Adversarial Robustness under Long-Tailed Distribution

Tong Wu[1,5], Ziwei Liu[2], Qingqiu Huang[3], Yu Wang[4], Dahua Lin[1,5,6]
[1]The Chinese University of Hong Kong, [2]S-Lab, Nanyang Technological University, [3]Huawei,
[4]Tsinghua University, [5]SenseTime-CUHK Joint Lab, [6]Centre of Perceptual and Interactive Intelligence
{wt020,dhlin,hq016}@ie.cuhk.edu.hk, ziwei.liu@ntu.edu.sg, yu-wang@mail.tsinghua.edu.cn

Abstract

Adversarial robustness has attracted extensive studies recently by revealing the vulnerability and intrinsic characteristics of deep networks. However, existing works on adversarial robustness mainly focus on balanced datasets, while real-world data usually exhibits a long-tailed distribution. To push adversarial robustness towards more realistic scenarios, in this work we investigate the adversarial vulnerability as well as defense under long-tailed distributions. In particular, we first reveal the negative impacts induced by imbalanced data on both recognition performance and adversarial robustness, uncovering the intrinsic challenges of this problem. We then perform a systematic study on existing long-tailed recognition methods in conjunction with the adversarial training framework. Several valuable observations are obtained: 1) natural accuracy is relatively easy to improve, 2) fake gain of robust accuracy exists under unreliable evaluation, and 3) boundary error limits the promotion of robustness. Inspired by these observations, we propose a clean yet effective framework, RoBal, which consists of two dedicated modules, a scale-invariant classifier and data re-balancing via both margin engineering at training stage and boundary adjustment during inference. Extensive experiments demonstrate the superiority of our approach over other state-of-the-art defense methods. To our best knowledge, we are the first to tackle adversarial robustness under long-tailed distributions, which we believe would be a significant step towards real-world robustness. Our code is available at: https://github.com/wutong16/Adversarial_Long-Tail.

1. Introduction

Despite the great progress on a variety of computer vision tasks, deep neural networks are found to be vulnerable to minor adversarial perturbations [38], *i.e.*, easily misled to make incorrect predictions. The existence of adversarial examples reveals a non-negligible security risk to modern computer vision models, with extensive efforts devoted to improving adversarial robustness.

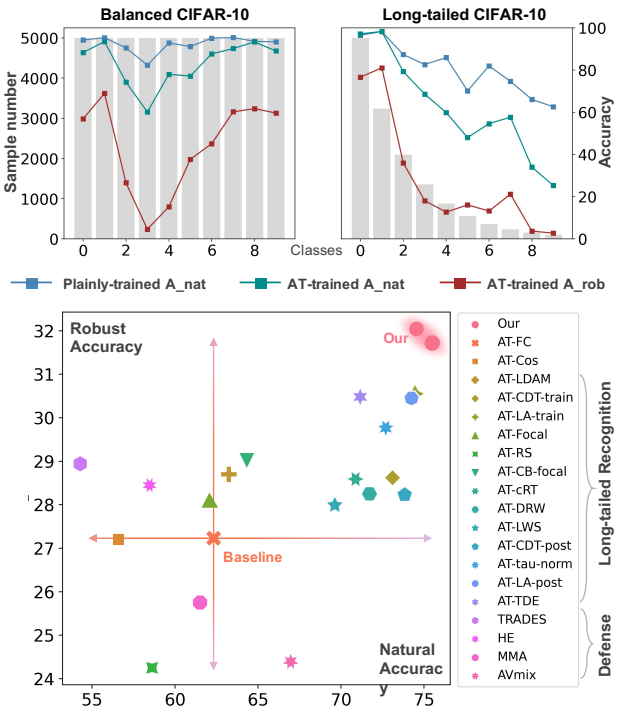

Figure 1. **Upper**: A long-tailed data distribution induces decreasing natural and robust accuracy from head to tail and a magnified "sacrifice" of natural accuracy especially to tail classes when adversarial training is applied. **Lower**: Evaluation results on two metric dimensions, including a number of long-tailed recognition methods combined with adversarial training, several state-of-the-art defense methods, and our RoBal in a region with trade-off.

Existing adversarial robustness research mainly focuses on balanced datasets such as MNIST, CIFAR, and ImageNet [12]. Nevertheless, real-world data usually exhibit a long-tailed distribution [40, 11], which brings challenges not only to the recognition tasks themselves but also to robustness against adversarial attacks. The former has been attracting increasing attention recently, with a number of algorithms [24, 16, 51, 7, 2, 44] proposed to tackle the issue; On the other hand, the latter remains largely unexplored.

To cast light on the challenges of adversarial robustness in **long-tailed recognition (LT)**, we first perform an intuitive comparison between networks trained on the balanced and long-tailed versions of CIFAR-10, respectively. Apart from normally trained models, we also adopt the **adversarial training (AT)** framework [25], which is one of the most effective and widely used defense methods, to provide the basic adversarial robustness for the networks. Per-class classification recalls are evaluated on clean images and images permuted by PGD attack [25], denoted by *natural accuracy* $\mathbf{A}_{nat}$ and *robust accuracy* $\mathbf{A}_{rob}$, respectively. $\mathbf{A}_{nat}$ is evaluated on both **plain** models and **AT-trained** models, while $\mathbf{A}_{rob}$ is performed only on the latter. Results are visualized in Fig. 1. There are three main observations from the comparison: 1) $\mathbf{A}_{nat}$ on plain models drops from head to tail, which is exactly what traditional long-tailed recognition aims to solve. 2) A similar decreasing tendency reasonably occurs in $\mathbf{A}_{rob}$. 3) It is worth noting that $\mathbf{A}_{nat}$ drops more significantly at the tail when adversarial training is applied, indicating that the well-known "sacrifice" of the natural accuracy induced by adversarial training is further magnified for tail classes under a long-tailed distribution.

To form a better understanding of the problem, the relationship between natural and robust accuracy can be connected by *boundary error* $\mathbf{R}_{bdy}$ [50] as:

$$\mathbf{A}_{rob} = \mathbf{A}_{nat} - \mathbf{R}_{bdy}, \tag{1}$$

where $\mathbf{R}_{bdy}$ represents how likely the features of clean and correctly predicted inputs are close to the ϵ-extension of the decision boundary. It represents the gap between the two forms of accuracy and indicates the vulnerability of samples against adversarial attacks.

Hence, to achieve improvement on both recognition performance and adversarial robustness, a natural idea is to raise $\mathbf{A}_{nat}$ while keeping a small value of $\mathbf{R}_{bdy}$. Specifically, on the one hand, we are able to address the issue of imbalance in data distribution via re-balancing strategies, thus we conduct a systematic study of currently widely used long-tailed recognition approaches to explore the proper combinations of these methods and the adversarial training framework. On the other hand, we would analyze why a normalized embedding space promotes model resistance against attacks, and then a scale-invariant classifier is introduced to replace the final linear layer. The idea of data re-balancing is then well aligned with the cosine classifier by the cooperation of class-aware and pair-aware margins during training and boundary adjustment at inference.

Note that the imbalance in data distribution and the deficiency in sample numbers are two issues induced simultaneously when we turn to long-tailed datasets instead of the artificially balanced ones. Although the importance of data scale in adversarial robustness has been widely studied [33], we mainly focus on the problem of imbalance in this paper. We study the effect of them separately in Sec 5, verifying that eliminating prediction priors is crucial to reducing the vulnerability of tail classes under attack.

Our contributions are as follows: **1)** To our best knowledge, we are the first to tackle adversarial robustness under long-tailed distribution, which we believe would be a significant step towards real-world robustness. **2)** We conduct a systematic study on existing long-tailed recognition methods and their adoption into the adversarial training procedure. Important insights are gained based on experimental observations. **3)** We further develop a clean yet effective approach, RoBal, that achieves state-of-the-art performance on both natural and robust accuracy.

2. Related Works

Long-Tailed Recognition. To tackle the long-tailed recognition problem, traditional re-balancing approaches include re-sampling [5, 15, 12, 36, 1] and re-weighting [7, 1]. However, these methods may suffer from the issue of under-representing major classes and over-fitting minor ones. To mitigate these negative impacts, more flexible usages of the basic methods were proposed, such as decoupled training [16, 51] and deferred re-balancing schedule [2], respectively, and they are proved to be more effective. Further, recently proposed approaches address class-specific properties by perspectives like margin [2], bias [30, 27], temperature [47] or weight scale [16, 19], and some of these methods can be either adopted to the whole training process or in a post-processing manner. Another trend of works focuses on sample-specific properties via hard example mining [22] or sample-aware re-weighting strategies leveraging meta-learning [31, 14, 37]. Besides, several recent approaches propose to transfer knowledge from head to tail through memory module [24], inter-class feature transferring [23], and "major-to-minor" translation [20]. In this paper, we would revisit and summarize a number of these methods and explore their effective combination with adversarial training in Sec. 3.

Adversarial Robustness. Plenty of adversarial defense methods have been proposed to tackle the problem of adversarial vulnerability. Among them, adversarial training [25] is one of the most effective and reliable strategies. Improvements have been made based on the AT framework via theoretical analysis and loss function examination [50, 42, 8]. Many efforts have also been devoted to exploring different training mechanisms such as metric learning [26], self-supervised learning [13], and semi-supervised learning [4]. Since AT is of the high computational cost and time consumption, another line of works [35, 43, 49] was proposed to accelerate the training procedure. Besides, some general strategies were revealed to be critical to the robustness performance such as label smoothing [34], early stop [32], different activation functions [46], batch normalization [45], and embedding space [29]. Most recently, Pang *et al*. [28]

Table 1. A systematic study of current LT strategies combined with AT framework, detailed explanations are to be included in Sec.A.3. Green, red, and blue denote impressive A_{nat}, unreliable evaluation of A_{rob} under PGD attack, and the smallest R_{bdy}, respectively.

Stage	Methods	Formulation	Clean	PGD	AA	Gap
Train	Vanilla FC	$g_i = W_i^T f(x)$	62.33	29.30	28.15	34.18
	Vanilla Cos	$g_i = \widetilde{W}_i^T \widetilde{f}(x)$	56.59	29.38	27.23	29.36
	Class-aware margin [2]	$g_i = W_i^T f(x) - \mathbb{1}\{i = y\} \cdot \delta_i$	63.24	29.81	28.70	34.54
	Cosine with margin [41]	$g_i = \widetilde{W}_i^T \widetilde{f}(x) - \mathbb{1}\{i = y\} \cdot m$	58.47	31.73	28.45	30.02
	Class-aware temperature [48]	$g_i = W_i^T f(x) \cdot (n_i/n_{max})^\gamma$	73.11	30.12	28.62	44.49
	Class-aware bias [27, 30]	$g_i = W_i^T f(x) + \tau \log(n_i)$	74.46	32.45	30.55	43.91
	Hard-exmaple mining [22]	$r(y) = (1 - p_y)^\gamma$, applyed with BCE loss	62.07	30.73	28.12	33.95
	Re-sampling [36]	$r_s(i) \propto 1/n_i$	58.62	25.06	24.25	34.37
	Re-weighting [7]	$r(y) = (1 - \beta)/(1 - \beta_y^n)$	64.33	34.53	29.01	35.32
Fine-tune	One-epoch re-sampling [16]	$h_i = {W'}_i^T f(x)$, ${W'}_i$ re-trained with RS	70.88	29.81	28.59	42.29
	One-epoch re-weighting [2, 7]	$h_i = {W'}_i^T f(x)$, ${W'}_i$ fine-tuned with RW	71.72	32.34	28.25	43.47
	Learnable classifier scale [16]	$h_i = s_i \cdot W_i^T f(x)$, where s_i is learnable	69.63	28.81	27.99	41.64
Inference	Classifier re-scaling [48, 19]	$h_i = (W_i/n_i^\tau)^T f(x)$	73.84	39.05	28.23	45.61
	Classifier normalization [16]	$h_i = (W_i/\left\|W_i\right\|^\tau)^T f(x)$	72.70	36.57	29.77	42.93
	Class-aware bias [27]	$h_i = W_i^T f(x) - \tau \log(n_i)$	74.25	31.95	30.45	43.80
	Feature disentangling [39]	$h_i = W_i^T (f(x) - \alpha \cos(f(x), d) \cdot d)$	71.16	32.69	30.48	40.68

and Gowal *et al*. [10] made systematic studies on the effect of basic training settings and some other choices, including model size, data, loss, and activation functions, respectively. Our method is also built on the AT framework, while we focus on the long-tailed training data distribution to explore how it affects the accuracy and ways for improvement.

3. Long-tailed Recognition with Defense

In this section, we first briefly introduce the adversarial training (AT) framework. Then we conduct a systematic study on some popular long-tailed recognition (LT) strategies and explore their proper combination with the AT framework, where the effectiveness is evaluated by $\mathbf{A}_{nat}$, $\mathbf{A}_{rob}$, and $\mathbf{R}_{bdy}$. We further reveal a fake increase of robustness that could be induced under unreliable evaluation. Finally, valuable knowledge from the study is summarized and inspires us to develop our method in Sec. 4.

3.1. Adversarial Training Preliminaries

Adversarial training, as one of the most effective defense methods, is adopted as the basic framework to maintain basic robustness in this paper. The standard AT and its variants can be formulated as a mini-max problem:

$$\begin{aligned} &\min_{\theta} E_{(x,y)\sim\mathcal{D}} \left[\mathcal{L}_T(\theta; x+\delta, y)\right], \\ &where \ \ \delta = \underset{\delta\in\mathcal{B}(\epsilon)}{\operatorname{argmax}} \mathcal{L}_A(\theta; x+\delta, y). \end{aligned} \tag{2}$$

The inner optimization aims to find effective adversarial examples by maximizing $\mathcal{L}_A$, and the outer optimization updates network parameters to minimize the training loss $\mathcal{L}_T$. The standard AT proposed by Madry *et al*. [25] uses Cross-Entropy loss(CE) for both $\mathcal{L}_A$ and $\mathcal{L}_T$, while we would explore the effects of different choices in this paper.

3.2. Revisiting Long-tailed Recognition Methods

Preliminaries. The LT methods, who could be naturally combined with the AT framework, can be categorized into three phases based on different applying stages: **training**, **fine-tuning**, and **inference**, as summarized in Table 1. The evaluation metrics include the accuracy of the clean images and permuted images under PGD-20 [25] and Auto-Attack (AA) [6]. Details are to be introduced in Sec. 5. We also report the gap between clean accuracy and Auto-Attack accuracy for a better view of boundary error.

Notifications. Suppose there are C classes in total with $n_i, i \in \{1, 2, ..., C\}$ samples for class i. We denote $f(x)$ as the deep feature extracted from image x and $W = [W_1, ..., W_C]$ as the classifier weight vectors. And the normalized weight vectors and features are denoted as $\widetilde{W}_i = W_i/\left\|W_i\right\|$ and $\widetilde{f}(x) = f/\left\|f\right\|$.

Training Stage. Methods applied to training stage include class-aware **re-sampling** [36, 24, 16, 51], and several **cost-sensitive learning** approaches. We denote the sampling frequency for class i as $r_s(i)$ in Table 1. Cost-sensitive learn-

ing methods usually modify the loss function by introducing class-specific parameters like *weight* (CB [7]), *margin* (LDAM [2]), *bias* (LA-train [27], Balanced Softmax [30]), and *temperature* (CDT-train [47]). A *hard example examining* method (Focal [22]) is also included here although it is applied with binary cross entropy loss. A general loss function for the cost-sensitive methods above with CE loss can be formulated as:

$$\begin{aligned} \mathcal{L}'_{CE}(W; f(x), y) &= -r_w(y) \cdot log(\frac{e^{z_y}}{\sum_i e^{z_i}}), \\ where\ \ z_i &= g_i(W_i, f(x)), \end{aligned} \tag{3}$$

where $g(W, f(x))$ denotes the logit before softmax, and $r_w(y)$ is a re-weighting factor for class y. The widely used linear classifier would have $g_i(W_i, f(x)) = W_i^T f(x) + b_i$. Different methods on training would have different g functions, which are listed in Table 1 (b_i is omit for simplicity).

$\mathcal{L}'_{CE}$ can be adopted to AT procedure in three modes: replacing the CE in $\mathcal{L}_A$, $\mathcal{L}_T$, or both of them, where $\mathcal{L}_A$ and $\mathcal{L}_T$ would affect the optimization of the adversarial examples and network parameter updating, respectively. We empirically observed that modifying $\mathcal{L}_T$ has a more conspicuous influence on the results. Thus results reported here are conducted with the second mode. We present a more detailed study in the supplementary material Sec. **??**.

Fine-tuning Stage. Fine-tuning based methods propose to re-train [16] or fine-tune the classifier via data re-balancing techniques with the backbone frozen, which take advantage of the idea of decoupling the learning of representation and classifier. We empirically find out that one-epoch of fine-tuning with class-aware sampling or re-weighting would remarkably raise A_{nat} while more steps make little difference. A similar conclusion is drawn when only weight scales s_i are learned at this stage (LWS [16]).

Inference Stage. LT methods applied at the inference stage based on a vanilla trained model would usually conduct a different forwarding process from the training stage to address shifted data distributions from train-set to test-set. Specifically, we denote $h(W, f(x))$ as the logit producing function on inference, and the prediction is performed by:

$$\underset{i \in [C]}{\operatorname{argmax}}\, h_i(W_i, f(x)), \tag{4}$$

We consider four post processing methods in this paper including classifier normalization (τ-norm [16]), classifier re-scaling based on sample numbers (CDT-post [48, 19]), feature disentangling (TDE [39]), and logit adjustment (LA-post [27]), as shown in Table 1.

3.3. Analysis and Takeaways

Here we summarize some of the key observations and knowledge revealed by the empirical study, including the effectiveness of LT methods on natural accuracy, the challenge to robustness evaluation reliability induced by some LT strategies, and the importance of boundary error for robust accuracy.

Natural accuracy is easy to improve. According to the results in Table 1, a number of LT strategies are proved effective in improving A_{nat}, as marked in green. We can either apply a cost-sensitive loss in outer minimization $\mathcal{L}_T$ during the whole training stage, or leverage fine-tuning and post-processing to boost the performance with little extra computational cost. This indicates that both re-balancing strategies applied during the training process and boundary adjustment at inference time positively impact A_{nat}, which inspires the development of our method in Sec. 4.

Fake improvement exists for robust accuracy evaluation. It is noticeable that methods based on classifier normalization [16] and re-scaling [48, 19] achieve impressive robust accuracy under PGD-attack, as marked in red, while the AA evaluated results remain ordinary. This is due to the sensitivity of PGD attack to both logits stretching and compressing, which is worth attention.

Consider a uniformly re-scaled classifier $W'_i = W_i / 10^{\kappa}$ at inference time, where logit scales and κ are negatively correlated. As shown in Fig. 2, A_{nat} is not effected since the re-scaling operation by $10^{-\kappa} > 0$ does not change the ordering; AA evaluated A_{rob} is also invariant to scaling which can be seen as a reliable evaluation of robustness; while PGD robustness exhibits a minimum at $\kappa \approx 0$ and increases as κ leaves zero on both sides.

The reason lies in the updating of PGD attack, which is based on the gradient produced by the inner CE loss:

$$\begin{aligned} \nabla_{f(x)} \mathcal{L}_{CE}(W; f(x), y) &= -\nabla_{f(x)} z_y + \sum_i p_i \nabla_{f(x)} z_i \\ &= \sum_{i \neq y} p_i (W_i - W_y). \end{aligned} \tag{5}$$

The vectors W_y to W_i with $i \neq y$ are weighted by *softmax* produced prediction confidence, p_i, which is effective by focusing more on easily confused classes. But the sensitivity of softmax to both the absolute and relative values of its components leads to two kinds of fail cases of PGD.

1) Gradient vanishing. The false sense of security due to gradient vanishing is not new knowledge [3, 6]: a correctly predicted clean image with $y = \operatorname{argmax}_i z_i$ would gain $p_y \approx 1$ and $p_i \approx 0 (i \neq y)$ when all logits are scaled up, leading to zero gradient as in Eqn. 5. We calculate the ratio of zero in gradient at pixel level during the PGD updating following [6], and it converges to the same level as A_{nat} rather than 100% in their paper (Fig. 2).

2) Direction averaging. On the other side, compressed logits lead to averaged softmax outputs, where $p_i \approx 1/C$, and the updating direction becomes $\nabla_{f(x)} \mathcal{L}_{CE} = \frac{1}{C} \sum_i W_i - W_y = \overline{W} - W_y$, pointing from W_y to the averaged weight vector $\overline{W}$. It only depends on y and fails

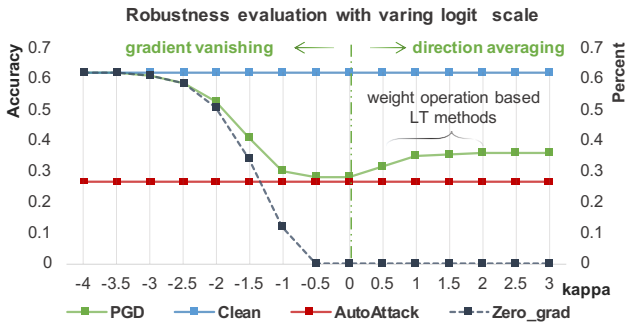

Figure 2. A_{nat} and A_{rob} under AA are invariant to logit scaling, while A_{rob} under PGD is easily over estimated when using weight operation based LT methods. The zero in gradients occurs on originally correctly predicted images with logit stretching.

to take sample-specific properties into consideration; and since it is fixed throughout the inner maximization procedure of Eqn. 2, the iterative attack actually degenerates to be single-step. Consequently, the attack is weakened to some extent so that A_{rob} gradually converges to a fixed value as κ goes up, leading to fake gain of performance that weight operation based methods suffer from.

Boundary error matters for robust accuracy. When erasing the false sense of security, the reliable A_{rob} under AA in Table 1 seems not significantly affected as A_{nat} raises. In fact, a huge gap between natural and robust accuracy is observed, and it continuously widens as the former improves, which can be reflected by the term R_{bdy}[1] in Eqn. 1. The phenomenon exposes the importance of controlling R_{bdy} in order to improve both A_{nat} and A_{rob}, which is an issue that many LT methods do not promise to solve. However, we found that cosine classifier based methods could benefit from a relatively smaller R_{bdy} compared with linear classifiers. One evidence is that in Table 1, the model trained with a vanilla cosine classifier exhibits the lowest gap between A_{nat} and A_{rob} under AA, marked in blue in the last column. We would analyze the reason behind it in the next section and how to take advantage of its good property.

4. Methodology

Earlier discoveries cast light on two key factors of solving this challenging problem: 1) a proper feature and classifier embedding helps to achieve a lower boundary error R_{bdy}, and 2) the combination of long-tailed recognition (LT) methods with adversarial training (AT) framework would benefit A_{nat}. Hence, we propose a clean yet effective approach which consists of two components, *i.e.* scale-invariant classifier and two-stage re-balancing, to achieve **Ro**bust and **Bal**anced predictions, namely **RoBal**.

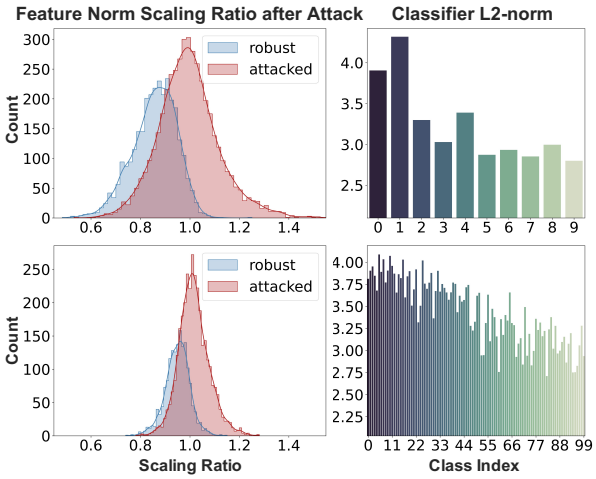

Figure 3. **Left:** the feature norm distribution shifts after adversarial attack, the successfully attacked samples statistically have a larger scaling factor than those that stay robust; **Right:** the classifier weight norm roughly decreases towards the tail classes.

4.1. Scale-invariant Classifier

In a basic classification task using a standard linear classifier, the predicted logit of class i can be represented as:

$$W_i^T f(x) + b_i = ||W_i|| \cdot ||f(x)|| \cos\theta_i + b_i, \qquad (6)$$

where we can see that the prediction depends on three factors: 1) the scale of weight vector $||W_i||$ and feature vector $||f(x)||$; 2) the angle $\cos\theta_i$ between them; and 3) the bias of the classifier b_i. In this section we would focus on the first factor to show the importance of being scale-invariant.

Firstly, the decomposition above indicates that the prediction of a sample can be changed by simply scaling its norm in the feature space. We consider this to be one of the schemes adversarial examples use to confuse the model, leading to different feature norm distributions between "successfully attacked" and "robust" samples. To be specific, the originally correctly predicted images can be separated into two groups by whether the attack is successful. We then calculate the scaling ratio $\|f(x+\delta)\| \,/\, \|f(x)\|$ between each attacked and clean input pair. We could observe different distributions of the two groups in Fig. 3, where a successful attack is more likely to happen with a relatively higher scaling ratio.

Besides the feature embedding, the scales of the weight vectors $||W_i||$ in a linear classifier would also induce problems to the long-tail scenario. Specifically, they usually decrease towards the tail classes as shown in Fig. 3, which is also observed in previous works [16, 19]. Different weight scales result in biased decision boundaries (Fig. 4) and hurt recognition performance. This could be alleviated by adjusting the class-specific bias. But it still suffers from a high

[1]The case that a wrong prediction being "corrected" after the attack rarely happens, so $\mathbf{R}_{bdy}$ can be basically reflected by $\mathsf{A}_{nat} - \mathsf{A}_{rob}$.

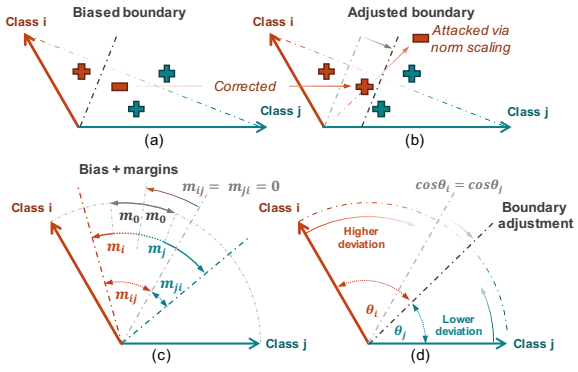

Figure 4. **(a)** and **(b)**: biased decision boundary is induced by imbalanced weight norms while boundary adjustment helps correct some mistakes; feature norm scaling can result in a successful attack; **(c)**: margin construction with *hyper-parameters ignored* for simplicity; **(d)**: boundary adjustment at inference stage.

adversarial risk even after boundary adjustment considering the varying feature norm, as shown in Fig. 4

Based on the observation and analysis, a natural idea is to remove the influence of scales from both the features and weights. Therefore, a scale-invariant classifier, *e.g.*, cosine classifier that limits the vectors to a hyper-sphere [29], could be a proper choice. The benefit of it reducing R_{bdy} has been revealed in Table 1. And we would further introduce how to address the issue of imbalance via its combination with re-balancing strategies.

4.2. Two-stage Re-balancing

Based on the normalized features and classifier weights, we further consider the problem of long-tailed data distribution. Recall the knowledge we gain from Sec. 3, where significant natural accuracy gain via effective elimination of imbalance can happen either during training or simply at inference time. Accordingly, we propose a two-stage re-balancing framework in this section, which exactly focuses on the other two factors in Eqn. 6, namely $\cos\theta_i$ and b_i: 1) margins introduced to the cosine classifier at training stage promote a more compact representation learning, where both class-aware and pair-aware margin engineering would boost network performance in our long-tailed scenario; 2) boundary adjustment at inference stage further tackles the issue of higher variance and deviation in tail classes. Different cases of the compensation and cooperation between them is explored in ablation study in Sec. 5.

Class-aware Margin on Training. A straight forward and widely used idea to take class imbalance into consideration is to assign class-specific bias in the CE loss during training. Following Ren *et al.* [30] and Menon *et al.* [27], we adopt the form of $b_i = \tau_b \log(n_i)$, and the modified CE Loss becomes:

$$\mathcal{L}_0 = -\log(\frac{e^{z_y+b_y}}{\sum_i e^{z_i+b_i}}) = \log(1+\sum_{i\neq y} e^{z_i-z_y+\tau_b \log(\frac{n_i}{n_y})}), \tag{7}$$

where τ_b is a hyper-parameter controlling the bias value calculation. However, on considering the formulation in the manner of margin, we find that the margin from the ground truth class y to class i, namely $\tau_b \log(n_i/n_y)$, would become negative when $n_y > n_i$, leading to less discriminating representation and classifier learning of head classes. To deal with the issue, we further add a class-aware margin term together with the pre-defined b_i, which assigns a larger margin value to the head class in compensation:

$$m_i = \frac{\tau_m}{s} \log \frac{n_i}{n_{min}} + m_0. \tag{8}$$

Here the first term would increase along with with n_i while achieving its lowest at zero when $n_i = n_{min}$, and τ_m is the hyper-parameter to control the trend; the second term $m_0 > 0$ is a uniform margin for all classes, as is a commonly used strategy for cosine classifier based networks [41]; s represents a temperature here to expand the value range of the cosine outputs, which helps to present a more clearly formulated loss function as below:

$$\begin{aligned}\mathcal{L}_1 &= -\log\left(\frac{e^{s(\cos\theta_y - m_y)+b_y}}{e^{s(\cos\theta_y - m_y)+b_y} + \sum_{i\neq y} e^{s\cos\theta_i + b_i}}\right) \\ &= \log(1+\sum_{i\neq y} e^{s(\cos\theta_i - \cos\theta_y + m_{yi})}),\end{aligned} \tag{9}$$

where $\cos\theta_i = \widetilde{W}_i^T \widetilde{f}(x)$, and

$$\begin{aligned}m_{yi} &= \frac{\tau_b}{s}\log(\frac{n_i}{n_y}) + \frac{\tau_m}{s}\log(\frac{n_y}{n_{min}}) + m_0 \\ &= \frac{(\tau_b - \tau_m)}{s}\log(\frac{n_i}{n_y}) + \frac{\tau_m}{s}\log(\frac{n_i}{n_{min}}) + m_0.\end{aligned} \tag{10}$$

Notice that we here adopt $\widetilde{W}_i = W_i/(||W_i|| + \gamma)$ here, which is slightly different from Sec. 3 while we empirically find it able to produce slightly better performance. The **first** line of formulation in Eqn. 10 constructs the margin between ground truth class y and a negative class i: a composition of a pair-aware margin $\log(n_i/n_y)$ scaled by τ_b, a class-aware margin $\log(n_i/n_{min})$ scaled by τ_m, and a uniform m_0, as shown in Fig. 4. While The **second** line reveals a more direct relationship to the data distribution: n_y occurs only in the first term to assign a larger margin to tail classes when $\tau_b - \tau_m > 0$. It encourages a more compact and discriminating learning on them and especially benefits the imbalance learning process. We would show how each term effects the training in the ablation study.

Class-specific Bias on Inference. With a normalized classifier, the decision boundary is naturally unbiased at inference time. However, the sparse data distribution in the tail

Table 2. Experimental results on CIFAR-10-LT and CIFAR-100-LT with WideResnet-34-10.

Dataset	CIFAR-10-LT						CIFAR-100-LT					
Methods	**Clean**	**FGSM**	**PGD**	**MIM**	**CW**	**AA**	**Clean**	**FGSM**	**PGD**	**MIM**	**CW**	**AA**
Plain	77.16	7.01	0.00	0.00	0.00	0.00	62.29	3.94	0.00	0.00	0.00	0.00
AT [25]	62.33	33.57	29.30	30.02	30.31	28.15	48.96	21.06	17.26	17.80	17.65	16.26
TRADES [50]	54.29	32.80	30.20	30.53	29.58	28.94	43.71	23.06	**21.13**	**21.42**	19.49	18.68
HE [29]	58.47	35.17	31.73	32.26	29.96	28.45	48.63	23.06	19.56	20.28	19.20	17.60
MMA [8]	61.51	36.40	29.29	30.38	29.59	25.91	**54.98**	19.65	13.52	13.98	12.35	14.54
AVmixup [21]	66.97	33.90	28.40	29.67	26.43	24.39	52.45	23.28	19.04	20.78	14.82	12.60
RoBal-N	**75.52**	40.04	33.50	34.57	33.68	31.72	51.63	22.81	19.01	19.50	19.42	18.16
RoBal-R	74.51	**40.55**	**33.87**	**34.92**	**34.12**	**32.04**	50.38	**23.59**	19.48	20.13	**20.16**	**18.69**

leads to higher uncertainty [18] and feature deviation [48], while head classes would benefit from a more compact feature embedding and concise classifier learning. As a result, when using a uniform margin m_0 with $\tau_b = \tau_m = 0$ during training, we can still observe an obvious decrease in the recall of the tail classes. Thus a post processing strategy to adjust the cosine boundary is still needed, as can be formulated as a dual process to the pre-defined b_i during training in Eqn. 8. τ_p is introduced here and the inference becomes:

$$\underset{i\in[C]}{\operatorname{argmax}}\, s\cdot\cos\theta_i - \tau_p \log(\frac{n_i}{\sum_j n_j}) \tag{11}$$

Actually, when class-specific margins are added along with the uniform one, the dependence on boundary adjustment is eliminated, as to be explored in Sec. 5.

Regularization Term. Finally, inspired by the some of the well-known AT variants [17, 50, 8], an additional regularization term between the paired features or logits produced by the clean and perturbed images would further promote the robustness performance. Different kinds of regularization terms can be easily adopted into the training framework via modification on the loss function, and we follow Zhang *et al*. [50] to take advantage of a KL-divergence term, and the overall loss function become:

$$\mathcal{L} = \mathcal{L}_1(x+\delta, y) + \alpha \cdot KL(\widetilde{W}_i^T \widetilde{f}(x+\delta),\ \widetilde{W}_i^T \widetilde{f}(x)) \tag{12}$$

where δ is the perturbation generated by inner maximization guided by a plain cross-entropy loss, performed on the direct outputs of the cosine operation without margins.

5. Experiments

Datasets. We conduct experiments on the long-tailed versions of CIFAR-10 and CIFAR-100 following [7]. Imbalance Ratio (IR) in the main experiments is set as 50 and 10, respectively. Experimental results with various IRs are also provided in Table 4.

Evaluation Metrics. On evaluating model robustness, the allowed l_∞ norm-bounded perturbation is $\epsilon = 8/255$. Attacks conducted include the single-step attack FGSM [9] and several iterative attacks including PGD, MIM, and C&W performed for 20 steps with a step size of $2/255$. We also use the recently proposed Auto Attack (AA) [6] which is an ensemble of different attacks and is parameter-free.

Comparison Methods. We compare our method with several state-of-the-art defense methods besides the standard AT [25], including TRADES [50], MMA [8], HE [29], and AVmixup [21], among which AVmixup [21] is re-implemented and the others are evaluated with the officially released code. **Implementation details** on our network training, hyper-parameter setting, and attacking algorithms are included in the supplementary material.

5.1. Comparison Results

The comparison with other defense methods is reported in Table 2. Since a trade-off between **N**atural and **R**obust accuracy usually exists, we report our results with different emphasis, denoted by *RoBal-N* and *RoBal-R*, respectively. The setting of hyper-parameters to control the trade-off can be found in Sec.A. On CIFAR-10-LT, our method significantly outperforms all the compared ones on both A_{nat} of clean images and A_{rob} under five different attacks. On CIFAR-100-LT, TRADES [50] and RoBal achieve comparable results on robust accuracy. However, they significantly sacrifice the performance on A_{nat}, while our method also consistently boosts A_{nat} compared with AT baseline. AVmixup [21] and MMA [8] (on CIFAR-100-LT) achieve decent A_{nat} while suffering from the poor robust performance under AA. It is noted that the overall improvement in CIFAR-100-LT is less significant than CIFAR-10-LT, possibly due to the smaller imbalance ratio or the increase of class number; thus we also provide preliminary results on ImageNet-LT [24] with 1000 classes in Sec.B.4.

5.2. Ablation Study

Here we explore the effect of hyper-parameters during the two-stage re-balancing. We mainly discuss three terms according to Eqn. 10, namely m_0, $\tau_b - \tau_m$, and τ_m. We change the critical variable while fix the others for analysis as shown in Table 3. Several interesting observations include: 1) a proper $m_0 > 0$ leads to higher A_{rob} yet lower

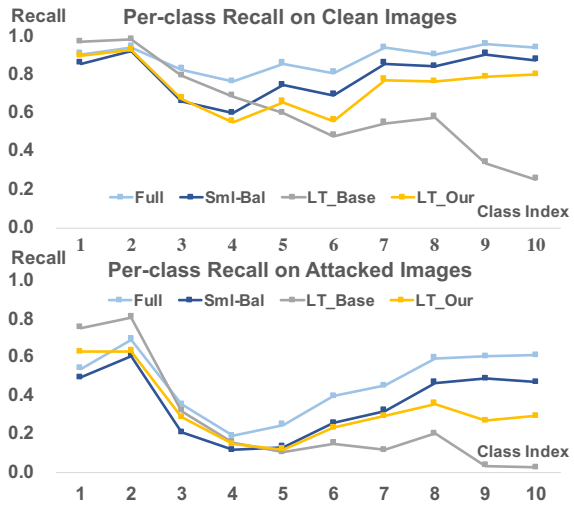

Figure 5. Results under different data scales and distributions.

A_{nat} at both stages; 2) a higher $\tau_b - \tau_m$ promotes natural accuracy at the end of training stage, yet a peak is observed for robustness; 3) a larger τ_m helps reduce the boundary error R_{bdy}, while it could hurt A_{nat} if set too large. *Please refer to the supplementary for more experimental results.*

Table 3. Effect of hyper-parameters. We fix $\tau_m = \tau_b = 0$ in block 1, $m_0 = 0.1, \tau_m = 0$ in block 2, and $m_0 = 0.1, \tau_b - \tau_m = 1.2$ in block 3. ** denotes the key metric that is worth noting.

End of training stage			**Inference with τ_p***	
m_0	**Clean**	**AA**	**Clean****	**AA**
0	63.51	28.47	75.98	30.46
0.1	63.17	29.13	75.51	31.24
0.2	60.6	29.04	74.83	32.04
0.3	56.59	28.63	71.31	31.28
$\tau_b - \tau_m$	**Clean****	**AA**	**Clean**	**AA**
0	63.51	28.47	75.51	31.24
0.5	68.58	30.55	75.77	30.94
1	73.93	31.52	75.42	31.61
1.5	74.67	30.95	74.67	30.95
τ_m	**Clean**	**AA**	**Training gap****	
-0.3	74.88	30.66	44.22	
0	75.08	31.79	43.29	
0.3	74.51	32.04	42.47	
0.6	71.84	31.41	40.43	

5.3. Further Analysis

Effect of Data Scale. Since a long-tailed dataset differs from the full dataset by both class-wise sample numbers and the overall data scale, we conduct a comparison with 1) the original full dataset (Full) and 2) a dataset with the same number of samples as the long-tailed version but is uniformly distributed (Sml-Bal). From the per-class recall shown in Fig. 5, we can see that: (1) Performance of Sml-Bal are uniformly lower than that of the full dataset. (2) The basic AT framework produces an apparent decrease in recall from head to tail on both clean and attacked images. Specifically, head classes gain even higher robustness than balanced baselines, indicating that the intrinsic prediction bias raise their resistance to attack. (3) Our RoBal (LT-Our) applied to the long-tailed dataset efficiently re-balances the per-class recall compared with the baseline (LT-base).

Table 4. Experimental results on CIFAR-10-LT with different IRs

IR	**Methods**	**Clean**	**PGD**	**MIM**	**CW**	**AA**
100	AT	56.72	27.27	27.87	27.80	25.97
	TRADES	45.67	26.97	27.29	26.53	25.93
	Our	**68.07**	**30.35**	**31.25**	**30.55**	**28.97**
50	AT	62.33	29.30	29.72	30.31	28.15
	TRADES	54.29	30.20	30.53	29.58	28.94
	Our	**73.93**	**34.24**	**35.37**	**34.58**	**32.70**
20	AT	74.09	33.59	34.=65	34.27	32.02
	TRADES	65.17	34.65	35.29	34.07	33.06
	Our	**78.49**	**39.17**	**40.44**	**39.38**	**37.58**
10	AT	79.45	37.11	37.93	38.31	35.51
	TRADES	72.92	39.15	39.95	38.41	37.33
	Our	**81.20**	**40.22**	**41.75**	**40.91**	**38.90**

Effect of Imbalance Ratio. We also constructed long-tailed datasets with different imbalanced ratios (**IR**) following [7] to evaluate the performance of AT, TRADES [50], and our methods. As shown in Table 4, our method outperforms the baseline and TRADES [50] remarkably on both natural accuracy and robust accuracies over different IRs.

6. Conclusion

In this paper, 1) we first reveal the negative impacts induced by long-tailed data distribution on both recognition performance and adversarial robustness, uncovering the intrinsic challenges of this problem. 2) Then, a systematic study on existing long-tailed recognition approaches and their combination with the adversarial training framework contributes several valuable observations. 3) Finally, inspired by them, we propose a clean yet effective framework that benefits from the norm-invariant property of cosine classifier and a two-stage re-balancing framework, which outperforms existing state-of-the-art defense methods. To our best knowledge, we are the first to tackle adversarial robustness under long-tailed distribution, which we believe would be a significant step towards real-world robustness.

Acknowledgements. This research was conducted in collaboration with SenseTime. This work is supported by GRF 14203518, ITS/431/18FX, CUHK Agreement TS1712093, NTU NAP and A*STAR through the Industry Alignment Fund - Industry Collaboration Projects Grant, and the Shanghai Committee of Science and Technology, China (Grant No. 20DZ1100800).

References

[1] Mateusz Buda, Atsuto Maki, and Maciej A Mazurowski. A systematic study of the class imbalance problem in convolutional neural networks. *Neural Networks*, 106:249–259, 2018. 2

[2] Kaidi Cao, Colin Wei, Adrien Gaidon, Nikos Arechiga, and Tengyu Ma. Learning imbalanced datasets with label-distribution-aware margin loss. In *Advances in Neural Information Processing Systems (NIPS)*, pages 1565–1576, 2019. 1, 2, 3, 4

[3] Nicholas Carlini and David Wagner. Towards evaluating the robustness of neural networks. In *IEEE Symposium on Security and Privacy (SP)*, pages 39–57. IEEE, 2017. 4

[4] Yair Carmon, Aditi Raghunathan, Ludwig Schmidt, John C Duchi, and Percy S Liang. Unlabeled data improves adversarial robustness. In *Advances in Neural Information Processing Systems (NIPS)*, pages 11192–11203, 2019. 2

[5] Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. Smote: synthetic minority oversampling technique. *Journal of artificial intelligence research*, 16:321–357, 2002. 2

[6] Francesco Croce and Matthias Hein. Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks. In *International Conference on Machine Learning (ICML)*, 2020. 3, 4, 7

[7] Yin Cui, Menglin Jia, Tsung-Yi Lin, Yang Song, and Serge Belongie. Class-balanced loss based on effective number of samples. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 1, 2, 3, 4, 7, 8

[8] Gavin Weiguang Ding, Yash Sharma, Kry Yik Chau Lui, and Ruitong Huang. MMA training: Direct input space margin maximization through adversarial training. In *International Conference on Learning Representations (ICLR)*, 2020. 2, 7

[9] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. 2015. 7

[10] Sven Gowal, Chongli Qin, Jonathan Uesato, Timothy Mann, and Pushmeet Kohli. Uncovering the limits of adversarial training against norm-bounded adversarial examples, 2020. 3

[11] Agrim Gupta, Piotr Dollar, and Ross Girshick. LVIS: A dataset for large vocabulary instance segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 1

[12] Haibo He and Edwardo A Garcia. Learning from imbalanced data. *IEEE Transactions on Knowledge and Data Engineering*, 21(9):1263–1284, 2009. 1, 2

[13] Dan Hendrycks, Kimin Lee, and Mantas Mazeika. Using pre-training can improve model robustness and uncertainty. In *International Conference on Machine Learning (ICML)*, pages 2712–2721, 2019. 2

[14] Muhammad Abdullah Jamal, Matthew Brown, Ming-Hsuan Yang, Liqiang Wang, and Boqing Gong. Rethinking class-balanced methods for long-tailed visual recognition from a domain adaptation perspective. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7610–7619, 2020. 2

[15] Nathalie Japkowicz and Shaju Stephen. The class imbalance problem: A systematic study. *Intelligent Data Analysis*, 6(5):429–449, 2002. 2

[16] Bingyi Kang, Saining Xie, Marcus Rohrbach, Zhicheng Yan, Albert Gordo, Jiashi Feng, and Yannis Kalantidis. Decoupling representation and classifier for long-tailed recognition. In *Eighth International Conference on Learning Representations (ICLR)*, 2020. 1, 2, 3, 4, 5

[17] Harini Kannan, Alexey Kurakin, and Ian Goodfellow. Adversarial logit pairing. *arXiv preprint arXiv:1803.06373*, 2018. 7

[18] Salman Khan, Munawar Hayat, Syed Waqas Zamir, Jianbing Shen, and Ling Shao. Striking the right balance with uncertainty. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 7

[19] Byungju Kim and Junmo Kim. Adjusting decision boundary for class imbalanced learning. *IEEE Access*, pages 81674–81685, 2020. 2, 3, 4, 5

[20] Jaehyung Kim, Jongheon Jeong, and Jinwoo Shin. M2m: Imbalanced classification via major-to-minor translation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 2

[21] Saehyung Lee, Hyungyu Lee, and Sungroh Yoon. Adversarial vertex mixup: Toward better adversarially robust generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 272–281, 2020. 7

[22] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollar. Focal loss for dense object detection. *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017. 2, 3, 4

[23] Jialun Liu, Yifan Sun, Chuchu Han, Zhaopeng Dou, and Wenhui Li. Deep representation learning on long-tailed data: A learnable embedding augmentation perspective. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2970–2979, 2020. 2

[24] Ziwei Liu, Zhongqi Miao, Xiaohang Zhan, Jiayun Wang, Boqing Gong, and Stella X. Yu. Large-scale long-tailed recognition in an open world. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 1, 2, 3, 7

[25] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations (ICLR)*, 2018. 2, 3, 7

[26] Chengzhi Mao, Ziyuan Zhong, Junfeng Yang, Carl Vondrick, and Baishakhi Ray. Metric learning for adversarial robustness. In *Advances in Neural Information Processing Systems (NIPS)*, pages 480–491, 2019. 2

[27] Aditya Krishna Menon, Sadeep Jayasumana, Ankit Singh Rawat, Himanshu Jain, Andreas Veit, and Sanjiv Kumar. Long-tail learning via logit adjustment, 2020. 2, 3, 4, 6

[28] Tianyu Pang, Xiao Yang, Yinpeng Dong, Hang Su, and Jun Zhu. Bag of tricks for adversarial training, 2020. 2

[29] Tianyu Pang, Xiao Yang, Yinpeng Dong, Kun Xu, Hang Su, and Jun Zhu. Boosting adversarial training with hypersphere embedding. 2020. 2, 6, 7

[30] Jiawei Ren, Cunjun Yu, Shunan Sheng, Xiao Ma, Haiyu Zhao, Shuai Yi, and Hongsheng Li. Balanced meta-softmax for long-tailed visual recognition. *Advances in Neural Information Processing Systems (NIPS)*, 2020. 2, 3, 4, 6

[31] Mengye Ren, Wenyuan Zeng, Bin Yang, and Raquel Urtasun. Learning to reweight examples for robust deep learning. *arXiv preprint arXiv:1803.09050*, 2018. 2

[32] Leslie Rice, Eric Wong, and J Zico Kolter. Overfitting in adversarially robust deep learning. *International Conference on Machine Learning (ICML)*, 2020. 2

[33] Ludwig Schmidt, Shibani Santurkar, Dimitris Tsipras, Kunal Talwar, and Aleksander Madry. Adversarially robust generalization requires more data. In *Advances in Neural Information Processing Systems (NIPS)*, pages 5014–5026, 2018. 2

[34] Ali Shafahi, Amin Ghiasi, Furong Huang, and Tom Goldstein. Label smoothing and logit squeezing: A replacement for adversarial training?, 2019. 2

[35] Ali Shafahi, Mahyar Najibi, Mohammad Amin Ghiasi, Zheng Xu, John Dickerson, Christoph Studer, Larry S Davis, Gavin Taylor, and Tom Goldstein. Adversarial training for free! In *Advances in Neural Information Processing Systems (NIPS)*, pages 3358–3369, 2019. 2

[36] Li Shen, Zhouchen Lin, and Qingming Huang. Relay backpropagation for effective learning of deep convolutional neural networks. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 467–482. Springer, 2016. 2, 3

[37] Jun Shu, Qi Xie, Lixuan Yi, Qian Zhao, Sanping Zhou, Zongben Xu, and Deyu Meng. Meta-weight-net: Learning an explicit mapping for sample weighting. In *Advances in Neural Information Processing Systems (NIPS)*, 2019. 2

[38] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013. 1

[39] Kaihua Tang, Jianqiang Huang, and Hanwang Zhang. Long-tailed classification by keeping the good and removing the bad momentum causal effect. In *Advances in Neural Information Processing Systems (NIPS)*, 2020. 3, 4

[40] 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 *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8769–8778, 2018. 1

[41] Hao Wang, Yitong Wang, Zheng Zhou, Xing Ji, Dihong Gong, Jingchao Zhou, Zhifeng Li, and Wei Liu. Cosface: Large margin cosine loss for deep face recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5265–5274, 2018. 3, 6

[42] Yisen Wang, Difan Zou, Jinfeng Yi, James Bailey, Xingjun Ma, and Quanquan Gu. Improving adversarial robustness requires revisiting misclassified examples. In *International Conference on Learning Representations (ICLR)*, 2020. 2

[43] Eric Wong, Leslie Rice, and J Zico Kolter. Fast is better than free: Revisiting adversarial training. In *International Conference on Learning Representations (ICLR)*, 2019. 2

[44] Tong Wu, Qingqiu Huang, Ziwei Liu, Yu Wang, and Dahua Lin. Distribution-balanced loss for multi-label classification in long-tailed datasets. In *European Conference on Computer Vision (ECCV)*, 2020. 1

[45] Cihang Xie, Mingxing Tan, Boqing Gong, Jiang Wang, Alan L Yuille, and Quoc V Le. Adversarial examples improve image recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 819–828, 2020. 2

[46] Cihang Xie, Mingxing Tan, Boqing Gong, Alan Yuille, and Quoc V Le. Smooth adversarial training. *arXiv preprin:2006.14536*, 2020. 2

[47] Cihang Xie and Alan Yuille. Intriguing properties of adversarial training at scale. In *International Conference on Learning Representations (ICLR)*, 2019. 2, 4

[48] Han-Jia Ye, Hong-You Chen, De-Chuan Zhan, and Wei-Lun Chao. Identifying and compensating for feature deviation in imbalanced deep learning. *arXiv preprin:2001.01385*, 2020. 3, 4, 7

[49] Dinghuai Zhang, Tianyuan Zhang, Yiping Lu, Zhanxing Zhu, and Bin Dong. You only propagate once: Accelerating adversarial training via maximal principle. In *Advances in Neural Information Processing Systems (NIPS)*, pages 227–238, 2019. 2

[50] Hongyang Zhang, Yaodong Yu, Jiantao Jiao, Eric Xing, Laurent El Ghaoui, and Michael Jordan. Theoretically principled trade-off between robustness and accuracy. In *International Conference on Machine Learning (ICLR)*, pages 7472–7482, 2019. 2, 7, 8

[51] Boyan Zhou, Quan Cui, Xiu-Shen Wei, and Zhao-Min Chen. BBN: Bilateral-branch network with cumulative learning for long-tailed visual recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–8, 2020. 1, 2, 3