Revisiting Adversarial Training under Long-Tailed Distributions

Xinli Yue, Ningping Mou, Qian Wang, Lingchen Zhao*
School of Cyber Science and Engineering, Wuhan University, Wuhan 430072, China
{xinliyue,ningpingmou,qianwang,lczhaocs}@whu.edu.cn

Abstract

Deep neural networks are vulnerable to adversarial attacks, leading to erroneous outputs. Adversarial training has been recognized as one of the most effective methods to counter such attacks. However, existing adversarial training techniques have predominantly been evaluated on balanced datasets, whereas real-world data often exhibit a long-tailed distribution, casting doubt on the efficacy of these methods in practical scenarios. In this paper, we delve into the performance of adversarial training under long-tailed distributions. Through an analysis of the prior method "RoBal"(Wu et al., CVPR'21), we discover that utilizing Balanced Softmax Loss (BSL) alone can obtain comparable performance to the complete RoBal approach while significantly reducing the training overhead. Then, we reveal that adversarial training under long-tailed distributions also suffers from robust overfitting similar to uniform distributions. We explore utilizing data augmentation to mitigate this issue and unexpectedly discover that, unlike results obtained with balanced data, data augmentation not only effectively alleviates robust overfitting but also significantly improves robustness. We further identify that the improvement is attributed to the increased diversity of training data. Extensive experiments further corroborate that data augmentation alone can significantly improve robustness. Finally, building on these findings, we demonstrate that compared to RoBal, the combination of BSL and data augmentation leads to a +6.66% improvement in model robustness under AutoAttack on CIFAR-10-LT. Our code is available at: https://github.com/NISPLab/AT-BSL.

1. Introduction

It is well-known that deep neural networks (DNNs) are vulnerable to adversarial attacks, where attackers can induce errors in the recognition results of DNNs by adding slight perturbations to the inputs [12, 38]. Many researchers have focused on defending against such attacks. Among the various defense methods proposed, adversarial training is recognized as one of the most effective approaches. Its insight is integrating adversarial examples into the training set to improve the generalization capability of the model against these examples [19, 30, 41, 44, 52, 53]. In recent years, significant progress has been made in the field of adversarial training. However, we observe that almost all works utilize balanced datasets such as CIFAR-10, CIFAR-100 [22], and Tiny-ImageNet [23] for performance evaluation. In contrast, real-world datasets often exhibit an imbalanced, typically long-tailed distribution. Hence, the efficacy of adversarial training in practice should be reassessed using long-tailed datasets [14, 39].

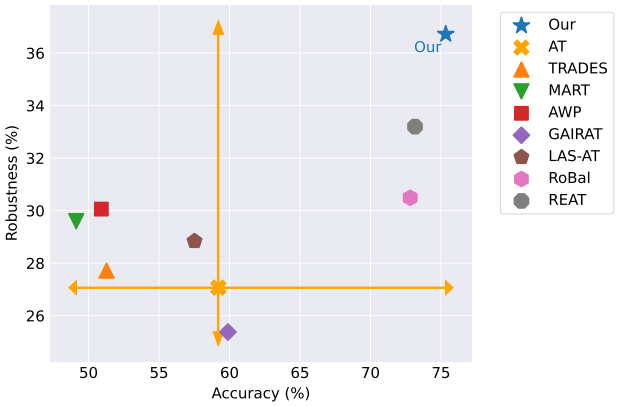


Figure 1. The clean accuracy and robustness under AutoAttack (AA) [5] of various adversarial training methods using WideResNet-34-10 [50] on CIFAR-10-LT [22]. Our method, building upon AT [30] and BSL [35], leverages data augmentation to improve robustness, achieving a +6.66% improvement over the SOTA method RoBal [45]. REAT [25] is a concurrent work with ours, yet to be published.

To our knowledge, RoBal [45] is the sole published work investigating the adversarial robustness under the long-tailed distribution. However, its complex design causes extensive training time and GPU memory, somewhat limiting its usability. Upon revisiting the design of RoBal, we notice that its most critical component is the Balanced Softmax Loss (BSL) [35]. We observe that combining AT [30] with

*Corresponding author.

BSL to form AT-BSL can match the effectiveness of RoBal while significantly reducing its training overhead. Hence, we advocate using AT-BSL as a substitute for RoBal.

In addition, we encounter another important observation: similar to training under balanced datasets, adversarial training with long-tailed distribution data also leads to the issue of robust overfitting [36]. Some prior works on balanced datasets employed data augmentation to mitigate this issue [4, 13, 34, 36, 44]. Hence, a straightforward approach is to attempt to introduce data augmentation in adversarial training under long-tailed distribution. Our results partially align with the results on balanced datasets, indicating that data augmentation can also mitigate robust overfitting. However, contrary to results on balanced datasets where only utilizing data augmentation alone is unhelpful for improving robustness [34, 36, 44], we find that data augmentation techniques, including MixUp [51], Cutout [9], CutMix [49], AugMix [17], AutoAugment (AuA) [6], RandAugment (RA) [7] and TrivialAugment (TA) [32], can significantly improve robustness. Hence, we further introduce the question: Why does data augmentation improve robustness? We hypothesize that data augmentation can increase the diversity of training data, enabling the model to learn richer representations and thereby improving robustness. We validate this hypothesis through ablation studies.

Our contributions are summarized as follows:

- Through ablation studies, we discover that BSL is the most critical component of RoBal, and the streamlined method AT-BSL can significantly reduce the training time and memory usage of RoBal.
- We observe that data augmentation not only mitigates robust overfitting in adversarial training under long-tailed distributions but also substantially improves robustness.
- We propose a hypothesis about how data augmentation improves robustness and validate this hypothesis through experiments.
- Comprehensive empirical evidence demonstrates that our discoveries generalize across multiple common data augmentation strategies, model architectures, and datasets.

2. Related Works

Long-Tailed Learning. Long-tailed distributions refer to a common imbalance in training datasets where a small portion of classes (head) have massive examples, while other classes (tail) have very few examples [14, 39]. Models trained under such distribution tend to exhibit a bias towards the head classes, resulting in poor performance for the tail classes. Traditional rebalancing techniques aim at addressing the long-tailed recognition problem include re-sampling [20, 37, 40, 55] and cost-sensitive learning [8, 28], which often improve the performance of tail classes at the expense of head classes. To mitigate these adverse effects, some methods handle class-specific attributes through perspectives such as margins [42] and biases [35]. Recently, more advanced techniques like class-conditional sharpness-aware minimization [56], feature clusters compression [26], and global-local mixture consistency cumulative learning [10] have been introduced, further improving the performance of long-tailed recognition. However, these works have been devoted to improving clean accuracy, and investigations into the adversarial robustness of long-tailed recognition remain scant.

Adversarial Training. The insight of adversarial training is integrating adversarial examples into the training set, thereby improving the generalizability of the model to such examples. Theoretically, adversarial training addresses a min-max problem, where the inner maximization generates the most powerful adversarial examples, and the outer minimization optimizes the model parameters. One of the most typical adversarial training methods is AT [30], which can be mathematically represented as follows:

$$\begin{aligned} &\underset{\theta_m}{\operatorname{argmin}} \mathcal{L}_{\min}\left(\theta_m; x', y\right), \\ &\text{where } x' = \underset{\|x'-x\|_p \leq \epsilon}{\operatorname{argmax}} \mathcal{L}_{\max}\left(\theta_m; x', y\right). \end{aligned} \tag{1}$$

where x' is an adversarial example constrained by the ℓ_p norm for the clean example x, y is the label of x, θ_m is the parameter of the model m, ϵ is the perturbation size, $\mathcal{L}_{\max}$ is the internal maximization loss, and $\mathcal{L}_{\min}$ is the external minimization loss.

Building upon the foundation of AT [30], subsequent works developed advanced adversarial training techniques such as TRADES [52], MART [41], AWP [44], GAIRAT [53], and LAS-AT [19].

Robustness under Long-Tailed Distribution. Unfortunately, previous works about adversarial training were mainly concerned with balanced datasets, but data in the real world more commonly follow long-tailed distributions [14, 39]. Therefore, a critical criterion for assessing the practical utility of adversarial training should be its performance on long-tailed distributions. To our knowledge, RoBal [45] is the only published work that investigates adversarial training on long-tailed datasets. In Section 3, we will conduct a detailed analysis of the design of RoBal.

Data Augmentation. Data augmentation has been recognized as an effective tool to mitigate overfitting and improve the generalization capability of models, irrespective of whether the distribution of the training data is balanced or long-tailed [2, 10, 47, 54]. Commonly employed data augmentation techniques in image classification tasks include random flips, rotations, and crops [15]. Some more advanced augmentation methods, such as MixUp [51], Cutout [9], and CutMix [49] may deliver better performance in standard training scenarios. Moreover, augmentation strategies such as Augmix [17], AuA [6], RA [7],

and TA [32], which integrate a learned or random combination of multiple augmentations, demonstrate superior performance.

3. Analysis of RoBal

3.1. Preliminaries

RoBal [45], compared to AT[30], introduces four additional components: 1) Cosine Classifier; 2) Balanced Softmax Loss[35]; 3) Class-aware Margin; and 4) TRADES Regularization [52].

Cosine Classifier. In basic classification tasks employing a standard linear classifier, the predicted logit for class i can be represented as follows:

$$\begin{aligned} g(f(x))_i &= W_i^T f(x) + b_i \\ &= \|W_i\| \cdot \|f(x)\| \cos\theta_i + b_i \\ &= z_i + b_i, \end{aligned} \tag{2}$$

where $g(\cdot)$ is the liner classifer. This formulation indicates that three factors influence the prediction result: 1) the magnitude of the weight vector $\|W_i\|$ and the feature vector $\|f(x)\|$; 2) the angle between them, denoted as $\cos\theta_i$; and 3) the bias term of the classifier b_i.

This decomposition highlights that the prediction results of the examples can be altered by adjusting the norm of examples in the feature space. In linear classifiers, the scale of the weight vector $\|W_i\|$ tends to decrease in tail classes, thereby impacting the accuracy for these classes. Consequently, [45] aims to employ a cosine classifier [33] to minimize the scale effects of features and weights. In the cosine classifier, the predicted logit for class i can be represented as follows:

$$\begin{aligned} h(f(x))_i &= s \cdot \left(\frac{W_i^T f(x)}{\|W_i\| \, \|f(x)\|} \right) + b_i \\ &= s \cdot \cos\theta_i + b_i, \end{aligned} \tag{3}$$

where $h(\cdot)$ is the cosine classifier, $\|\cdot\|$ denotes the ℓ_2 norm of the vector, s is the scaling factor.

Balanced Softmax Loss. An intuitive and widely adopted approach to address class imbalance is assigning class-specific biases during training for cross-entropy loss. [45] employs the approach outlined in [31, 35], where the bias is defined as $b_i = \tau_b \log(n_i)$. This modification leads to the Balanced Softmax Loss (**BSL**), which is formulated as:

$$\begin{aligned} &\mathcal{L}_0(h(f(x)), \mathrm{y}) = -\log\left(\frac{e^{s\cdot\cos\theta_y + b_y}}{\sum_i e^{s\cdot\cos\theta_i + b_i}} \right) \\ &= \log\left(1 + \sum_{i\neq y} e^{s\cdot(\cos\theta_i - \cos\theta_y) + \tau_b \log\left(\frac{n_i}{n_y}\right)} \right), \end{aligned} \tag{4}$$

where n_i is the number of examples in the i-th class, and τ_b is a hyperparameter controling the calculation of the bias. BSL dynamically adjusts to the label distribution shift between training and testing by incorporating specific biases for each class based on their respective example counts, thereby enhancing the performance in long-tailed learning [35].

Class-Aware Margin. Yet, when the margin from the true class y to any other class i, expressed as $\tau_b \log(n_i/n_y)$, becomes negative (i.e., when $n_y > n_i$), it may degrade the quality of discriminative representations and the learning performance of the classifier, particularly for head classes. To address this, [45] designs a class-aware margin term [33], which assigns a larger margin value to head classes as a form of compensation:

$$m_i = \frac{\tau_m}{s} \log \frac{n_i}{n_{\min}} + m_0. \tag{5}$$

The first term increases with n_i and reaches its minimum of zero when $n_i = n_{\min}$, with τ_m as the hyperparameter controlling the trend. The second term, $m_0 > 0$, establishes a universal boundary for all classes. Add this class-aware margin m_i to $\mathcal{L}_0$ to become $\mathcal{L}_1$:

$$\begin{aligned} &\mathcal{L}_1(h(f(x)), \mathrm{y}) \\ &= -\log\left(\frac{e^{s(\cos\theta_y - m_y) + b_y}}{e^{s(\cos\theta_y - m_y) + b_y} + \sum_{i\neq y} e^{s\cos\theta_i + b_i}} \right). \end{aligned} \tag{6}$$

TRADES Regularization. [45] incorporates a Kullback-Leibler (KL) regularization term following TRADES [52], thereby modifying the overall loss function to:

$$\mathcal{L}_{\min} = \mathcal{L}_1\left(h\left(f\left(x'\right)\right), \mathrm{y}\right) + \beta \cdot \mathrm{KL}\left(h\left(f\left(x'\right)\right), h(f(x))\right), \tag{7}$$

where β is a hyperparameter for controlling the intensity of the TRADES regularization.

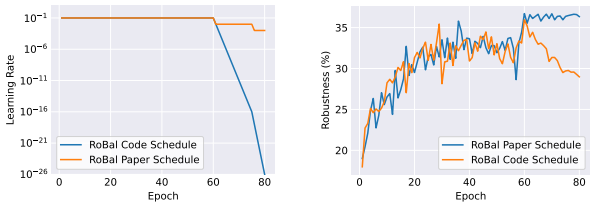
3.2. Ablation Studies of RoBal

We conduct ablation studies to investigate the contribution of each component contribution within RoBal [45]. Specifically, each component of RoBal is sequentially incorporated into AT [30] to examine their impact on clean accuracy, adversarial robustness, training time per epoch, and memory usage. The results are summarized in Table 1. Note that the selected hyperparameters selected strictly conform to the default settings of [45]. Further details about the adversarial attacks are provided in Section 5.1.

We observe that the AT augmented with Balanced Softmax Loss (AT-BSL) outperforms the vanilla AT in both clean accuracy and adversarial robustness. Nevertheless, the addition of a cosine classifier to AT-BSL improves robustness against PGD attacks[30], but robustness against adaptive attacks such as CW[3], LSA [18], and AA [5] significantly declines. This observation aligns with insights

Table 1. The clean accuracy, robustness, time (average per epoch) and memory (GPU) of ResNet-18 [15] on CIFAR-10-LT following the integration of components from RoBal [45] into AT [30]. The best results are **bolded**. The second best results are underlined. Cos: Cosine Classifier; BSL: Balanced Softmax Loss [35]; CM: Class-aware Margin [45]; TRADES: TRADES Regularization [52].

Method	Components				Accuracy						Efficiency	
	Cos	BSL	CM	TRADES	Clean	FGSM	PGD	CW	LSA	AA	Time (s)	Memory (MiB)
AT [30]					54.91	32.21	28.05	28.28	28.73	26.75	21.36	**946**
AT-BSL		✓			70.21	37.44	31.91	**31.45**	**32.25**	29.48	**21.00**	**946**
AT-BSL-Cos	✓	✓			**71.99**	39.41	34.73	30.27	29.94	28.43	22.39	**946**
AT-BSL-Cos-TRADES	✓	✓		✓	69.31	39.62	34.87	30.19	30.15	28.64	38.91	1722
RoBal [45]	✓	✓	✓	✓	70.34	**40.50**	**35.93**	31.05	31.10	**29.54**	39.03	1722



(a) (b)

Figure 2. Analysis of learning rate scheduling for RoBal [45]. (a) Comparison of the learning rate schedules: 'RoBal Code Schedule' from the source code and 'RoBal Paper Schedule' as described in the publication. (b) The evolution of test robustness under PGD-20 [30] using ResNet-18 on CIFAR-10-LT across training epochs.

from REAT [25], which suggests that the scale-invariant nature of the cosine classifier in RoBal may induce gradient vanishing during the crafting of adversarial examples with cross-entropy loss. This phenomenon is attributed to the normalization of weights and features in the classification layer, which considerably diminishes the magnitude of the gradient, thus hampering the generation of effective adversarial examples [25]. Subsequent integration of TRADES regularization and class-aware margin do not yield significant improvements in robustness against AA, yet substantially increase both training time and memory usage. In fact, AT-BSL in solation competes with the complete RoBal scheme in terms of clean accuracy and robustness against AA. Hence, we advocate using AT-BSL, which renders adversarial training more efficient without sacrificing significant performance. The $\mathcal{L}_{min}$ formulation for AT-BSL is presented as follows:

$$\begin{aligned} \mathcal{L}_{\text{min}} &= \mathcal{L}_0\left(g\left(f\left(x'\right)\right), \text{y}\right) \\ &= -\log\left(\frac{e^{z_y+b_y}}{\sum_i e^{z_i+b_i}}\right) \\ &= -\log\left(\frac{n_y^{\tau_b}\cdot e^{z_y}}{\sum_i n_i^{\tau_b}\cdot e^{z_i}}\right). \end{aligned} \tag{8}$$

3.3. Robust Overfitting and Unexpected Discoveries

Discrepancy in Learning Rate Scheduling: Paper Description vs. Code Implementation. RoBal [45] asserts that early stopping is not employed, and the reported results are from the last training epoch, i.e., the 80th epoch. The learning rate schedule described in the publication starts at 0.1, with decays at the 60th and 70th epochs, each by a factor of 0.1. However, upon running the source code of RoBal, we observe, as depicted by the blue line in Fig. 2(b), that test robustness remains essentially unchanged after the first learning rate decay (60th epoch), indicating an absence of robust overfitting. It is well-known that adversarial training techniques on balanced datasets, e.g., CIFAR-10, exhibit the robust overfitting issue [36], and CIFAR-10-LT, having fewer data points than CIFAR-10, should theoretically exacerbate this issue, contradicting the notion that more data mitigates robust overfitting as suggested in [34].

Upon a meticulous examination of the official code provided by RoBal, we discover inconsistencies between the implemented learning rate schedule and what was claimed in the publication. The code follows a schedule that begins at 0.1, with a decay of 0.1 per epoch after the 60th epoch and then 0.01 per epoch after the 75th epoch (depicted by the blue line in Fig. 2(a)). This leads to a learning rate as low as 1e-26 by the 80th epoch, potentially limiting the training after the 60th epoch and contributing to the similar performance of the model at the 60th and 80th epochs as shown in Fig. 2(b).

We then adjust the learning rate schedule to match the one declared in [45] (depicted by the orange line in Fig. 2(a)) and redraw the robustness curve, shown by the orange line in Fig. 2(b). After this adjustment, we observe a decline in test robustness post the first learning rate decay, consistent with the expected robust overfitting phenomenon typically observed on CIFAR-10.

Thus, adversarial training under long-tailed distributions also exhibits robust overfitting, similar to that under balanced distributions. The question then arises: How can we address this issue? Several studies [4, 13, 34, 36, 44] have

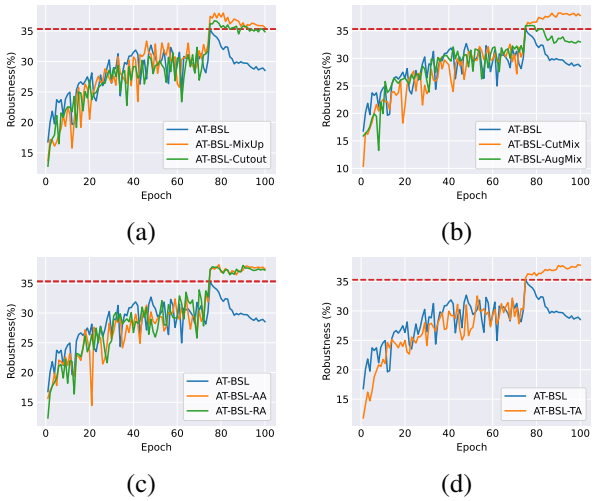


Figure 3. The evolution of test robustness against PGD-20 using ResNet-18 on CIFAR-10-LT for AT-BSL with various data augmentation. To facilitate comparison, the robustness of the most effective checkpoint of AT-BSL is marked by red dashed lines within each graph. Due to the density of the illustrations, the results are divided into four separate panels: (a), (b), (c), and (d).

attempted to mitigate this using data augmentation on balanced datasets, but whether data augmentation remains effective on long-tailed distributions is currently unknown.

Testing MixUp. Some prior works [34, 36, 44] have suggested that on CIFAR-10, MixUp [51] can help alleviate robust overfitting. Thus, we hypothesize that MixUp could also alleviate robust overfitting on CIFAR-10-LT, a long-tailed version of CIFAR-10. According to Fig. 3(a), it is clear that AT-BSL-MixUp significantly reduces robust overfitting compared to AT-BSL. Moreover, we unexpectedly discovered that MixUp also markedly improves robustness. This result is inconsistent with prior observations on balanced datasets [34, 36, 44], which concluded that data augmentation alone does not improve robustness.

Exploring data augmentation. Inspired by the validation of the MixUp hypothesis, our investigation extends to evaluate whether other data augmentation techniques can similarly alleviate robust overfitting and improve robustness. This examination includes methods such as Cutout [9], CutMix [49], AugMix [17], TA [32], AuA [6], and RA [7]. The robustness achieved with these augmentation techniques during training is presented in Fig. 3. Our results reveal that each augmentation technique reduces robust overfitting, with CutMix, AuA, RA, and TA showing particularly strong performance in almost entirely preventing it. Moreover, we note that the robustness achieved with each augmentation surpasses that of the vanilla AT-BSL, thereby supporting the assertion that data augmentation alone can indeed improve robustness.

4. Why Data Augmentation Can Improve Robustness

Formulating Hypothesis. Our hypothesis posits that data augmentation improves robustness by increasing the diversity of the training data, thus enabling models to learn richer representations. We take RA [7] as a case study, where for each training image, RA randomly applies a series of augmentations from a pool of 14 options-Identity, ShearX, ShearY, TranslateX, TranslateY, Rotate, Brightness, Color, Contrast, Sharpness, Posterize, Solarize, AutoContrast, and Equalize. We conduct an ablation study to evaluate the individual impact of each augmentation. Specifically, we limit the pool to just one type, forcing RA to use the same single augmentation for all training examples. According to the results in Fig. 4(a), no single augmentation independently improves robustness except for the Contrast. In fact, augmentations such as Solarize, AutoContrast, and Equalize significantly underperform compared to AT-BSL. This suggests that the diversity provided by a single augmentation is insufficient for improving robustness.

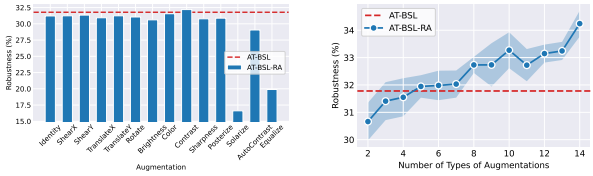
Validating Hypothesis. Subsequently, we examine how increasing the variety of augmentation types impacts robustness. In each experiment, we randomly select n types of augmentations to constitute the pool of RA, with $n \in \{2, 14\}$. Each configuration is tested five times. As shown in Fig. 4(b), robustness consistently improves when more types of augmentations are added to the pool. This trend shows that a richer assortment of augmentations can increase example diversity. Consequently, the model learns more comprehensive representations, which improves robustness and supports our hypothesis.

Moreover, to further reinforce our hypothesis, we conduct an ablation study on three types of augmentations—Solarize, AutoContrast, and Equalize—which, when used individually, impair robustness. We start with a baseline excluding these three augmentations, denoted as RA-11, and incrementally reintroduce them. The results presented in Table 2 show that robustness incrementally improves as more augmentation types are included. Although using these three methods individually may have negative effects, incorporating them into the pool can still improve robustness. This further validates our hypothesis that data augmentation improves robustness by increasing the diversity of training examples.

5. Experiments

5.1. Settings

Datasets. Following [45], we conduct experiments on CIFAR-10-LT and CIFAR-100-LT [22]. Due to space limitations, partial results for CIFAR-100-LT are provided in the appendix. The main experiments focus on CIFAR-10-



(a) (b)

Figure 4. The robustness under AA for AT-BSL with different augmentations using ResNet-18 on CIFAR-10-LT. (a) Change the augmentation space of RA [7] to a single augmentation, and the horizontal axis represents the name of the single augmentation. (b) The horizontal axis represents the number of types of augmentations in the search space of RA.

Table 2. The clean accuracy and robustness against AA for AT-BSL utilizing various augmentations with ResNet-18 on CIFAR-10-LT. The best results are highlighted in **bolded**. "RA-11" refers to utilizing only the initial 11 augmentations in the pool. The lines below RA-11 represent additional augmentations added to RA-11, with the last line employing the full pool. SO denotes Solarize; AC refers to AutoContrast; EQ means Equalize.

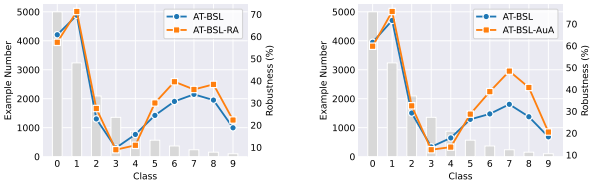
Method	Clean	FGSM	PGD	CW	LSA	AA
RA-11	67.80	40.68	35.88	34.01	33.89	32.12
SO	67.60	41.43	37.04	34.52	34.05	32.76
AC	68.57	41.20	36.60	34.24	34.07	32.51
EQ	68.33	41.64	36.80	34.33	34.17	32.59
SO+AC	68.43	42.10	37.23	34.62	34.37	33.02
SO+EQ	68.53	41.89	37.42	35.07	34.83	33.49
AC+EQ	68.36	41.88	37.42	34.91	34.49	33.15
SO+AC+EQ	**70.86**	**43.06**	**37.94**	**36.24**	**36.04**	**34.24**

LT with an imbalance ratio (IR) set to 50. Table 6 further provides results across various IRs.

Evaluation Metrics. Model robustness is evaluated under an l_∞ norm-bounded perturbation of $\epsilon = 8/255$. Employed attacks include the single-step FGSM [12] and several iterative attacks, such as PGD [30], CW [3] and LSA [18], executed over 20 steps with a step size of $2/255$. We also employ AutoAttack (AA) [5], regarded as the strongest attack to date. For all methods, the evaluations consider both the best checkpoint (chosen based on robustness under PGD-20) and the final checkpoint.

Comparison Methods. We consider two adversarial training methods under long-tailed distributions, including RoBal [45] and REAT [25], as well as defenses designed for balanced distributions, such as AT [30], TRADES [52], MART [41], AWP [44], GAIRAT [53], and LAS-AT [19].

Training Details. We train the models using the Stochastic Gradient Descent (SGD) optimizer with an initial learning rate of 0.1, momentum of 0.9, and weight decay of 5e-4. The batch size is set to 128. The training lasts for 100 epochs, with the learning rate reduced by 10 at the 75th and 90th epochs following [52]. Adversarial examples are generated with a maximum perturbation of $8/255$ and a step size of $2/255$, utilizing 10 iterations for the internal maximization, denoted PGD-10. The impact of PGD steps on robustness is detailed in Table 15. For all experiments about AT-BSL, we adopt $\tau_b = 1$, with results for different τ_b values shown in Fig. 7. Note that the AT-BSL version in Tables 3 and 4 represents our implementation and may differ in training parameters from that of RoBal [45]. Detailed discussions about these differences are provided in the appendix.



(a) (b)

Figure 5. The class-wise example number and robustness against AA for various algorithms on CIFAR-10-LT at the best checkpoint. (a) ResNet-18; (b) WideResNet-34-10.

5.2. Main Results

The results in Tables 3 and 4 indicate that on CIFAR-10-LT, AT-BSL with data augmentation obtains the highest clean accuracy and adversarial robustness across both ResNet-18 and WideResNet-34-10 models. Specifically, on the WideResNet-34-10 model, our AT-BSL-AuA method marks a significant improvement of +6.66% in robustness against AA over RoBal. Moreover, in terms of robustness at the final checkpoint, our method significantly outperforms others, showcasing that data augmentation effectively mitigates robust overfitting.

We detail the class-wise robustness of various methods in Fig. 5. Notably, apart from a few exceptions, our method improves robustness across nearly all classes, particularly in the tail classes (classes 5 to 9). This improvement illustrates the efficacy of our method in addressing long-tailed distribution challenges. Moreover, in line with observations on balanced datasets [29, 43, 46, 48], there is a significant variance in class-wise robustness. Interestingly, Class 3 exhibits the lowest robust level despite having more training examples than subsequent classes. This phenomenon suggests that the intrinsic characteristics of Class 3 may play a significant role in its vulnerability, as highlighted in previous research [45].

5.3. Futher Analysis

Effect of Augmentation Strategies and Parameters. We present the impact of various augmentation strategies and

Table 3. The clean accuracy and robustness for various algorithms using ResNet-18 on CIFAR-10-LT. The best results are **bolded**.

Method	Best Checkpoint						Last Checkpoint					
	Clean	FGSM	PGD	CW	LSA	AA	Clean	FGSM	PGD	CW	LSA	AA
AT [30]	49.35	30.09	27.30	26.93	27.08	25.76	52.91	29.29	25.15	25.58	27.13	24.23
TRADES [52]	43.61	29.18	27.81	26.73	26.58	26.41	43.75	29.06	27.05	26.10	25.93	25.78
MART [41]	48.61	32.75	30.29	28.82	28.46	27.73	48.80	32.60	29.78	28.45	28.12	27.30
AWP [44]	49.29	33.78	31.20	30.53	30.36	29.53	47.75	32.77	30.83	30.01	29.68	29.12
GAIRAT [53]	50.83	30.20	27.46	21.65	21.23	20.41	50.66	28.44	25.60	19.68	19.22	18.26
LAS-AT[19]	52.81	33.35	30.32	29.57	29.15	28.53	53.50	33.14	30.09	29.13	28.84	28.30
RoBal [45]	70.34	40.50	35.93	31.05	31.10	29.54	70.00	36.18	29.00	27.67	26.98	25.63
REAT [25]	67.38	40.13	35.83	33.88	33.66	32.20	67.58	36.99	30.93	30.83	31.62	28.61
AT-BSL	68.89	40.08	35.27	33.47	33.46	31.78	67.63	35.20	28.65	28.91	31.35	26.97
AT-BSL-RA	**70.86**	**43.06**	**37.94**	**36.24**	**36.04**	**34.24**	**71.83**	**42.62**	**37.15**	**35.37**	**35.50**	**33.44**

Table 4. The clean accuracy and robustness of various algorithms using WideResNet-34-10 on CIFAR-10-LT. The best results are highlighted in **bolded**.

Method	Best Checkpoint						Last Checkpoint					
	Clean	FGSM	PGD	CW	LSA	AA	Clean	FGSM	PGD	CW	LSA	AA
AT [30]	59.21	31.88	27.88	28.19	29.81	27.07	58.25	29.77	25.29	25.71	29.83	24.94
TRADES [52]	51.28	31.58	28.70	28.45	28.36	27.72	53.85	30.44	26.23	26.57	26.77	25.59
MART [41]	49.13	34.33	32.32	30.73	30.13	29.60	52.48	33.95	31.09	29.64	29.43	28.67
AWP [44]	50.91	34.28	31.85	31.23	31.01	30.06	48.65	33.21	31.07	30.33	30.14	29.40
GAIRAT [53]	59.89	33.47	30.40	26.69	26.71	25.38	56.37	29.41	27.25	23.94	23.95	23.15
LAS-AT [19]	57.52	33.66	29.86	29.60	29.44	28.84	58.19	32.98	28.89	28.75	28.58	27.90
RoBal [45]	72.82	41.34	36.42	32.48	31.95	30.49	70.85	35.95	27.74	27.59	26.76	25.71
REAT [25]	73.16	41.32	35.94	35.28	35.67	33.20	67.76	34.51	27.75	28.17	31.82	26.66
AT-BSL	73.19	41.84	35.60	34.86	35.99	32.80	65.95	33.29	27.23	27.87	31.00	26.45
AT-BSL-AuA	**75.17**	**46.18**	**40.84**	**38.82**	**39.23**	**37.15**	**77.27**	**44.73**	**38.06**	**37.14**	**39.05**	**35.11**

their parameters on robustness through Table 5 and Fig. 6. These experiments focus on the robustness at the best checkpoint. Specifically, Table 5 employs the optimal hyper-parameters for each strategy: a mixing rate of $\alpha = 0.3$ for Mixup, a window length of 17 for Cutout, a mixing rate of $\alpha = 0.1$ for CutMix, and a magnitude of 8 for RA. The results illustrate that different strategies can improve robustness beyond the vanilla AT-BSL. Notably, AuA and RA not only improve robustness but also contribute to higher clean accuracy. Fig. 6 indicates that for MixUp and CutMix, lower α values lead to better robustness; for Cutout, longer window lengths generally correlate with better robustness; and for RA, an optimal level of transformation improves robustness, reaching its peak at magnitude $= 8$, suggesting that overly aggressive augmentation may not yield further benefits.

Effect of Hyperparameter τ_b. To investigate the sensitivity of AT-BSL to τ_b, we evaluate the performance of the trained models under varying τ_b values. Specifically, we utilize ResNet-18 with τ_b ranging from 0 to 20. When $\tau_b = 0$, the bias $b_i = \tau_b \log(n_i)$ added by AT-BSL becomes zero, reverting BSL to the cross-entropy loss and transforming AT-BSL into vanilla AT [30]. The results, depicted in Fig.

Table 5. The clean accuracy and robustness for AT-BSL with different augmentations using ResNet-18 on CIFAR-10-LT. The best results are **bolded**.

Method	Clean	FGSM	PGD	CW	LSA	AA
Vanilla	68.89	40.08	35.27	33.47	33.46	31.78
MixUp [51]	65.82	41.33	**38.05**	34.29	33.63	32.92
Cutout [9]	65.12	40.25	36.68	34.81	34.51	33.35
CutMix [49]	64.54	41.13	37.86	34.10	33.46	32.83
AugMix [17]	67.12	40.31	35.95	34.19	34.02	32.51
TA [32]	67.14	41.56	37.75	34.34	33.90	32.62
AuA [6]	**71.63**	42.69	37.78	35.60	35.47	33.69
RA [7]	70.86	**43.06**	37.94	**36.24**	**36.04**	**34.24**

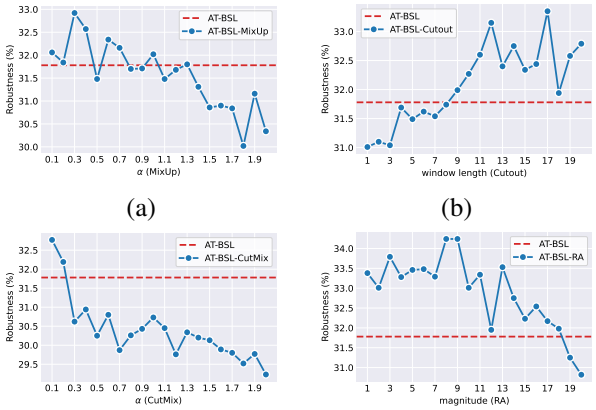


Figure 6. The robustness against AA employing ResNet-18 on CIFAR-10-LT as we variables include (a) the mixing rate α for MixUp, (b) the window length for Cutout, (c) the mixing rate α for CutMix, and (d) the magnitude of transformations for RA.

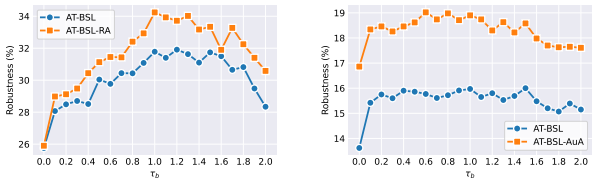


Figure 7. The robustness under AA for various algorithms with different τ_b using ResNet-18. (a): CIFAR-10-LT; (b): CIFAR-100-LT.

7, reveal that on CIFAR-10-LT, AT-BSL is quite sensitive to variations in τ_b, with the model performing optimally at $\tau_b = 1$. In addition, on CIFAR-100-LT, the performance of AT-BSL exhibits a lesser degree of sensitivity to changes in τ_b. Moreover, including data augmentation strategies across all datasets and τ_b values consistently results in a significant robustness improvement compared to the vanilla AT-BSL. This emphasizes the benefits of data augmentation in adversarial training, especially within the context of long-tailed distributions.

Effect of Imbalance Ratio. To assess how our method performs with varying IRs, we create long-tailed versions of datasets following [8, 45]. Our results, presented in Table 6, demonstrate that RA consistently improves the robustness of AT-BSL in various IR settings. This further supports the conclusion that data augmentation can improve robustness.

Effect of PGD Step Size. To investigate the impact of the PGD step size on robustness, we fine-tune the step size from the standard $2/255$ to smaller values of $1/255$ and $0.5/255$, simultaneously increasing the number of PGD steps from 10 to 20 and then to 40. As shown in Table 7, it is clear that RA consistently improves the robustness of AT-BSL across all tested PGD step sizes. However, we also note a decrease in robustness relative to the original baseline performance achieved at a PGD step size of $2/255$.

Table 6. The clean accuracy and robustness for various algorithms using ResNet-18 on CIFAR-10-LT with different imbalance ratios. Better results are **bolded**.

IR	Method	Clean	FGSM	PGD	CW	LSA	AA
10	AT-BSL	73.29	47.33	42.04	40.77	41.05	39.12
	AT-BSL-RA	**79.00**	**50.98**	**44.19**	**42.82**	**43.10**	**40.56**
20	AT-BSL	71.89	44.76	39.40	38.47	38.68	36.74
	AT-BSL-RA	**75.84**	**47.62**	**41.68**	**39.92**	**39.82**	**37.78**
50	AT-BSL	68.89	40.08	35.27	33.47	33.46	31.78
	AT-BSL-RA	**70.86**	**43.06**	**37.94**	**36.24**	**36.04**	**34.24**
100	AT-BSL	62.03	35.06	30.95	29.41	29.56	28.01
	AT-BSL-RA	**66.85**	**38.75**	**33.69**	**31.77**	**31.50**	**30.00**

Table 7. The clean accuracy and robustness of various algorithms employing ResNet-18 on CIFAR-10-LT, adjusted for various imbalance ratios (IRs).

Size	Method	Clean	FGSM	PGD	CW	LSA	AA
0.5	AT-BSL	68.57	39.65	35.10	32.92	32.97	31.28
	AT-BSL-RA	**68.68**	**41.97**	**37.60**	**34.81**	**34.36**	**33.26**
1	AT-BSL	68.63	39.98	35.09	33.02	33.00	31.18
	AT-BSL-RA	**68.93**	**42.71**	**37.85**	**35.30**	**34.79**	**33.51**
2	AT-BSL	68.89	40.08	35.27	33.47	33.46	31.78
	AT-BSL-RA	**70.86**	**43.06**	**37.94**	**36.24**	**36.04**	**34.24**

6. Conclusion

In this paper, we first investigate the design of RoBal and identify Balanced Softmax Loss as the critical component. We then propose the issue of robust overfitting in adversarial training under long-tailed distributions and attempt to mitigate this using data augmentation. We discover that data augmentation not only mitigates robust overfitting but also improves robustness, and we validate that the improved robustness is due to the expanded training example diversity brought by data augmentation. Finally, we conduct extensive experiments with various data augmentation strategies, model architectures, and datasets, affirming the generalizability of our findings. Through our research, we contribute to the advancement of adversarial training, making it more adaptable and effective in real-world data scenarios.

Acknowledgements

This work was partially supported by the NSFC under Grants U20B2049, U21B2018, and 62302344.

References

[1] Sravanti Addepalli, Samyak Jain, et al. Efficient and effective augmentation strategy for adversarial training. In *NeurIPS*, 2022. 4

[2] Sumyeong Ahn, Jongwoo Ko, and Se-Young Yun. CUDA: Curriculum of data augmentation for long-tailed recognition. In *ICLR*, 2023. 2, 3

[3] Nicholas Carlini and David Wagner. Towards evaluating the robustness of neural networks. In *S&P*, 2017. 3, 6, 1

[4] Yair Carmon, Aditi Raghunathan, Ludwig Schmidt, John C Duchi, and Percy S Liang. Unlabeled data improves adversarial robustness. In *NeurIPS*, 2019. 2, 4

[5] Francesco Croce and Matthias Hein. Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks. In *ICML*, 2020. 1, 3, 6

[6] Ekin D Cubuk, Barret Zoph, Dandelion Mane, Vijay Vasudevan, and Quoc V Le. Autoaugment: Learning augmentation strategies from data. In *CVPR*, 2019. 2, 5, 7, 1

[7] Ekin Dogus Cubuk, Barret Zoph, Jon Shlens, and Quoc Le. Randaugment: Practical automated data augmentation with a reduced search space. In *NeurIPS*, 2020. 2, 5, 6, 7, 1

[8] Yin Cui, Menglin Jia, Tsung-Yi Lin, Yang Song, and Serge Belongie. Class-balanced loss based on effective number of samples. In *CVPR*, 2019. 2, 8

[9] Terrance DeVries and Graham W Taylor. Improved regularization of convolutional neural networks with cutout. *arXiv preprint arXiv:1708.04552*, 2017. 2, 5, 7, 1

[10] Fei Du, Peng Yang, Qi Jia, Fengtao Nan, Xiaoting Chen, and Yun Yang. Global and local mixture consistency cumulative learning for long-tailed visual recognitions. In *CVPR*, 2023. 2

[11] Wang et al. Better diffusion models further improve adversarial training. In *ICML*, 2023. 5

[12] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In *ICLR*, 2015. 1, 6

[13] Sven Gowal, Chongli Qin, Jonathan Uesato, Timothy Mann, and Pushmeet Kohli. Uncovering the limits of adversarial training against norm-bounded adversarial examples. *arXiv preprint arXiv:2010.03593*, 2020. 2, 4

[14] Agrim Gupta, Piotr Dollar, and Ross Girshick. Lvis: A dataset for large vocabulary instance segmentation. In *CVPR*, 2019. 1, 2

[15] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 2, 4, 3

[16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Identity mappings in deep residual networks. In *ECCV*, 2016. 2

[17] Dan Hendrycks, Norman Mu, Ekin Dogus Cubuk, Barret Zoph, Justin Gilmer, and Balaji Lakshminarayanan. Augmix: A simple data processing method to improve robustness and uncertainty. In *ICLR*, 2020. 2, 5, 7, 1

[18] Dorjan Hitaj, Giulio Pagnotta, Iacopo Masi, and Luigi V Mancini. Evaluating the robustness of geometry-aware instance-reweighted adversarial training. *arXiv preprint arXiv:2103.01914*, 2021. 3, 6

[19] Xiaojun Jia, Yong Zhang, Baoyuan Wu, Ke Ma, Jue Wang, and Xiaochun Cao. Las-at: adversarial training with learnable attack strategy. In *CVPR*, 2022. 1, 2, 6, 7, 4

[20] Bingyi Kang, Saining Xie, Marcus Rohrbach, Zhicheng Yan, Albert Gordo, Jiashi Feng, and Yannis Kalantidis. Decoupling representation and classifier for long-tailed recognition. In *ICLR*, 2020. 2

[21] Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the design space of diffusion-based generative models. In *NeurIPS*, 2022. 5

[22] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. *Technical report*, 2009. 1, 5

[23] Ya Le and Xuan Yang. Tiny imagenet visual recognition challenge. *CS 231N*, 2015. 1, 2

[24] Saehyung Lee, Hyungyu Lee, and Sungroh Yoon. Adversarial vertex mixup: Toward better adversarially robust generalization. In *CVPR*, 2020. 2

[25] Guanlin Li, Guowen Xu, and Tianwei Zhang. Adversarial training over long-tailed distribution. *arXiv preprint arXiv:2307.10205*, 2023. 1, 4, 6, 7, 2, 3

[26] Jian Li, Ziyao Meng, Daqian Shi, Rui Song, Xiaolei Diao, Jingwen Wang, and Hao Xu. Fcc: Feature clusters compression for long-tailed visual recognition. In *CVPR*, 2023. 2

[27] Lin Li and Michael W. Spratling. Data augmentation alone can improve adversarial training. In *ICLR*, 2023. 4

[28] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *ICCV*, 2017. 2

[29] Xinsong Ma, Zekai Wang, and Weiwei Liu. On the tradeoff between robustness and fairness. In *NeurIPS*, 2022. 6

[30] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *ICLR*, 2018. 1, 2, 3, 4, 6, 7

[31] Aditya Krishna Menon, Sadeep Jayasumana, Ankit Singh Rawat, Himanshu Jain, Andreas Veit, and Sanjiv Kumar. Long-tail learning via logit adjustment. In *ICLR*, 2021. 3

[32] Samuel G Müller and Frank Hutter. Trivialaugment: Tuning-free yet state-of-the-art data augmentation. In *CVPR*, 2021. 2, 3, 5, 7, 1

[33] Tianyu Pang, Xiao Yang, Yinpeng Dong, Kun Xu, Jun Zhu, and Hang Su. Boosting adversarial training with hypersphere embedding. In *NeurIPS*, 2020. 3

[34] Sylvestre-Alvise Rebuffi, Sven Gowal, Dan Andrei Calian, Florian Stimberg, Olivia Wiles, and Timothy A Mann. Data augmentation can improve robustness. In *NeurIPS*, 2021. 2, 4, 5, 1

[35] Jiawei Ren, Cunjun Yu, Xiao Ma, Haiyu Zhao, Shuai Yi, et al. Balanced meta-softmax for long-tailed visual recognition. In *NeurIPS*, 2020. 1, 2, 3, 4

[36] Leslie Rice, Eric Wong, and Zico Kolter. Overfitting in adversarially robust deep learning. In *ICML*, 2020. 2, 4, 5

[37] Li Shen, Zhouchen Lin, and Qingming Huang. Relay backpropagation for effective learning of deep convolutional neural networks. In *ECCV*, 2016. 2

[38] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. In *ICLR*, 2014. 1

[39] Grant Van Horn, Oisin Mac Aodha, Yang Song, Yin Cui, Chen Sun, Alex Shepard, Hartwig Adam, Pietro Perona, and Serge Belongie. The inaturalist species classification and detection dataset. In *CVPR*, 2018. 1, 2

[40] Tao Wang, Yu Li, Bingyi Kang, Junnan Li, Junhao Liew, Sheng Tang, Steven Hoi, and Jiashi Feng. The devil is in classification: A simple framework for long-tail instance segmentation. In *ECCV*, 2020. 2

[41] Yisen Wang, Difan Zou, Jinfeng Yi, James Bailey, Xingjun Ma, and Quanquan Gu. Improving adversarial robustness requires revisiting misclassified examples. In *ICLR*, 2020. 1, 2, 6, 7, 4

[42] Yu-Xiong Wang, Deva Ramanan, and Martial Hebert. Learning to model the tail. In *NeurIPS*, 2017. 2

[43] Zeming Wei, Yifei Wang, Yiwen Guo, and Yisen Wang. Cfa: Class-wise calibrated fair adversarial training. In *CVPR*, 2023. 6

[44] Dongxian Wu, Shu-Tao Xia, and Yisen Wang. Adversarial weight perturbation helps robust generalization. In *NeurIPS*, 2020. 1, 2, 4, 5, 6, 7

[45] Tong Wu, Ziwei Liu, Qingqiu Huang, Yu Wang, and Dahua Lin. Adversarial robustness under long-tailed distribution. In *CVPR*, 2021. 1, 2, 3, 4, 5, 6, 7, 8

[46] Han Xu, Xiaorui Liu, Yaxin Li, Anil Jain, and Jiliang Tang. To be robust or to be fair: Towards fairness in adversarial training. In *ICML*, 2021. 6

[47] Zhengzhuo Xu, Zenghao Chai, and Chun Yuan. Towards calibrated model for long-tailed visual recognition from prior perspective. In *NeurIPS*, 2021. 2

[48] Xinli Yue, Ningping Mou, Qian Wang, and Lingchen Zhao. Revisiting adversarial robustness distillation from the perspective of robust fairness. In *NeurIPS*, 2023. 6

[49] Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoon Yoo. Cutmix: Regularization strategy to train strong classifiers with localizable features. In *ICCV*, 2019. 2, 5, 7, 1

[50] Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. In *BMVC*, 2016. 1, 2, 3

[51] Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. In *ICLR*, 2018. 2, 5, 7, 1

[52] Hongyang Zhang, Yaodong Yu, Jiantao Jiao, Eric Xing, Laurent El Ghaoui, and Michael Jordan. Theoretically principled trade-off between robustness and accuracy. In *ICML*, 2019. 1, 2, 3, 4, 6, 7

[53] Jingfeng Zhang, Jianing Zhu, Gang Niu, Bo Han, Masashi Sugiyama, and Mohan Kankanhalli. Geometry-aware instance-reweighted adversarial training. In *ICLR*, 2021. 1, 2, 6, 7, 4

[54] Yongshun Zhang, Xiu-Shen Wei, Boyan Zhou, and Jianxin Wu. Bag of tricks for long-tailed visual recognition with deep convolutional neural networks. In *AAAI*, 2021. 2

[55] Zizhao Zhang and Tomas Pfister. Learning fast sample reweighting without reward data. In *ICCV*, 2021. 2

[56] Zhipeng Zhou, Lanqing Li, Peilin Zhao, Pheng-Ann Heng, and Wei Gong. Class-conditional sharpness-aware minimization for deep long-tailed recognition. In *CVPR*, 2023. 2