-level balanced performance, rather than solely focusing on overall robustness accuracy. Additionally, to better evaluate the effectiveness of SAT, we introduce the balanced robustness metric from TAET to assess model performance under long-tailed distributions. Experimental results show that SAT outperforms existing state-of-the-art methods on multiple long-tailed datasets, particularly demonstrating significant advantages in the adversarial robustness of tail classes. The main contributions of this work are summarized as follows:

- We propose a novel framework based on adversarial training, the Semantic-guided Adversarial Training (SAT) framework, which guides the model's long-tail training by leveraging the text encoder of CLIP to construct fixed semantic embeddings for each class.
- By introducing a semantic alignment loss, we ensure that the model's visual features are consistent with the

corresponding textual semantic representations, thereby applying balanced class constraints in the feature space and mitigating the model's bias towards head classes.
- Extensive experiments on multiple long-tailed datasets demonstrate that our method significantly improves both overall robustness and the performance of tail classes under long-tailed distributions.

II. Related Work

A. Adversarial Training

To resolve the vulnerability of deep learning models to adversarial attacks, researchers have explored various defense strategies, including adversarial training [11], defensive distillation [12], and adversarial example detection [13]. Among these, adversarial training stands out as one of the most effective and widely used methods, offering stable performance in enhancing adversarial robustness. It can be formulated as the following min-max optimization problem:

$$\theta^* = \arg\min_{\theta} \mathbb{E}_{(x,y)\in\mathcal{D}} \left[\max_{\delta\in S} \mathcal{L}(f_\theta(x+\delta), y; \theta) \right], \quad (1)$$

where $\mathcal{D}$ is data distribution of the dataset. The function $f_\theta(\cdot)$ represents a DNN model with parameters θ. The variable δ denotes a small perturbation confined to an ℓ_p-norm ball with radius ϵ satisfying $\| \delta \|_p \leq \epsilon$, x refers to a benign example and y is the ground-truth label for inputs x and (x, y) subject to $\mathcal{D}$ distribution, $\mathcal{L}(\cdot)$ is loss function.

Recently, several adversarial training variants have been introduced by various researchers [14]–[17]. *Wang et al.* [15] proposed MART, which separates misclassified and correctly classified samples during training, including only benign examples and misclassified adversarial examples in the training data. *Kuang et al.* [16] introduced a method that leverages topological information, encouraging the model to maintain topological consistency between clean and adversarial examples in the feature space. *Li et al.* [17] proposed Core Feature-aware Adversarial Training (CoFAT), which improves robustness by guiding the model to focus on core features and minimize reliance on spurious features during training.

B. Long-Tailed Learning

In many real-world recognition tasks, training datasets often exhibit imbalanced class distributions, typically following a long-tailed pattern: a few classes have abundant examples, while most classes are underrepresented with limited training instances [18]. Models trained on such distributions tend to be biased toward frequently occurring classes, leading to skewed decision boundaries and significant performance degradation on rare categories. This imbalance not only affects overall classification accuracy but also weakens the model's ability to generalize on underrepresented classes, which are often more important in real-world applications. As a result, learning robust and unbiased feature representations under severe class imbalance remains a critical challenge in visual recognition, driving ongoing research in long-tailed learning [19].

To address these challenges, various methods have been proposed from different angles. These include data resampling strategies to rebalance class frequencies [8], loss reweighting and cost-sensitive learning approaches that prioritize minority classes [20], and decoupled training or classifier redesign techniques to mitigate representation bias between head and tail classes [21]. Recent works further enhance long-tailed recognition by refining feature learning or optimizing classifier behavior, yielding notable improvements in classification performance. However, most existing methods focus on improving standard accuracy under imbalanced data, with limited attention to robustness issues. Specifically, research on adversarial robustness [9], [10], [22] under long-tailed distributions remains scarce, leaving a significant gap in addressing model vulnerability in realistic, imbalanced scenarios.

III. The Proposed Method

In this section, we propose a novel Semantic-Guided Adversarial Training framework (SAT). This framework aims to enhance the model's robustness on long-tail distributions and balance its performance on tail classes. The overall process of the method is illustrated in Fig. 1. SAT leverages the semantic consistency of text to improve the model's robustness. Specifically, by incorporating the vision-language model CLIP to obtain semantic embeddings for each class label, and combining it with a semantic alignment loss, SAT ensures that the model's visual features align with the corresponding textual semantic representations. This imposes a balanced class constraint in the feature space, mitigating the model's bias toward the head classes. Next, we will detail the two key components of our approach: (1) Adversarial Robust Learning and (2) Adversarial Semantic Alignment.

A. Adversarial Robust Learning

In many real-world recognition tasks, training data often follows a long-tailed distribution, where a few classes have many samples, while most classes are underrepresented with limited instances. This leads to models being biased toward more frequent classes, causing skewed decision boundaries and significant performance degradation on underrepresented classes. Inspired by [10], we introduce a balanced cross-class loss, hierarchical bias loss, and rare-class emphasis loss, working together with an adversarial example generation module to balance accuracy and robustness. These losses help alleviate challenges posed by long-tailed distributions, reducing bias toward head classes and enhancing robustness in such settings. Comparing with the AT-BSL method [9], we find that AT-BSL dynamically adjusts the sample distribution based on the training progress and prediction accuracy for each class. This adaptive adjustment of the sample ratio between frequent and rare classes helps the model focus more on rare class samples. In contrast, TAET identifies weak classes through the average loss during training, strengthening both tail and weak classes to prevent neglecting the latter.

First, we use a 10-step PGD method to generate adversarial examples, then perform adversarial training by inputting both

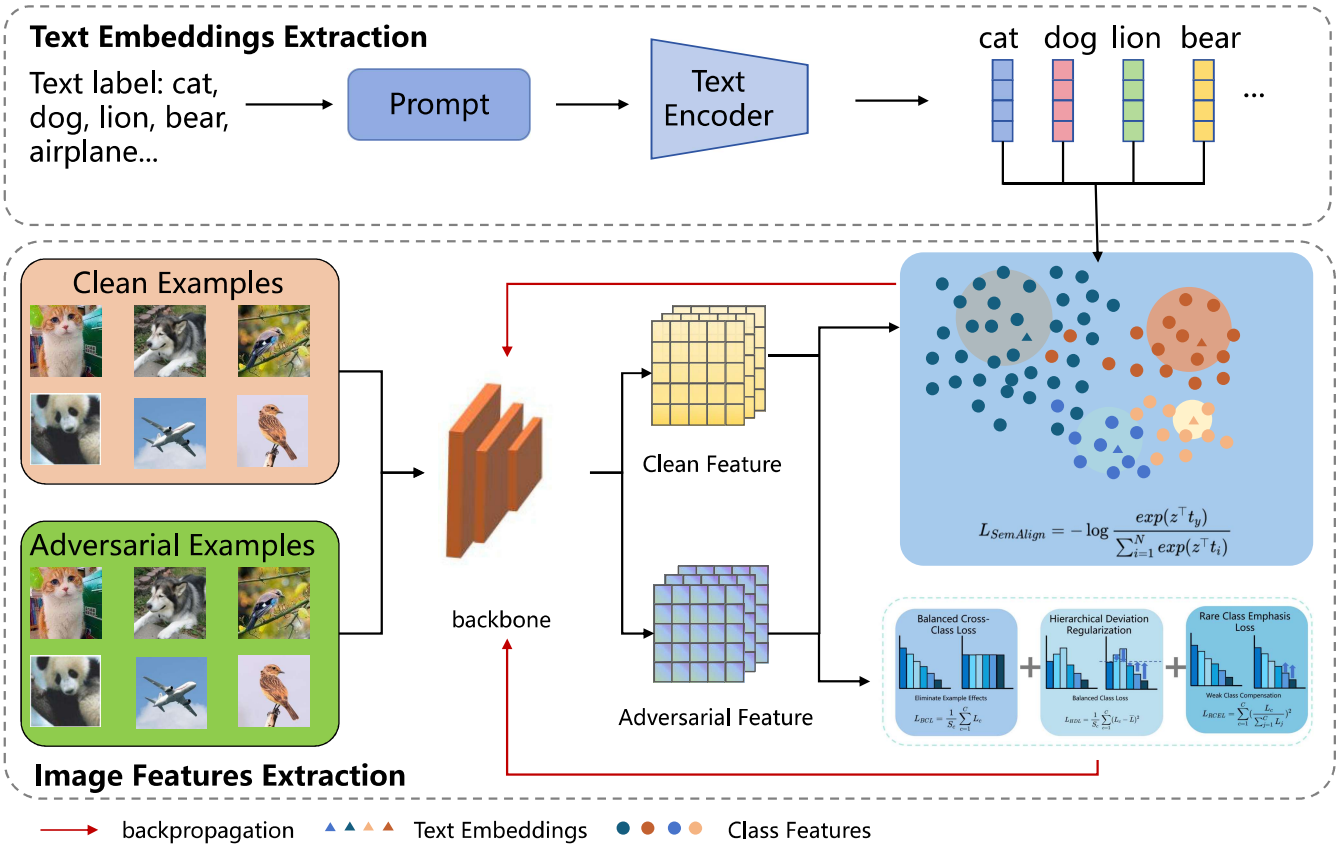


Fig. 1. The pipeline of our approach. The black arrows represent the forward propagation and inference process during training, while the red arrows represent the backpropagation process.

clean and adversarial examples into the model. For examples feeding into the model, we introduce the **balanced cross-class loss**, **hierarchical bias loss**, and **rare-class emphasis loss** to further optimize the model. The formalized loss is as follows:

$$L_{ARL} = \alpha L_{BCL} + \beta L_{HBL} + \gamma L_{RCEL}, \tag{2}$$

where α, β, and γ are hyperparameters for loss reweighting.

The balanced cross-class loss (BCL) reweights the loss to prevent the model from overemphasizing the head classes in long-tailed datasets, ensuring that the contribution of each class to the overall loss is more evenly distributed. BCL achieves this by averaging the class-specific losses, thereby improving the classifier's performance across all classes in a long-tailed distribution. The formula is as follows:

$$L_{BCL} = \frac{1}{S_i} \sum_{i=1}^{N} L_i \tag{3}$$

where S_i represents the total number of samples of the i-th class, and L_i denotes the Cross-Entropy loss associated with the i-th class.

The hierarchical bias loss (HBL) quantifies the deviation between each class's loss and the average loss, applying a quadratic penalty to reduce this gap, thereby enhancing the model's robustness to imbalanced data distributions. The formula is as follows:

$$L_{HBL} = \frac{1}{S_i} \sum_{i=1}^{N} (L_i - \overline{L})^2. \tag{4}$$

The rare class emphasis loss (RCEL) normalizes the class-specific losses and assigns higher weights to rare classes, encouraging the model to prioritize learning from these underrepresented classes, thereby improving overall classification performance. The formula is as follows:

$$L_{RCEL} = \sum_{i=1}^{N} \left(\frac{L_i}{\sum_{j=1}^{N} L_j}\right)^2. \tag{5}$$

B. Adversarial Semantic Alignment

Although the three loss constraints in Section III-A can alleviate the model's bias toward head classes, relying solely on sample-level loss reweighting and balancing strategies is insufficient to fundamentally overcome the feature representation bias caused by long-tailed data distributions. During the learning process, the model tends to over-rely on surface features that may carry data bias, learned from the rich samples of the head classes, and struggles to extract robust and highly discriminative semantic features from the few samples of the

tail classes. To address this limitation, we introduce prior knowledge, independent of sample quantity, to guide and strengthen the model's feature learning.

We leverage the semantic consistency knowledge embedded in large-scale vision-language models to construct a stable and discriminative semantic reference point for each class, namely text embeddings. This imposes a balanced constraint at the feature representation level, alleviating the model's reliance on head classes. We utilize the pre-trained CLIP text encoder to build semantic anchors for all classes. Specifically, for each class name y_n in the training set, we use a prompt template ("a photo of $\{y_n\}$") to convert it into a textual description, and extract its l_2-normalized feature vector t_n using the frozen text encoder, where $t_n = 1$. These text embeddings encode the abstract semantics of the class and preserve semantic similarity relationships in the feature space, providing the model with a stable, data-distribution-independent semantic reference framework. To align visual feature learning with the aforementioned semantic priors, we design a core semantic alignment loss. This loss function, during adversarial training, forces the model to align the visual features learned from adversarial samples with the text embedding of the corresponding class in the semantic space, while distancing itself from embeddings of other classes. Given an adversarial example x' and its true label y, the normalized feature z extracted by the visual encoder, the **semantic alignment loss** is defined as follows:

$$L_{SemAlign} = -\log \frac{exp(z^\top t_y)}{\sum_{i=1}^{N} exp(z^\top t_i)}. \tag{6}$$

Unlike the explicit balancing approach in the adversarial robust learning module that directly adjusts the loss weights of each class, this loss provides a "relative alignment" optimization objective based on similarity comparison through the softmax function. This is particularly crucial for tail classes: it does not require the tail class features to perfectly match their text embeddings, but rather encourages them to have a higher relative match to their own text embedding than to other class embeddings. This more flexible objective mitigates the optimization challenges posed by the scarcity of tail class samples, guiding the model to focus on learning more discriminative and robust semantic features, rather than surface features that are easily disrupted.

C. Overall Loss

Finally, we achieve collaborative optimization between the adversarial robust learning module and the adversarial semantic alignment module. The adversarial robust learning module addresses class imbalance from the data level by using a loss balancing mechanism, while the adversarial semantic alignment module intervenes at the knowledge level, correcting feature representation bias and enhancing semantic robustness by introducing external semantic priors. Their joint optimization objective is as follows:

$$L = L_{ARL} + \eta L_{SemAlign}, \tag{7}$$

TABLE I

BALANCED ACCURACY (%) AND BALANCED ROBUSTNESS (%) OF VARIOUS ALGORITHMS USING RESNET-18 ON CIFAR-10-LT WITH AN IMBALANCE RATIO (IR) OF 10. THE BEST RESULTS ARE HIGHLIGHTED IN **BOLD**.

Method	Clean	FGSM	PGD-20	PGD-100	CW	AA
AT-BSL	72.74	34.13	26.86	25.62	15.67	25.26
RoBal	73.18	33.14	27.12	26.98	13.44	24.13
TRADES	65.77	25.73	20.23	19.59	27.42	19.63
AT	69.00	32.53	25.69	24.55	15.22	24.28
MART	58.33	33.67	30.10	29.36	**48.12**	26.04
ADT	68.08	32.70	26.00	25.27	10.05	25.06
REAT	74.56	31.42	24.02	22.52	11.07	22.69
LAS-AT	64.04	33.04	27.80	26.74	36.77	24.71
HE	68.18	32.40	25.66	24.59	12.94	24.32
TAET	74.67	39.59	35.15	34.29	32.04	30.57
SAT(ours)	**77.85**	**45.80**	**40.99**	**39.51**	38.68	**36.76**

TABLE II

THE STANDARD ACCURACY (%) USING RESNET-18 PERFORMANCE ON CIFAR100-LT AND MINI-IMAGENET WITH AN IMBALANCE RATIO (IR) OF 10. THE BEST RESULTS ARE HIGHLIGHTED IN **BOLD**.

Dataset	Method	Clean	FGSM	PGD-20	CW	AA
CIFAR-100	TAET	50.06	25.13	21.23	20.17	18.81
	SAT(ours)	**50.88**	**25.64**	**22.65**	**21.41**	**20.02**
Mini-ImageNet	TAET	**42.96**	15.82	13.27	12.40	10.86
	SAT(ours)	41.77	**16.59**	**14.74**	**13.31**	**12.24**

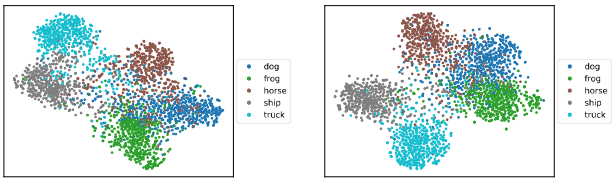
where η is the hyperparameter to balance the two parts of our method.

IV. EXPERIMENTS

A. Experimental Settings

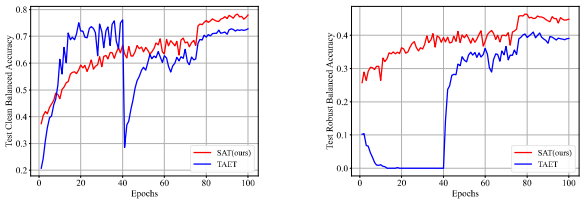
Datasets. In our experiments, we evaluate the proposed method on several widely used long-tailed classification benchmarks, including CIFAR-10-LT, CIFAR-100-LT, and Mini-ImageNet-LT. Specifically, CIFAR-10-LT and CIFAR-100-LT are constructed by resampling the original CIFAR-10 and CIFAR-100 datasets, respectively, resulting in 10 and 100 classes to simulate varying degrees of class imbalance. Mini-ImageNet-LT is a long-tailed variant built upon the Mini-ImageNet dataset, which consists of 100 classes and approximately 60,000 images with a resolution of 84×84. Compared to the CIFAR benchmarks, Mini-ImageNet-LT contains richer semantic information and greater visual diversity across categories, providing a more challenging evaluation setting.

Training Details. For all experiments, we adopt ResNet-18 as the backbone network. The model is trained using the SGD optimizer with a momentum of 0.9 and a weight decay of 5×10^{-4}, and the batch size is set to 128. The initial learning rate is set to 0.1 and decayed by a factor of 10 at the 75th and 90th epochs, respectively, with a total of 100 training epochs. During training, adversarial examples are generated using PGD-10, with a maximum perturbation budget $\epsilon = 8/255$ and a step size of $2/255$. In addition, for hyperparameter settings, we follow the configurations in [10], where α, β, γ are set to 0.1, respectively. the hyperparameter η is fixed to 9 in all experiments.



(a) TAET (b) SAT(ours)

Fig. 2. t-SNE visualization of latent space logits extracted from (a) TAET and (b) our proposed SAT method on tail classes (last five classes).



(a) Clean Accuracy (b) Robust Accuracy

Fig. 3. Results of clean accuracy and robust accuracy (under PGD-10) over epochs on CIFAR-10-LT of TAET and our method in the training process. The models are trained under l_∞ norm with $\epsilon = 8/255$ on ResNet-18.

TABLE III
BALANCED ACCURACY (%) AND BALANCED ROBUSTNESS (%) USING RESNET-18 ON CIFAR-10-LT UNDER DIFFERENT LONG-TAIL IMBALANCE RATIOS(IRS). BETTER RESULTS ARE HIGHLIGHTED IN **BOLD**.

IR	Method	Clean	FGSM	PGD-20	CW	AA
10	AT-BSL	72.74	34.13	26.86	15.67	25.26
	TAET	74.67	39.59	35.15	32.04	30.57
	SAT(ours)	**77.85**	**45.80**	**40.99**	**38.68**	**36.76**
20	AT-BSL	66.40	28.82	22.56	15.43	21.46
	TAET	66.93	36.82	33.98	31.75	27.84
	SAT(ours)	**70.62**	**41.24**	**37.10**	**35.70**	**33.96**
50	AT-BSL	58.23	26.37	20.87	16.74	19.79
	TAET	59.37	31.69	29.35	28.27	25.30
	SAT(ours)	**64.18**	**38.51**	**34.08**	**32.92**	**31.05**
100	AT-BSL	54.73	22.51	19.17	17.34	18.29
	TAET	52.68	28.68	26.53	24.55	22.98
	SAT(ours)	**58.68**	**30.41**	**28.75**	**26.99**	**26.12**

TABLE IV
PERFORMANCE (%) OF VARIOUS METHODS ON THE LAST TAIL CLASSES (LAST FIVE CLASSES) UNDER CLEAN AND ATTACKED CONDITIONS USING RESNET-18 ON CIFAR-10-LT(IR=10). THE BEST RESULTS ARE HIGHLIGHTED IN **BOLD**, AND THE SECOND-BEST RESULTS ARE UNDERLINED.

Method	Clean					Attacked(PGD-20)				
	Dog	Frog	Horse	Ship	Truck	Dog	Frog	Horse	Ship	Truck
AT-BSL	58.63	73.48	70.48	78.29	66	23.02	26.97	21.7	21.7	19
RoBal	59.13	74.73	68.52	78.96	65	23.74	25.78	20.79	19.5	20
TRADES	53.96	65.11	63.25	47.28	47	11.87	13.2	16.27	3.87	4
AT	51.07	62.32	63.85	64.34	54	14.74	17.67	19.27	17.05	10
MART	37.76	52.09	58.43	47.25	41	17.62	19.06	25.3	8.52	9
ADT	52.15	65.58	65.66	56.59	55	16.18	16.74	18.67	10.07	10
REAT	58.63	73.02	72.89	78.29	69	16.18	13.48	18.67	20.15	17
LAS-AT	50.72	55.34	56.02	48.83	34	19.06	13.02	20.48	12.4	8
HE	57.31	67.90	65.66	51.16	54	14.74	20	18.07	10.85	12
TAET	56.11	79.06	72.28	80.60	77	23.74	34.83	32.53	36.34	27
SAT(ours)	**64.39**	**86.05**	**81.33**	**82.17**	**82**	**30.58**	**45.12**	**44.58**	**46.51**	**33**

Evaluation Set. To better evaluate the adversarial robustness under long-tailed distributions, we introduce the balanced accuracy and balanced robustness accuracy from [10] as robustness metrics to assess our method, while the degree of class imbalance is measured by the imbalance ratio (IR). Model robustness is assessed under l_∞-bounded perturbations with a maximum budget of $8/255$. Specifically, FGSM and PGD attacks are employed for evaluation, where PGD is conducted with 20 and 100 iterations using a step size of $1/255$. In addition, a 100-step CW attack and AutoAttack (AA), a strong ensemble-based evaluation protocol, are included to provide a comprehensive assessment of robustness under varying attack strengths. The proposed approach is compared with a range of representative adversarial training and long-tailed robustness baselines, including AT-BSL [9], RoBal [22], TRADES [14], AT [23], MART [15], ADT [24], REAT [25], LAS-AT [26], HE [27], and TAET [10].

B. Adversarial Robustness Analysis

As shown in Table I, SAT achieves the best performance on CIFAR-10-LT (IR = 10) under both clean and adversarial settings. Under clean conditions, SAT reaches a balanced accuracy of 77.85%, improving over TAET by 3.18%. The advantage becomes more pronounced under adversarial attacks. SAT achieves 45.80% robustness under FGSM, which is 6.21% higher than TAET. Under stronger iterative attacks, SAT attains 40.99% and 39.51% robustness under PGD-20 and PGD-100, improving over TAET by 5.84% and 5.22%, respectively. Even under AutoAttack, SAT achieves 36.76% robustness, exceeding TAET by 6.19%. These results indicate that SAT substantially enhances adversarial robustness while also improving clean accuracy, rather than trading off accuracy for robustness. While some competing methods obtain strong results under certain attacks, their performance varies notably across attack strengths, whereas SAT remains consistently superior across all evaluated settings. Additionally, we visualized the results of our training process. As shown in Fig. 2, we visualized the t-SNE plot of tail class samples processed by the model, where it is evident that our method achieves clearer decision boundaries compared to TAET. Similarly, as shown in Fig. 3, we visualized the accuracy change during the training process, and it is clear that our method outperforms TAET. From Fig. 2 and Fig. 3, we can observe that our method significantly outperforms the baseline in terms of performance.

As further reported in Table II, SAT is evaluated on CIFAR-100-LT and Mini-ImageNet-LT. On CIFAR-100-LT, SAT achieves 50.88% clean accuracy, surpassing TAET by 0.82%, and consistently improves robustness under different attacks. In particular, SAT improves robustness by 1.42% under PGD-20 and by 1.21% under AutoAttack. On Mini-ImageNet-LT, SAT still improves robustness by 1.47% under PGD-20 and by 1.38% under AutoAttack, but its clean accuracy is slightly lower than that of TAET. This suggests that on long-tailed datasets with higher semantic complexity and greater visual diversity, SAT remains effective in improving robustness, yet it becomes more challenging to simultaneously maintain clean accuracy, leaving room for further improvement

in more complex and large-scale long-tailed scenarios.

We further conduct experiments on CIFAR-10-LT under different imbalance ratios, and the results are reported in Table III. These experiments systematically characterize how clean accuracy and adversarial robustness evolve as the imbalance ratio increases from 10 to 100. Overall, increasing class imbalance negatively affects all methods, while the rate and magnitude of performance degradation differ substantially, particularly under strong adversarial attacks. From the perspective of clean accuracy, TAET maintains a relatively high level at IR=10, but its accuracy drops to 52.68% when the imbalance ratio increases to 100, whereas SAT still preserves a clean accuracy of 58.68% under the same setting. This gap indicates that under extreme class scarcity, TAET is more prone to bias toward head classes, while SAT suppresses such bias more effectively and maintains more stable classification performance. This difference is more pronounced in terms of adversarial robustness. Although TAET achieves reasonable robustness under PGD-20 and AutoAttack at IR=10, its robustness degrades more rapidly as the imbalance ratio increases to 50 and 100. Under AutoAttack at IR=100, TAET drops to 22.98%, while SAT remains at 26.12%, suggesting that SAT experiences a slower robustness degradation when strong adversarial perturbations are combined with severe class imbalance. A further comparison with AT-BSL shows that under moderate and high imbalance ratios, AT-BSL suffers more severe robustness degradation under strong attacks, with consistently lower performance than SAT under PGD-20 and AutoAttack. This observation indicates that loss reweighting or output-level adjustment alone is insufficient in extreme long-tailed settings, whereas SAT establishes a more effective interaction between long-tailed structure and adversarial training, leading to improved robustness generalization across different imbalance levels.

We additionally evaluate the tail classes of CIFAR-10-LT, as shown in Table IV. This analysis focuses on the five classes with the fewest samples under clean conditions and PGD-20 attacks. Under clean settings, several tail classes already achieve relatively high accuracy with TAET; however, this behavior does not persist once adversarial perturbations are introduced, indicating that clean accuracy alone is insufficient to reflect model stability on tail classes. Under PGD-20 attacks, TAET exhibits substantial performance degradation on tail classes. The accuracy of the Horse class drops from 72.28% under clean conditions to 32.53%, and the Frog class retains only 34.83% robustness, showing that the decision boundaries learned by TAET for tail classes are easily disrupted by strong adversarial perturbations. In contrast, SAT maintains higher robustness under the same attack setting, reaching 44.58% on Horse and 45.12% on Frog, indicating reduced sensitivity to adversarial perturbations. The response to SAT varies across different tail classes: classes with higher intra-class variability, such as Horse and Ship, exhibit larger robustness gains, whereas relatively simpler classes, such as Dog and Truck, show more moderate improvements. These results indicate that SAT more effectively mitigates bias toward head classes and slows down performance degradation on tail classes under adversarial attacks, enabling more stable predictions in challenging long-tailed scenarios.

TABLE V
BALANCED ROBUSTNESS (%) RESULTS OF RESNET-18 ON CIFAR-10-LT (IR=10) UNDER DIFFERENT HYPERPARAMETER SETTINGS.

η	Clean	FGSM	PGD-20	PGD-100	CW	AA
0.5	76.89	46.00	40.51	39.29	39.16	36.48
0.8	75.97	45.01	38.90	37.94	38.32	36.13
1.00	77.54	45.96	40.08	39.08	39.09	37.25
2.00	77.05	45.99	40.17	39.22	38.77	37.07
3.00	76.74	46.10	40.87	39.24	39.14	37.16
4.00	76.68	45.45	40.42	38.80	39.87	37.26
5.00	76.90	45.34	38.61	37.58	37.68	35.69
6.00	78.56	45.78	39.52	37.97	38.04	36.07
7.00	75.26	45.33	39.91	38.68	38.38	36.58
8.00	76.04	45.21	39.09	37.87	37.52	35.77
9.00	77.85	45.80	40.99	39.51	38.68	36.76
10.00	77.68	44.59	39.33	38.14	38.50	36.14

C. Ablation Study

We conduct ablation experiments on the hyperparameter η to analyze its impact on model performance, and the results are reported in Table V. In this ablation study, all training settings are kept fixed except for the value of η, and the model is evaluated on CIFAR-10-LT (IR = 10) under clean conditions as well as multiple adversarial attacks using balanced robustness metrics. From the overall trend, model performance varies within a relatively narrow range as η changes across a wide interval. The clean accuracy consistently remains between 75% and 78%, while the robustness under PGD-20 and AutoAttack does not exhibit drastic fluctuations, indicating that the training process is reasonably stable with respect to this hyperparameter. A closer inspection shows that when η is set to very small or excessively large values, robustness under certain attack settings tends to decrease, whereas moderately larger values of η lead to a more balanced trade-off across different attacks. Considering the results under varying attack strengths, $\eta = 9$ achieves the highest robustness under PGD-20 and PGD-100, reaching 40.99% and 39.51%, respectively. At the same time, it attains the second-highest clean accuracy among all tested hyperparameter settings and maintains competitive performance under AutoAttack. Therefore, we adopt $\eta = 9$ as the default setting in all our experiments.

V. CONCLUSION

In this paper, we propose a semantic-guided adversarial training framework. This method introduces semantic prior knowledge, independent of sample quantity, embedded in large-scale vision-language models, as an additional supervisory signal into the adversarial training process. By utilizing a pre-trained text encoder to generate semantically consistent text embeddings for each class and designing a corresponding semantic alignment loss, we impose more balanced constraints on all classes under long-tailed distributions. This effectively alleviates the model's inherent bias toward head samples and

significantly enhances its robustness on tail classes. Experimental results demonstrate that the proposed method achieves state-of-the-art performance on multiple long-tailed benchmark datasets. Compared to existing advanced methods, our approach not only improves overall adversarial robustness but, more importantly, exhibits superior robustness on tail classes, achieving a fairer distribution of robustness.

VI. Acknowledgment

This work was supported in part by the National Key R&D Program of China (2023YFE0116300), in part by the National Natural Science Foundation of China (Grant No.62006097, U1836218).

References

[1] J. Qian, L. Xu, X. Ren, and X. Wang, "Structured deep neural network-based backstepping trajectory tracking control for lagrangian systems," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 36, no. 6, pp. 11 650–11 656, 2025.

[2] D. Chen, J. Liu, and X. Che, "On the upper bounds of number of linear regions and generalization error of deep convolutional neural networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 47, no. 6, pp. 5014–5022, 2025.

[3] P. Gao, J. Chen, X. Che, F. Liu, Y. Lu, Y. Guan, and J. He, "Attention-driven deep neural networks with cross-channel temporal modeling for robust cybersecurity situational awareness," *IEEE Transactions on Information Forensics and Security*, vol. 20, pp. 11 223–11 237, 2025.

[4] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," *CoRR*, vol. abs/1412.6572, 2014.

[5] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. J. Goodfellow, and R. Fergus, "Intriguing properties of neural networks," *CoRR*, vol. abs/1312.6199, 2013.

[6] J. Yang, X. Wang, M. Zhang, L. Liu, and J. Li, "Adaptive dynamic causal meta graph-task network for remaining useful life prediction with extreme long-tailed distribution condition," *IEEE Transactions on Industrial Informatics*, vol. 21, no. 9, pp. 7232–7242, 2025.

[7] Z. Li, H. Zhao, A. Gao, D. Guo, T.-H. Chang, and X. Wan, "Prototype-oriented clean subset extraction for noisy long-tailed classification," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 35, no. 8, pp. 7953–7965, 2025.

[8] N. Chang, Z. Yu, Y.-X. Wang, A. Anandkumar, S. Fidler, and J. M. Alvarez, "Image-level or object-level? a tale of two resampling strategies for long-tailed detection," in *International conference on machine learning*. PMLR, 2021, pp. 1463–1472.

[9] X. Yue, N. Mou, Q. Wang, and L. Zhao, "Revisiting adversarial training under long-tailed distributions," in *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024, pp. 24 492–24 501.

[10] W. Yu-Hang, J. Guo, A. Liu, K. Wang, Z. Wu, Z. Liu, W. Yin, and J. Liu, "Taet: Two-stage adversarial equalization training on long-tailed distributions," in *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025, pp. 15 476–15 485.

[11] L. Zhang, C. Wu, and N. Yang, "Weakly supervised contrastive adversarial training for learning robust features from semi-supervised data," in *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025, pp. 25 718–25 727.

[12] N. Papernot and P. McDaniel, "Extending defensive distillation," *arXiv preprint arXiv:1705.05264*, 2017.

[13] H. Sun, T. Zhu, J. Li, and W. Zhou, "Average and strict gan-based reconstruction for adversarial example detection," *IEEE Transactions on Dependable and Secure Computing*, vol. 22, no. 4, pp. 4344–4361, 2025.

[14] H. Zhang, Y. Yu, J. Jiao, E. P. Xing, L. E. Ghaoui, and M. I. Jordan, "Theoretically principled trade-off between robustness and accuracy," *ArXiv*, vol. abs/1901.08573, 2019.

[15] Y. Wang, D. Zou, J. Yi, J. Bailey, X. Ma, and Q. Gu, "Improving adversarial robustness requires revisiting misclassified examples," in *International Conference on Learning Representations*, 2020.

[16] H. Kuang, H. Liu, X. Lin, and R. Ji, "Defense against adversarial attacks using topology aligning adversarial training," *IEEE Transactions on Information Forensics and Security*, vol. 19, pp. 3659–3673, 2024.

[17] F. Li, K. Li, H. Wu, J. Tian, and J. Zhou, "Toward robust learning via core feature-aware adversarial training," *IEEE Transactions on Information Forensics and Security*, vol. 20, pp. 6236–6251, 2025.

[18] C. Zhang, G. Almpanidis, G. Fan, B. Deng, Y. Zhang, J. Liu, A. Kamel, P. Soda, and J. Gama, "A systematic review on long-tailed learning," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 36, no. 8, pp. 13 670–13 690, 2025.

[19] W. Chen, K. Yang, Z. Yu, Y. Shi, and C. P. Chen, "A survey on imbalanced learning: latest research, applications and future directions," *Artificial Intelligence Review*, vol. 57, no. 6, p. 137, 2024.

[20] B. K. Baloch, S. Kumar, S. Haresh, A. Rehman, and T. Syed, "Focused anchors loss: Cost-sensitive learning of discriminative features for imbalanced classification," in *Asian Conference on Machine Learning*. PMLR, 2019, pp. 822–835.

[21] C. Cong, S. Xuan, S. Liu, S. Zhang, M. Pagnucco, and Y. Song, "Decoupled optimisation for long-tailed visual recognition," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 38, no. 2, 2024, pp. 1380–1388.

[22] T. Wu, Z. Liu, Q. Huang, Y. Wang, and D. Lin, "Adversarial robustness under long-tailed distribution," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 8659–8668.

[23] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, "Towards deep learning models resistant to adversarial attacks," *ArXiv*, vol. abs/1706.06083, 2017.

[24] Y. Dong, Z. Deng, T. Pang, J. Zhu, and H. Su, "Adversarial distributional training for robust deep learning," *Advances in neural information processing systems*, vol. 33, pp. 8270–8283, 2020.

[25] G. Li, G. Xu, and T. Zhang, "Adversarial training over long-tailed distribution," *arXiv preprint arXiv:2307.10205*, vol. 1, no. 4, p. 6, 2023.

[26] X. Jia, Y. Zhang, B. Wu, K. Ma, J. Wang, and X. Cao, "Las-at: Adversarial training with learnable attack strategy," in *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 13 388–13 398.

[27] T. Pang, X. Yang, Y. Dong, K. Xu, J. Zhu, and H. Su, "Boosting adversarial training with hypersphere embedding," *Advances in Neural Information Processing Systems*, vol. 33, pp. 7779–7792, 2020.