Class-Conditional Sharpness-Aware Minimization for Deep Long-Tailed Recognition

Zhipeng Zhou[1,†], Lanqing Li[2,4,†], Peilin Zhao[3,⋆], Pheng-Ann Heng[2], Wei Gong[1,⋆]

[1]University of Science and Technology of China, [2]The Chinese University of Hong Kong,
[3]Tencent AI Lab, [4]Zhejiang Lab

zzp1994@mail.ustc.edu.cn, lanqingli1993@gmail.com, masonzhao@tencent.com,
pheng@cse.cuhk.edu.hk, weigong@ustc.edu.cn

Abstract

It's widely acknowledged that deep learning models with flatter minima in its loss landscape tend to generalize better. However, such property is under-explored in deep long-tailed recognition (DLTR), a practical problem where the model is required to generalize equally well across all classes when trained on highly imbalanced label distribution. In this paper, through empirical observations, we argue that sharp minima are in fact prevalent in deep long-tailed models, whereas naïve integration of existing flattening operations into long-tailed learning algorithms brings little improvement. Instead, we propose an effective two-stage sharpness-aware optimization approach based on the decoupling paradigm in DLTR. In the first stage, both the feature extractor and classifier are trained under parameter perturbations at a class-conditioned scale, which is theoretically motivated by the characteristic radius of flat minima under the PAC-Bayesian framework. In the second stage, we generate adversarial features with class-balanced sampling to further robustify the classifier with the backbone frozen. Extensive experiments on multiple long-tailed visual recognition benchmarks show that, our proposed ***C****lass-****C****onditional* ***S****harpness-****A****ware* ***M****inimization (CC-SAM), achieves competitive performance compared to the state-of-the-arts. Code is available at* https://github.com/zzpustc/CC-SAM.

1. Introduction

Modern deep learning models, composed of multiple neural network layers with millions of parameters, have achieved remarkable successes in computer vision [24, 33, 41, 48]. A key enabler of deep learning is the collection of large-scale datasets [29, 42, 64], which are normally split into training and testing sets with presumably i.i.d. samples. However, such scenario provides relatively trivial tests for the generalization of machine learning models. In practice, label [17, 25, 56] and domain [14, 18, 23] distribution shifts are prevalent, due to the disparity between the data preparation and evaluation protocols. A classical example is imbalanced [15] or long-tailed recognition [60], where a model is trained on highly imbalanced source label distribution $p_s(\boldsymbol{y})$ while evaluated on a uniform target label distribution $p_t(\boldsymbol{y})$.

In this paper, we focus on the practical yet challenging deep long-tailed recognition (DLTR) problem, which is inherent in the visual world [32, 60] with fundamental connections to many disciplines such as the power-law scaling in network science [2] and the Pareto principle in economics [39]. In computer vision, numerous deep long-tailed learning studies have emerged in recent years, which mainly belong to 5 categories: class re-balancing [7, 10, 28, 40, 45, 52, 54], information augmentation [22, 27, 31, 50, 55], decoupled training [20, 58, 62], representation learning [9, 32, 51, 59, 65] and ensemble learning [6, 53, 63].

In this work, we propose a novel approach to DLTR from a *distinct angle*, by seeking out flat minima in the loss landscape of modern neural networks to ensure model robustness under parameter perturbation. Such optimization strategy, termed *flattening* in our context, have been shown in a myriad of literature to effectively improve generalization of deep learning models in terms of supervised learning [13, 21, 35, 43], self-supervised learning [30] and continual learning [11, 44]. However, application and adaption of flattening in the context of DLTR remain under-explored. To fill this gap, we first show later in this paper

† Equal contribution. ⋆ Corresponding authors. Work was primarily done when Z. Zhou worked as L. Li's intern at Tencent AI Lab, Shenzhen.

(Section 2.2.2) that existing flattening methods are ineffective for long-tailed learning, consistent with the observation from a very recent paper [49], due to severe label distribution shifts. Accordingly, we present a new efficient variant of the sharpness-aware minimization (SAM) [13] technique based on the Decoupling paradigm [20] of DLTR, which leverages the invariance of the class conditional distribution between the source and the target domain. In a nutshell, our contributions are three-fold:

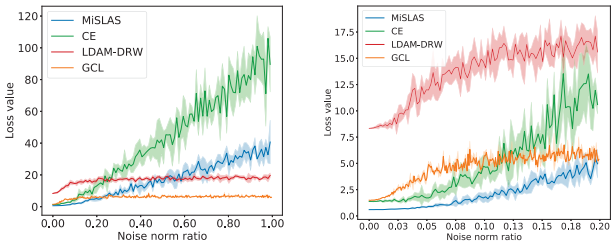
- Through the lens of flattening, we corroborate a very recent observation [49] that existing sharpness-aware minimization techniques are suboptimal for deep long-tailed learning, justifying the need of more effective approaches to this important and practical issue.
- We introduce **C**lass-**C**onditional **S**harpness-**A**ware **M**inimization (CC-SAM), a novel algorithm tailored for DLTR that improves model generalization by robust training against class-conditioned parameter perturbation. This technique is motivated by the characteristic radius of flat minima we derive under the PAC-Bayesian framework and can be implemented efficiently.
- We integrate CC-SAM with the two-stage decoupling paradigm of DLTR. Extensive experiments demonstrate that our method can achieve competitive performance on multiple public DLTR benchmarks, with remarkable robustness to out-of-distribution samples for open long-tailed recognition [32].

2. Preliminaries

2.1. Problem Setup

Throughout the paper, we denote scalars as s, vectors as $\boldsymbol{s}$, sets as $\mathcal{S}$ and equality by definition as $:=$. For a typical d-way classification task in DLTR, suppose we are given a training dataset $\mathcal{S} = \bigcup_{i=1}^{n}\{(\boldsymbol{x}_i, \boldsymbol{y}_i)\}$, where n is the total number of samples. Let's denote $\{n_1, n_2, ..., n_d\}$ as the sample number of each class, where $n = \sum_i n_i$. Without loss of generality, we assume $n_i < n_j$ if $i < j$, and usually $n_d \gg n_1$, following a highly imbalanced class distribution. We further write the training (source) distribution as $p_s(\boldsymbol{x}, \boldsymbol{y}) = p_s(\boldsymbol{x}|\boldsymbol{y})p_s(\boldsymbol{y})$ and the testing (target) distribution as $p_t(\boldsymbol{x}, \boldsymbol{y}) = p_t(\boldsymbol{x}|\boldsymbol{y})p_t(\boldsymbol{y})$. Consider a family of models parameterized by $\boldsymbol{w} \in \mathcal{W} \subseteq \mathbb{R}^k$; given a loss function l: $\mathcal{W} \times \mathcal{X} \times \mathcal{Y} \to \mathbb{R}_+$, we define the empirical training loss $L_{\mathcal{S}}(\boldsymbol{w}) := \frac{1}{n}\sum_{i=1}^{n} l(\boldsymbol{w}, \boldsymbol{x_i}, \boldsymbol{y_i})$ and the population testing loss $L_{\mathcal{T}}(\boldsymbol{w}) := \mathbb{E}_{(\boldsymbol{x},\boldsymbol{y})\sim p_t(\boldsymbol{x},\boldsymbol{y})}[l(\boldsymbol{w}, \boldsymbol{x}, \boldsymbol{y})]$. Having observed only $\mathcal{S}$, the goal is to optimize for model parameters $\boldsymbol{w}$ having lowest risk $L_{\mathcal{T}}(\boldsymbol{w})$ at test time.

In DLTR, a key challenge is the label distribution shift between the training and testing sets, i.e., $p_s(\boldsymbol{y}) \neq p_t(\boldsymbol{y})$ [17]. However, it's logical to assume *no distribution shift* within each individual class. That is, the training and testing samples are drawn i.i.d. from the same class-conditional distribution, which indicates $p_s(\boldsymbol{x}|\boldsymbol{y}) = p_t(\boldsymbol{x}|\boldsymbol{y})$. Here, we build our flattening algorithm upon the decoupling paradigm of DLTR [20], where the feature extractor $f(\boldsymbol{x}; \boldsymbol{\varphi})$ and classifier $h(\boldsymbol{z}; \boldsymbol{\theta})$ are trained by distinct sampling strategies in a two-stage manner, with $\boldsymbol{w} := (\boldsymbol{\varphi}, \boldsymbol{\theta})$.



(a) 1D global loss landscape. (b) 1D local loss landscape.

Figure 1. Loss value vs. noise norm ratio $||\boldsymbol{\epsilon}||_2/||\boldsymbol{\theta}||_2$. All experiments are conducted on the CIFAR-10-LT with an imbalance ratio of 100, implemented with the same backbone ResNet-32 and averaged over 5 random seeds.

2.2. Motivation

Previous studies have shown strong connection between the geometry of the loss landscape and generalization of deep learning models [1, 16, 19, 21]. In this section, we first empirically demonstrate that many deep long-tailed models exhibit sharp minima, and then show that naïve application of classical flattening operations results in limited improvement. Motivated by these observations, we theoretically analyze the characteristic radius of flat minima in DLTR to show that a class-conditional flattening procedure is required.

2.2.1 Sharpness in Deep Long-Tailed Models

To show the pervasive sharpness in current DLTR models, we select two baselines (empirical risk minimization of cross-entropy (CE), LDAM-DRW [7]) and two advanced methods (MiSLAS [62], GCL [26]) from recent literature to conduct experiments[1]. By perturbing the parameter $\boldsymbol{\theta}$ of the classifier h with increasing noise $\boldsymbol{\epsilon}$, we observe the loss of each model on CIFAR-10-LT in Figure 1.

Figure 1(a) shows that, at global scale, the loss of CE is the sharpest across a wide noise range. However, due to the fact that DLTR methods usually employ distinct loss functions, it's hard to compare the sharpness of their loss landscapes in a fair and meaningful way. Moreover, large noise ratio can easily flip the sign of parameters and deteriorate

[1] We re-implement all models via their publicly released code.

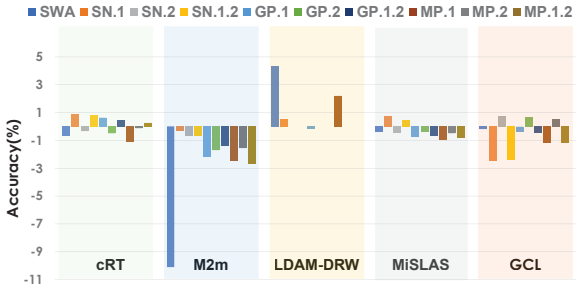


Figure 2. Performance gain of naïve integration of flattening operations with DLTR algorithms. The number represents the stage where flattening applies. For example, "SN.1.2" means we apply spectral normalization on both stages. Only the first-stage result of LDAM is reported since it's a one-stage approach. All experiments are conducted on CIFAR-100-LT with imbalance ratio of 100.

the classifier decision boundary. Therefore in this paper, we only consider locally flat minima, i.e., in the perturbative ($||\boldsymbol{\epsilon}||_2 \ll ||\boldsymbol{w}||_2$) regime shown in Figure 1(b). From this local view, *only MiSLAS* is observed to have a comparable or flatter minimum than CE, suggesting that the existing DLTR methods have poor generalization and robustness against model parameter perturbation. More detailed 2D loss landscapes covering overall, head and tail classes are depicted in **Appendix E**, and corroborate our observation in 1D.

2.2.2 Naïve Integration of Flattening Procedures

To improve the generalization of deep long-tailed models, we experiment to navigate flatter minima by integrating the existing flattening operations (stochastic weight averaging (**SWA**) [16], spectral normalization (**SN**) [4], gradients penalization (**GP**) [61], and model perturbation (**MP**) [44]) into five representative deep long-tailed learning baselines to see if they are generally beneficial for DLTR. Illustrated in Figure 2, these flattening methods bring little or negative gains in most cases. Even though in some cases they are beneficial, the corresponding baselines (mainly LDAM-DRW) are too weak so that the final performance are still sub-optimal, which calls for more effective flattening procedures for DLTR. More implementation details are provided in the **Appendix C**.

2.2.3 Characteristic Radius of Flat Minima

Let's re-consider flattening operations in DLTR from a theoretical perspective. The empirical loss $L_{\mathcal{S}}(\boldsymbol{w})$ is typically non-convex for deep neural network models, whose landscape may exhibit multiple local or global minima with similar values of $L_{\mathcal{S}}(\boldsymbol{w})$ while having drastically different generalization performance (i.e., significantly different values of $L_{\mathcal{T}}(\boldsymbol{w})$). In order to find a more effective approach for generalization in DLTR, motivated by the high correlation between sharpness of the loss landscape and model generalization [19], we follow the sharpness-aware minimization (SAM) framework [13] by optimizing an upper bound of $L_{\mathcal{T}}(\boldsymbol{w})$ derived from the PAC-Bayesian framework [34]. Given a prior over the parameters before observing any data and a posterior over the parameters dependent on the training set and learning algorithm, the PAC-Bayesian framework bounds the generalization error of any model in terms of the KL divergence between the two probability distributions. Assuming a Gaussian prior and posterior, we arrive at the following generalization bound (formal proof in **Appendix D**):

Theorem 1 *(Perturbative PAC-Bayesian Generalization Bound). For any $\rho > 0$, $0 < \delta < 1$, number of samples $n \in \mathbb{N}^+$, $k := dim(\boldsymbol{w})$, with probability at least $1-\delta$ over a training set $\mathcal{S}$ sampled i.i.d. from distribution $\mathcal{T}$,*

$$L_{\mathcal{T}}(\boldsymbol{w}) \leq \max_{||\boldsymbol{\epsilon}||_2 \leq \sqrt{k}\rho} L_{\mathcal{S}}(\boldsymbol{w}+\boldsymbol{\epsilon}) + \sqrt{\frac{\frac{||\boldsymbol{w}||_2^2}{4\rho^2} + \log(\frac{n}{\delta}) + \mathcal{O}(1)}{n-1}}.$$

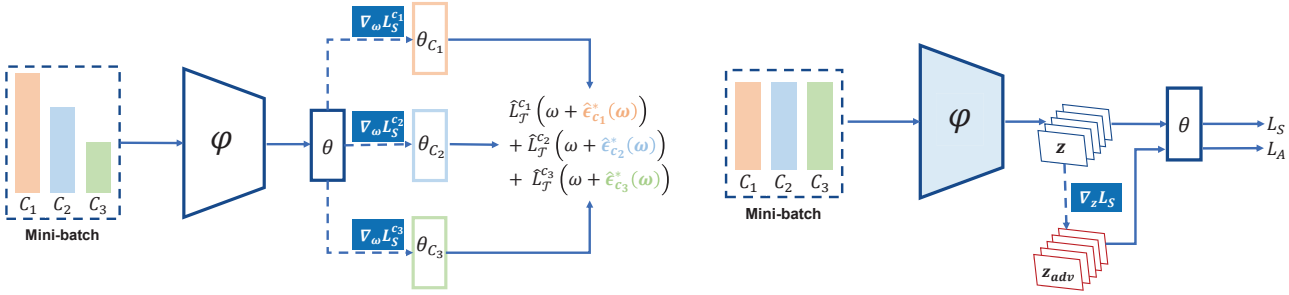
Intuitively, optimizing the perturbative bound as a surrogate function effectively seeks out parameter solutions whose neighborhoods have uniformly low empirical training loss, thus achieving a flat minimum. Now, our key insight is that the bound is convex with respect to the radius ρ of the flat minima, hence there exists an optimal ρ^* which gives the tightest generalization bound:

$$\rho^* := \arg\min_{\rho} \left[\max_{||\boldsymbol{\epsilon}||_2 \leq \sqrt{k}\rho} L_{\mathcal{S}}(\boldsymbol{w}+\boldsymbol{\epsilon}) + \sqrt{\frac{\frac{||\boldsymbol{w}||_2^2}{4\rho^2} + \log(\frac{n}{\delta})}{n-1}} \right]. \tag{1}$$

We denote ρ^* as the *characteristic radius* of the flat minima centered at $\boldsymbol{w}$. It's worth mentioning that such optimal perturbation radius is overlooked by previous literature on sharpness-aware optimization, mainly due to arguments that the PAC-Bayesian bound is too loose to capture the real generalization error [35]. However, our Theorem 1 inspired from [8, 12] ensures a non-vacuous bound (i.e., the square root term falls below 1, see empirical evidence in **Appendix B**) in the over-parameterized "deep learning" regime, rendering the characteristic radius relevant for optimization.

Moreover, the perturbative bound can be approximated by a first-order Taylor expansion, yielding

$$\begin{aligned} \max_{||\boldsymbol{\epsilon}||_2 \leq \sqrt{k}\rho} L_{\mathcal{S}}(\boldsymbol{w}+\boldsymbol{\epsilon}) &\approx \max_{||\boldsymbol{\epsilon}||_2 \leq \sqrt{k}\rho} \left[L_{\mathcal{S}}(\boldsymbol{w}) + \boldsymbol{\epsilon}^T \nabla_{\boldsymbol{w}} L_{\mathcal{S}}(\boldsymbol{w}) \right] \\ &= L_{\mathcal{S}}(\boldsymbol{w}) + \sqrt{k}\rho ||\nabla_{\boldsymbol{w}} L_{\mathcal{S}}(\boldsymbol{w})||_2 \end{aligned} \tag{2}$$



(a) **Stage 1**: Class-conditional sharpness-aware minimization. We take three classes and classifier perturbation as examples for illustration.

(b) **Stage 2**: Robust training of the classifier by progressively generating adversarial features.

Figure 3. The overall framework of CC-SAM.

and the optimal perturbation vector

$$\hat{\boldsymbol{\epsilon}}^*(\boldsymbol{w}) \approx \underset{||\boldsymbol{\epsilon}||_2 \leq \sqrt{k}\rho}{\arg\max} \left[L_{\mathcal{S}}(\boldsymbol{w}) + \boldsymbol{\epsilon}^T \nabla_{\boldsymbol{w}} L_{\mathcal{S}}(\boldsymbol{w}) \right] = \sqrt{k}\rho \frac{\nabla_{\boldsymbol{w}} L_{\mathcal{S}}(\boldsymbol{w})}{||\nabla_{\boldsymbol{w}} L_{\mathcal{S}}(\boldsymbol{w})||_2}. \tag{3}$$

Substitute Eqn 2 into 1, and ignore $\log(\frac{n}{\delta})$ by assuming $\rho \ll ||\boldsymbol{w}||_2$ as we discussed in Sec 2.2.1, obtaining

$$\rho^* \approx \left(\frac{||\boldsymbol{w}||_2}{2||\nabla_{\boldsymbol{w}} L_{\mathcal{S}}(\boldsymbol{w})||_2} \right)^{\frac{1}{2}} k^{-\frac{1}{4}} (n-1)^{-\frac{1}{4}} \tag{4}$$

with the approximated optimal generalization bound

$$\widehat{L}_{\mathcal{T}}(\boldsymbol{w}) \approx \max_{||\boldsymbol{\epsilon}||_2 \leq \sqrt{k}\rho^*} L_{\mathcal{S}}(\boldsymbol{w}+\boldsymbol{\epsilon}) + \frac{1}{2\sqrt{n-1}} \cdot \frac{||\boldsymbol{w}||_2}{\rho^*}. \tag{5}$$

The first term captures the sharpness of $L_{\mathcal{S}}$ at $\boldsymbol{w}$, and the second term serves as a regularizer on the magnitude of $\boldsymbol{w}$, akin to the L2 regularization in SAM [13]. A key difference here is that our regularization term is well motivated by theoretic interpretation from the PAC-Bayesian bound.

3. Methodology

In this section, we introduce our two-stage sharpness-aware optimization algorithm based on Decoupling [20] in detail. In the first stage of the decoupled training, both the feature extractor and classifier are trained under parameter perturbations at a class-conditioned scale. In the second stage, we generate adversarial features to further robustify the classifier while freezing the backbone. The complete pseudo-code is shown in Algorithm 1.

3.1. Stage 1: Class-Conditional Sharpness-Aware Minimization (CC-SAM)

Note that a key assumption we made in Theorem 1 when deriving the characteristic radius of flat minima is that the training data $\mathcal{S}$ are drawn i.i.d. from the target distribution $\mathcal{T}$. However, as we explained in Sec 2.1, such assumption is invalid due to the severe label distribution shift between training and testing sets in DLTR. To resolve this fundamental conflict, we turn to the class conditional distribution $p_s(\boldsymbol{x}|\boldsymbol{y})$ and $p_t(\boldsymbol{x}|\boldsymbol{y})$ instead, for which the i.i.d. assumption reasonably holds. Accordingly, we decompose the total loss $L_{\mathcal{T}}(\boldsymbol{w})$ and derive a generalization bound for each class separately, obtaining

$$L_{\mathcal{T}}(\boldsymbol{w}) = \sum_{c=1}^{k} L_{\mathcal{T}}^c(\boldsymbol{w}) = \sum_{c=1}^{k} \mathbb{E}_{\boldsymbol{x}\sim p_t(\boldsymbol{x}|c)} \left[l(\boldsymbol{w}, \boldsymbol{x}, c) \right], \tag{6}$$

$$\widehat{L}_{\mathcal{T}}^c(\boldsymbol{w}) = \max_{||\boldsymbol{\epsilon}||_2 \leq \sqrt{k}\rho_c^*} L_{\mathcal{S}}^c(\boldsymbol{w}+\boldsymbol{\epsilon}) + \frac{1}{2\sqrt{n-1}} \cdot \frac{||\boldsymbol{w}||_2}{\rho_c^*}, \approx (2||\boldsymbol{w}||_2 ||\nabla_w L_{\mathcal{S}}^c(\boldsymbol{w})||_2)^{\frac{1}{2}} k^{\frac{1}{4}} (n_c-1)^{-\frac{1}{4}} \tag{7}$$

$$\rho_c^* = \left(\frac{||\boldsymbol{w}||_2}{2||\nabla_w L_{\mathcal{S}}^c(\boldsymbol{w})||_2} \right)^{\frac{1}{2}} k^{-\frac{1}{4}} (n_c-1)^{-\frac{1}{4}}, \tag{8}$$

$$\hat{\boldsymbol{\epsilon}}_c^*(\boldsymbol{w}) \approx \sqrt{k}\rho_c^* \frac{\nabla_w L_{\mathcal{S}}^c(\boldsymbol{w})}{||\nabla_w L_{\mathcal{S}}^c(\boldsymbol{w})||_2}. \tag{9}$$

An immediate observation of Eqn 8 is that the characteristic radius ρ_c^* is class-dependent, more specifically, negatively correlated with the label frequency n_c. One possible interpretation is that the model is more confident regarding the majority classes with higher n_c, due to non-asymptotic estimation error given limited samples, hence it requires a smaller perturbative region in the parameter space to ensure flat loss landscape with respect to that class. Additionally, the class-wise generalization bound $\widehat{L}_{\mathcal{T}}^c(\boldsymbol{w})$ in Eqn 7 is positively related to the class-wise gradient norm $||\nabla_{\boldsymbol{w}} L_{\mathcal{S}}^c(\boldsymbol{w})||_2$, which effectively enforces hard example mining.

It follows that optimizing the approximated generalization bound $\widehat{L}_{\mathcal{T}}(\boldsymbol{w}) = \sum_{c=1}^{k} \widehat{L}_{\mathcal{T}}^c(\boldsymbol{w})$ yields an effective algorithm in the first stage, namely **C**lass-**C**onditional

Algorithm 1: Training Paradigm of CC-SAM

Input: Training Dataset
$\mathcal{S} \sim p_s(\boldsymbol{x}, \boldsymbol{y}) = p_s(\boldsymbol{x}|\boldsymbol{y})p_s(\boldsymbol{y})$
Output: Model trained with CC-SAM

Stage 1:
Initialize $\boldsymbol{w} = \{\boldsymbol{\varphi}, \boldsymbol{\theta}\}$ randomly
while not converged **do**
  foreach *batch* $\mathcal{B}_i$ *in* $\mathcal{S}$ **do**
    Compute empirical loss $L_{\mathcal{S}}$ with $\mathcal{B}_i$
    Estimate the class-specific gradient set $G = \{\boldsymbol{g_1}, \boldsymbol{g_2}, ..., \boldsymbol{g_k}\}$ with respect to $L_{\mathcal{S}}$
    Perturb $\boldsymbol{w}$ with G according to Eqn 8 and Eqn 9
    Update $\boldsymbol{w}$ via Eqn 10 and Eqn 11

Stage 2:
Freeze $\boldsymbol{\varphi}$
while not converged **do**
  Sample batches via class-balanced sampler $\mathcal{S}' = \{\mathcal{B}'_1, \mathcal{B}'_2, ..., \mathcal{B}'_m\} \sim p_s(\boldsymbol{x}|\boldsymbol{y}) \cdot \text{Uniform}(\boldsymbol{y})$
  foreach *batch* $\mathcal{B}'_i$ *in* $\mathcal{S}'$ **do**
    Computing empirical loss $L_{\mathcal{S}}$ with $\mathcal{B}'_i$
    Obtain the adversarial feature via Eqn 12
    Evaluate overall loss according to Eqn 13 and Eqn 14
    Update $\boldsymbol{\theta}$ accordingly

Sharpness-**A**ware **M**inimization (CC-SAM):

Given learning rate η,

$$\boldsymbol{w} \leftarrow \boldsymbol{w} - \eta \nabla_{\boldsymbol{w}} L_{\mathcal{S}}^{\text{CC-SAM}}(\boldsymbol{w}) \tag{10}$$

$$\approx \boldsymbol{w} - \eta \sum_{c=1}^{k} \nabla_{\boldsymbol{w}} \widehat{L}_{\mathcal{T}}^{c}(\boldsymbol{w})|_{\boldsymbol{w}+\hat{\boldsymbol{\epsilon}}_c^*(\boldsymbol{w})}, \tag{11}$$

which is computationally efficient since it only involves first-order gradients. Even when $\boldsymbol{w}$ is close to the optimal $\boldsymbol{w}^*$, where $\nabla_{\boldsymbol{w}} L_{\mathcal{S}}(\boldsymbol{w})|_{\boldsymbol{w}^*} \approx 0$, the gradient for each class $\nabla_{\boldsymbol{w}} L_{\mathcal{S}}^{c}(\boldsymbol{w})|_{\boldsymbol{w}^*}$ for estimating the optimal perturbation vector $\hat{\boldsymbol{\epsilon}}_c^*(\boldsymbol{w})$ in Eqn 9 is most likely far from zero, circumventing the need for computing high-order terms.

3.2. Stage 2: Robust Training of the Classifier

For the second stage, we freeze the backbone to maintain feature representation and concentrate on refining the decision boundary of the classifier. Following Decoupling, we adopt the class-balanced sampling strategy in this stage, to rectify the classifier with uniform label distribution. As depicted in the Figure 3(b), when taking original feature $\boldsymbol{z}$ as the input of classifier $h(\boldsymbol{z}; \boldsymbol{\theta})$, we can generate adversarial features in forward pass as follows:

$$\boldsymbol{z_{adv}} = \boldsymbol{z} + \lambda \frac{\nabla_{\boldsymbol{z}} L_{\mathcal{S}}(\boldsymbol{z}; \boldsymbol{\theta})}{||\nabla_{\boldsymbol{z}} L_{\mathcal{S}}(\boldsymbol{z}; \boldsymbol{\theta})||_2} \tag{12}$$

$$L_{\mathcal{A}} = h(\boldsymbol{z_{adv}}; \boldsymbol{\theta}) \tag{13}$$

where λ is the hyper-parameter to scale the adversarial gradient $\nabla_{\boldsymbol{z}} L_{\mathcal{S}}(\boldsymbol{z}; \boldsymbol{\theta}) / ||\nabla_{\boldsymbol{z}} L_{\mathcal{S}}(\boldsymbol{z}; \boldsymbol{\theta})||_2$. Moreover, we employ a progressive strategy to balance $L_{\mathcal{S}}$ and $L_{\mathcal{A}}$. At epoch t,

$$L = (1 - \frac{t}{T}) L_{\mathcal{S}} + \frac{t}{T} L_{\mathcal{A}} \tag{14}$$

where T is the total number of epochs in this stage, and $L_{\mathcal{A}}$ dominates the training progressively.

4. Evaluation

In this section, we evaluate our method on multiple mainstream DLTR datasets and compare it with popular baselines including the state-of-the-art (SOTA) methods. Moreover, we also report the result on open-set recognition [32, 47] tasks to demonstrate the excellent robustness of CC-SAM to out-of-distribution samples. In the end, an ablation study is presented to verify the effectiveness of our design choices.

4.1. Datasets and Baselines

Following the mainstream evaluation protocol [7, 63], we conduct experiments on five major long-tailed datasets, CIFAR-10-LT, CIFAR-100-LT, Places-LT [32], ImageNet-LT [32], and iNaturalist 2018 [46].

CIFAR-10-LT/CIFAR-100-LT: These two datasets are sampled from the original CIFAR with different imbalance ratios $\beta = N_{max}/N_{min}$, where N_{max} and N_{min} are the corresponding number of the most and least frequent classes. Following [26], we set the imbalance ratio as $\{200, 100, 50\}$ for evaluation.

Places-LT & ImageNet-LT: Both datasets were first proposed by OLTR [32]. Places-LT contains 62.5K training images spanning 365 classes in total, with imbalance ratio 996. ImageNet-LT has 115.8K training images covering 1000 categories, with imbalance ratio being 256.

iNaturalist 2018: As a naturally long-tailed classification dataset, iNaturalist 2018 contains 437.5K trainin images from 8142 categories, and its imbalance ratio is 512. We follow the official spilt in our evaluations.

For fair comparison, we exclude ensemble or pre-training models in our experiments. The baselines and state-of-the-art methods evaluated include (1) class re-balancing: LDAM-DRW [7], LDAM-DRW + SAM [38], De-confound-TDE [45], Lifted Loss [36], Focal Loss [28], OpenMax [3], BBN [63], LADE [17], DisAlign [58],

	CIFAR-10-LT			CIFAR-100-LT		
Imbalance Ratio	200	100	50	200	100	50
CE	65.68	70.70	74.81	34.84	38.43	43.90
CE + Mixup [57]	65.84	72.96	79.48	35.84	40.01	45.16
LDAM-DRW [7]	73.52	77.03	81.03	38.91	42.04	47.62
De-confound-TDE [45]	-	80.60	83.60	-	44.15	50.31
CE + Mixup + cRT [20]	73.06	79.15	84.21	41.73	45.12	50.86
BBN [63]	73.47	79.82	81.18	37.21	42.56	47.02
Contrastive Learning [51]	-	81.40	85.36	-	46.72	51.87
BGP [49]	-	-	-	41.20	45.20	50.50
MiSLAS [62]	77.31	82.06	85.16	42.33	47.50	52.62
VS + SAM [38]	-	82.40	-	-	46.60	-
GCL [26]	79.03	82.68	85.46	44.88	48.71	53.55
CC-SAM	**80.94**	**83.92**	**86.22**	**45.66**	**50.83**	**53.91**

Table 1. Top-1 Accuracy on CIFAR-10-LT and CIFAR-100-LT. All methods take ResNet-32 as the backbone. All baseline results except BGP are directly adopted from [26]. "-" means the original paper didn't report the corresponding results.

cRT	Stage 1 + dir	Stage 1 + mag	Stage 2	Acc
✔				37.8
✔	✔			38.9
✔		✔		37.5
✔	✔	✔		40.1
✔	✔	✔	✔	**40.6**

Table 2. Ablation studies on Places-LT. "Stage 1 + dir" enforces parameter perturbation along the recommended direction ($\nabla_w L_S^c(\boldsymbol{w})/||\nabla_w L_S^c(\boldsymbol{w})||_2$ in Eqn 9) with magnitude of 1, whereas "Stage 1 + mag" enforces perturbation with the recommended magnitude (ρ_c^* in Eqn 9) in a random direction.

LUNA [5], BGP [49], VS+SAM [38] (2) information augmentation: RSG [50], (3) decoupled training: Decouple-τ-norm [20], cRT [20], MisLAS [62], GCL [26], (4) representation learning: Range Loss [59], OLTR [32], IEM [65], Contrastive Learning [51], ResLT [9].

Our code is implemented with Pytorch 1.4.0 and all experiments are carried out on Tesla V100 GPUs. We train each model with batch size of 64 (for CIFAR-10-LT and CIFAR-100-LT) / 128 (for Places-LT) / 256 (for ImageNet-LT) / 512 (for iNaturalist 2018), SGD optimizer with momentum of 0.9. We apply similar tricks in Balanced Softmax [40] as the mainstream methods have done. Although we implemented CC-SAM under a two-stage framework, it serves as a general technique to improve the existing DLTR methods without model perturbation (**Appendix A.3**).

Intuitively, CC-SAM brings more computation overhead due to additional gradient descents. But we only perturbed the last several layers as an efficient version of CC-SAM in our evaluation. A simple training time comparison is presented in **Appendix A.5**.

4.2. Main Results

We evaluate CC-SAM on the five mainstream public benchmarks mentioned above, and the corresponding results are reported in Table 1 and Table 3. The best method is bolded and the second best is underlined. According to the evaluations, we observe that CC-SAM unanimously attains the best ranking on CIFAR-LT datasets. On large scale datasets (Places-LT, ImageNet-LT, and iNaturalist 2018), CC-SAM shows competitive performance compared to other advanced methods, too. Specifically, it consistently brings better gains to the medium and tail classes, demonstrating the effectiveness of class-conditional flattening for deep long-tailed recognition.

4.3. Open Long-Tailed Recognition

The ability to detect out-of-distribution samples from the open world, namely open-set recognition [47], provides a unique dimension for evaluating the robustness of deep learning models. Open-set long-tailed recognition, or OLTR [32], combines the open-set problems with deep long-tailed learning to offer even more challenging tasks for model generalization. Following the setting of OLTR [32], we enable CC-SAM to distinguish the close-set and open-set samples by applying a simple, non-learnable prototype-based metric. Results are presented in Table 4. We find that CC-SAM achieves the top F-measure performance and outperforms LUNA [5] on Places-LT, a SOTA with a sophisticated open-set detection method based on hierarchical metrics. The experiments demonstrate the remarkable gen-

Dataset	Method	Backbone	Many	Medium	Few	Overall
ImageNet-LT	CE	ResNeXt-50	**65.9**	37.5	7.7	44.4
	Decouple-τ-norm [20]	ResNet-50	56.6	44.2	27.4	46.7
	Balanced Softmax [40]	ResNeXt-50	64.1	48.2	33.4	52.3
	LADE [17]	ResNeXt-50	64.4	47.7	34.3	52.3
	RSG [50]	ResNeXt-50	63.2	48.2	32.2	51.8
	DisAlign [58]	ResNet-50	61.3	52.2	31.4	52.9
		ResNeXt-50	62.7	52.1	31.4	53.4
	ResLT [9]	ResNeXt-50	63.0	53.3	35.5	52.9
	BGP [49]	ResNet-50	-	-	-	51.5
	MiSLAS [62]	ResNet-50	-	-	-	52.7
	LDAM-DRW + SAM [38]	ResNet-50	62.0	52.1	34.8	53.1
	GCL [26]	ResNet-50	-	-	-	54.9
	CC-SAM	ResNet-50	61.4	49.5	37.1	52.4
		ResNeXt-50	63.1	**53.4**	**41.1**	**55.4**
Places-LT	CE	ResNet-152	**45.7**	27.3	8.2	30.2
	Decouple-τ-norm [20]	ResNet-152	37.8	40.7	31.8	37.9
	Balanced Softmax [40]	ResNet-152	42.0	39.3	30.5	38.6
	LADE [17]	ResNet-152	42.8	39.0	31.2	38.8
	RSG [50]	ResNet-152	41.9	41.4	32.0	39.3
	DisAlign [58]	ResNet-152	40.4	42.4	30.1	39.3
	ResLT [9]	ResNet-152	39.8	**43.6**	31.4	39.8
	MiSLAS [62]	ResNet-152	-	-	-	40.2
	GCL [26]	ResNet-152	-	-	-	40.6
	CC-SAM	ResNet-152	41.2	42.1	**36.4**	**40.6**
iNaturalist 2018	CE	ResNet-50	**72.2**	63.0	57.2	61.7
	Decouple-τ-norm [20]	ResNet-50	65.6	65.3	65.9	65.6
	Balanced Softmax [40]	ResNet-50	-	-	-	70.6
	LADE [17]	ResNet-50	-	-	-	70.0
	RSG [50]	ResNet-50	-	-	-	70.3
	DisAlign [58]	ResNet-50	-	-	-	70.6
	ResLT [9]	ResNet-50	-	-	-	70.2
	BGP [49]	ResNet-50	70.0	69.9	69.6	70.5
	MiSLAS [62]	ResNet-50	-	-	-	71.6
	LDAM-DRW + SAM [38]	ResNet-50	64.1	70.5	71.2	70.1
	GCL [26]	ResNet-50	-	-	-	**72.0**
	CC-SAM	ResNet-50	65.4	**70.9**	**72.2**	70.9

Table 3. Top-1 Accuracy on Places-LT, ImageNet-LT and iNaturalist 2018. As for Places-LT, we take a pre-trained ResNet-152 as the backbone for a fair comparison.

eralization of CC-SAM due to its capacity to learn highly robust representations.

4.4. Ablation Study

Since CC-SAM consists of multiple components, here we provide an ablation study to demonstrate their effectiveness. The result of experiments conducted on Places-LT are shown in Table 2. In particular, CC-SAM degrades to cRT without any of our proposed operations, which we take as a baseline. It is observed that CC-SAM benefits from the design choices of each individual stage, and the integration of both stages shows the best performance.

We also study the class-conditional perturbation by differentiating the impacts of choosing the right direction $\nabla_{\boldsymbol{w}} L_{\mathcal{S}}^{c}(\boldsymbol{w}) / ||\nabla_{\boldsymbol{w}} L_{\mathcal{S}}^{c}(\boldsymbol{w})||_2$ and the right magnitude ρ_c^* in Eqn 9, shown as "stage 1 + dir" and "stage 1 + mag" in Table 2 respectively. For the former, we choose a perturbation scale of 1. For the latter, we perturb the model parameters by a random direction sampled from a zero-mean Gaussian. It turns out the perturbation direction contributes more, and

	ImageNet-LT				Places-LT			
Method	Many	Medium	Few	F-measure	Many	Medium	Few	F-measure
CE	40.1	10.4	0.4	0.295	45.9	22.4	0.4	0.366
Lifted Loss [36]	34.8	29.3	17.4	0.374	41.0	35.2	23.8	0.459
Focal Loss [28]	35.7	29.3	15.6	0.371	41.0	34.8	22.3	0.453
Range Loss [59]	34.7	29.4	17.2	0.373	41.0	35.3	23.1	0.457
OpenMax [3]	35.8	30.0	17.6	0.368	41.1	35.4	23.2	0.458
OLTR [32]	41.9	33.9	17.4	0.474	44.6	36.8	25.2	0.464
IEM [65]	46.1	42.3	20.1	0.525	48.8	42.4	28.9	0.486
LUNA [5]	48.2	44.7	23.6	**0.579**	48.1	41.6	29.0	0.491
CC-SAM	61.4	49.5	37.1	0.552	41.2	41.8	36.4	**0.510**

Table 4. Open long-tail performance of top-1 accuracy on ImageNet-LT and Places-LT. F-measure is a balanced treatment of precision and recall. The backbone of CC-SAM is ResNet-50 for ImageNet-LT, while other compared methods are equipped with ResNet-10.

applying both attains the best improvement. For more ablation studies, please refer to **Appendix A.1**.

5. Related Work and Discussion

5.1. Deep Long-Tailed Recognition

A latest survey [60] systematically studies up-to-date DLTR algorithms published at the top-tier conferences and introduces a new taxonomy, dividing DLTR methodologies into three categories: class re-balancing, information augmentation, and module improvement. For class re-balancing, a branch of sampling strategies [20, 40, 52, 54] have been proposed to re-balance the training distribution. Meanwhile, some studies design dedicated loss functions [7, 10, 28, 45] to dynamically re-weight the gradients for optimization. As one representative of information augmentation, RSG [50] enriches the feature space of tail classes via dynamical prototypes and a maximized vector loss. Decoupled training [20, 58, 62] propose to learn effective representation and robust classifier according to different sampling strategies in two stages.

Our method follows the decoupled training paradigm. Moreover, to combat label distribution shift, our method seeks out flat minima by perturbing the model parameters at a class-conditional scale, which is unseen in existing DLTR literature.

5.2. Flattening for Deep Learning Models

Sharpness as a generalization measure of deep learning models has been extensively explored in previous literature [1, 13, 16, 19]. Keskar *et al*. [21] empirically found that a large batch SGD optimizer might lead to sharp minima and poor generalization. Inspired by this, He *et al*. [16] implicitly averaged along the SGD trajectory to find the asymmetric valleys, namely asymmetric flat minima. Due to the power of flattening, it has also been applied to many learning tasks to achieve better generalization. Rangwani *et al*. [37] introduced a domain adversarial training framework by adding an auxiliary loss to smooth the local minima for domain adaptation. Shi *et al*. [44] demonstrated that a flattened model was resilient to catastrophic forgetting in class-incremental few-shot learning.

In this work, we investigate flattening in the context of DLTR. To our best knowledge, there are only a few works which succeed in applying flattening to improve generalization in DLTR/OLTR. GBP [49] is built upon the observation of high correlation between logits and gradient norm, and proposes a gradient penalty loss as regularization. However, it lacks rigorous theoretical interpretation by employing a very loose (vacuous) PAC-Bayesian bound and meanwhile under-performs compared to our approach (Table 1, 3). Liu *et al*. [30] develops a variant of SAM that re-weights the perturbation at instance level to combat data imbalance, without specifying an optimal perturbation or giving any theoretical justification; thus, we regard it as a simplistic and incomplete version of our method. Besides, Rangwani *et al*. [38] observes that re-weighting techniques in DLTR lead to saddle points for tail classes, and directly leverages SAM to alleviate this problem. In contrast, CC-SAM further advances [38] by providing a theoretically motivated class-conditional SAM design for better generalization and robustness to label distribution shift.

Acknowledgment We thank anonymous reviewers for their valuable comments. This work was supported by the NSFC under Grants 61932017 and 61971390, and the following grants from the Research Grants Council of the Hong Kong Special Administrative Region, China (Project Reference Number: 14201620 and 14201321). Lanqing Li would like to thank Guangyong Chen and Danruo Deng for their helpful comments on the PAC-Bayesian framework.

References

[1] Carlo Baldassi, Fabrizio Pittorino, and Riccardo Zecchina. Shaping the learning landscape in neural networks around wide flat minima. *Proceedings of the National Academy of Sciences*, 117(1):161–170, 2020. 2, 8

[2] Albert-László Barabási and Réka Albert. Emergence of scaling in random networks. *science*, 286(5439):509–512, 1999. 1

[3] Abhijit Bendale and Terrance E Boult. Towards open set deep networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1563–1572, 2016. 5, 8

[4] Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale gan training for high fidelity natural image synthesis. *arXiv preprint arXiv:1809.11096*, 2018. 3

[5] Jiarui Cai, Yizhou Wang, Hung-Min Hsu, Jenq-Neng Hwang, Kelsey Magrane, and Craig S Rose. Luna: Localizing unfamiliarity near acquaintance for open-set long-tailed recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 131–139, 2022. 6, 8

[6] Jiarui Cai, Yizhou Wang, and Jenq-Neng Hwang. Ace: Ally complementary experts for solving long-tailed recognition in one-shot. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 112–121, 2021. 1

[7] Kaidi Cao, Colin Wei, Adrien Gaidon, Nikos Arechiga, and Tengyu Ma. Learning imbalanced datasets with label-distribution-aware margin loss. *Advances in neural information processing systems*, 32, 2019. 1, 2, 5, 6, 8

[8] Niladri S. Chatterji, Behnam Neyshabur, and Hanie Sedghi. The intriguing role of module criticality in the generalization of deep networks. In *8th International Conference on Learning Representations*, 2020. 3

[9] Jiequan Cui, Shu Liu, Zhuotao Tian, Zhisheng Zhong, and Jiaya Jia. Reslt: Residual learning for long-tailed recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022. 1, 6, 7

[10] Yin Cui, Menglin Jia, Tsung-Yi Lin, Yang Song, and Serge Belongie. Class-balanced loss based on effective number of samples. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9268–9277, 2019. 1, 8

[11] Danruo Deng, Guangyong Chen, Jianye Hao, Qiong Wang, and Pheng-Ann Heng. Flattening sharpness for dynamic gradient projection memory benefits continual learning. *Advances in Neural Information Processing Systems*, 34:18710–18721, 2021. 1

[12] Gintare Karolina Dziugaite and Daniel M. Roy. Computing nonvacuous generalization bounds for deep (stochastic) neural networks with many more parameters than training data. In *Proceedings of the Thirty-Third Conference on Uncertainty in Artificial Intelligence*, 2017. 3

[13] Pierre Foret, Ariel Kleiner, Hossein Mobahi, and Behnam Neyshabur. Sharpness-aware minimization for efficiently improving generalization. In *9th International Conference on Learning Representations*, 2021. 1, 2, 3, 4, 8

[14] Ishaan Gulrajani and David Lopez-Paz. In search of lost domain generalization. In *9th International Conference on Learning Representations*, 2021. 1

[15] Haibo He and Edwardo A Garcia. Learning from imbalanced data. *IEEE Transactions on knowledge and data engineering*, 21(9):1263–1284, 2009. 1

[16] Haowei He, Gao Huang, and Yang Yuan. Asymmetric valleys: Beyond sharp and flat local minima. *Advances in neural information processing systems*, 32, 2019. 2, 3, 8

[17] Youngkyu Hong, Seungju Han, Kwanghee Choi, Seokjun Seo, Beomsu Kim, and Buru Chang. Disentangling label distribution for long-tailed visual recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6626–6636, 2021. 1, 2, 5, 7

[18] Yuanfeng Ji, Lu Zhang, Jiaxiang Wu, Bingzhe Wu, Long-Kai Huang, Tingyang Xu, Yu Rong, Lanqing Li, Jie Ren, Ding Xue, et al. Drugood: Out-of-distribution (ood) dataset curator and benchmark for ai-aided drug discovery–a focus on affinity prediction problems with noise annotations. *arXiv preprint arXiv:2201.09637*, 2022. 1

[19] Yiding Jiang, Behnam Neyshabur, Hossein Mobahi, Dilip Krishnan, and Samy Bengio. Fantastic generalization measures and where to find them. In *8th International Conference on Learning Representations*, 2020. 2, 3, 8

[20] Bingyi Kang, Saining Xie, Marcus Rohrbach, Zhicheng Yan, Albert Gordo, Jiashi Feng, and Yannis Kalantidis. Decoupling representation and classifier for long-tailed recognition. *arXiv preprint arXiv:1910.09217*, 2019. 1, 2, 4, 6, 7, 8

[21] Nitish Shirish Keskar, Dheevatsa Mudigere, Jorge Nocedal, Mikhail Smelyanskiy, and Ping Tak Peter Tang. On large-batch training for deep learning: Generalization gap and sharp minima. In *5th International Conference on Learning Representations*, 2017. 1, 2, 8

[22] Jaehyung Kim, Jongheon Jeong, and Jinwoo Shin. M2m: Imbalanced classification via major-to-minor translation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 13896–13905, 2020. 1

[23] Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanas Phillips, Irena Gao, et al. Wilds: A benchmark of in-the-wild distribution shifts. In *International Conference on Machine Learning*, pages 5637–5664. PMLR, 2021. 1

[24] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems 25: 26th Annual Conference on Neural Information Processing Systems*, pages 1106–1114, 2012. 1

[25] Lanqing Li, Liang Zeng, Ziqi Gao, Shen Yuan, Yatao Bian, Bingzhe Wu, Hengtong Zhang, Chan Lu, Yang Yu, Wei Liu, et al. Imdrug: A benchmark for deep imbalanced learning in ai-aided drug discovery. *arXiv preprint arXiv:2209.07921*, 2022. 1

[26] Mengke Li, Yiu-ming Cheung, and Yang Lu. Long-tailed visual recognition via gaussian clouded logit adjustment. In *Proceedings of the IEEE/CVF Conference on Computer Vi-*

sion and Pattern Recognition, pages 6929–6938, 2022. 2, 5, 6, 7

[27] Shuang Li, Kaixiong Gong, Chi Harold Liu, Yulin Wang, Feng Qiao, and Xinjing Cheng. Metasaug: Meta semantic augmentation for long-tailed visual recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5212–5221, 2021. 1

[28] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017. 1, 5, 8

[29] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014. 1

[30] Hong Liu, Jeff Z HaoChen, Adrien Gaidon, and Tengyu Ma. Self-supervised learning is more robust to dataset imbalance. In *International Conference on Learning Representations*, 2021. 1, 8

[31] 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. 1

[32] 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/CVF Conference on Computer Vision and Pattern Recognition*, pages 2537–2546, 2019. 1, 2, 5, 6, 8

[33] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3431–3440, 2015. 1

[34] David A McAllester. Pac-bayesian model averaging. In *Proceedings of the twelfth annual conference on Computational learning theory*, pages 164–170, 1999. 3

[35] Behnam Neyshabur, Srinadh Bhojanapalli, David McAllester, and Nati Srebro. Exploring generalization in deep learning. *Advances in neural information processing systems*, 30, 2017. 1, 3

[36] Hyun Oh Song, Yu Xiang, Stefanie Jegelka, and Silvio Savarese. Deep metric learning via lifted structured feature embedding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4004–4012, 2016. 5, 8

[37] Harsh Rangwani, Sumukh K Aithal, Mayank Mishra, Arihant Jain, and Venkatesh Babu Radhakrishnan. A closer look at smoothness in domain adversarial training. In *International Conference on Machine Learning*, pages 18378–18399. PMLR, 2022. 8

[38] Harsh Rangwani, Sumukh K Aithal, Mayank Mishra, and Venkatesh Babu Radhakrishnan. Escaping saddle points for effective generalization on class-imbalanced data. In *Advances in Neural Information Processing Systems*. 5, 6, 7, 8

[39] William J Reed. The pareto, zipf and other power laws. *Economics letters*, 74(1):15–19, 2001. 1

[40] Jiawei Ren, Cunjun Yu, Xiao Ma, Haiyu Zhao, Shuai Yi, et al. Balanced meta-softmax for long-tailed visual recognition. *Advances in neural information processing systems*, 33:4175–4186, 2020. 1, 6, 7, 8

[41] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28, 2015. 1

[42] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3):211–252, 2015. 1

[43] Ikuro Sato, Yamada Ryota, Masayuki Tanaka, Nakamasa Inoue, and Rei Kawakami. Pof: Post-training of feature extractor for improving generalization. In *International Conference on Machine Learning*, pages 19221–19230. PMLR, 2022. 1

[44] Guangyuan Shi, Jiaxin Chen, Wenlong Zhang, Li-Ming Zhan, and Xiao-Ming Wu. Overcoming catastrophic forgetting in incremental few-shot learning by finding flat minima. *Advances in Neural Information Processing Systems*, 34:6747–6761, 2021. 1, 3, 8

[45] Jingru Tan, Changbao Wang, Buyu Li, Quanquan Li, Wanli Ouyang, Changqing Yin, and Junjie Yan. Equalization loss for long-tailed object recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11662–11671, 2020. 1, 5, 6, 8

[46] 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. 5

[47] Sagar Vaze, Kai Han, Andrea Vedaldi, and Andrew Zisserman. Open-set recognition: A good closed-set classifier is all you need. *arXiv preprint arXiv:2110.06207*, 2021. 5, 6

[48] Athanasios Voulodimos, Nikolaos Doulamis, Anastasios Doulamis, and Eftychios Protopapadakis. Deep learning for computer vision: A brief review. *Computational intelligence and neuroscience*, 2018. 1

[49] Dong Wang, Yicheng Liu, Liangji Fang, Fanhua Shang, Yuanyuan Liu, and Hongying Liu. Balanced gradient penalty improves deep long-tailed learning. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 5093–5101, 2022. 2, 6, 7, 8

[50] Jianfeng Wang, Thomas Lukasiewicz, Xiaolin Hu, Jianfei Cai, and Zhenghua Xu. Rsg: A simple but effective module for learning imbalanced datasets. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3784–3793, 2021. 1, 6, 7, 8

[51] Peng Wang, Kai Han, Xiu-Shen Wei, Lei Zhang, and Lei Wang. Contrastive learning based hybrid networks for long-tailed image classification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 943–952, 2021. 1, 6

[52] 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 *European conference on computer vision*, pages 728–744. Springer, 2020. 1, 8

[53] Xudong Wang, Long Lian, Zhongqi Miao, Ziwei Liu, and Stella Yu. Long-tailed recognition by routing diverse distribution-aware experts. In *International Conference on Learning Representations*, 2020. 1

[54] Yiru Wang, Weihao Gan, Jie Yang, Wei Wu, and Junjie Yan. Dynamic curriculum learning for imbalanced data classification. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 5017–5026, 2019. 1, 8

[55] Xi Yin, Xiang Yu, Kihyuk Sohn, Xiaoming Liu, and Manmohan Chandraker. Feature transfer learning for face recognition with under-represented data. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5704–5713, 2019. 1

[56] Liang Zeng, Lanqing Li, Ziqi Gao, Peilin Zhao, and Jian Li. Imgcl: Revisiting graph contrastive learning on imbalanced node classification. *arXiv preprint arXiv:2205.11332*, 2022. 1

[57] Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412*, 2017. 6

[58] Songyang Zhang, Zeming Li, Shipeng Yan, Xuming He, and Jian Sun. Distribution alignment: A unified framework for long-tail visual recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2361–2370, 2021. 1, 5, 7, 8

[59] Xiao Zhang, Zhiyuan Fang, Yandong Wen, Zhifeng Li, and Yu Qiao. Range loss for deep face recognition with long-tailed training data. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 5409–5418, 2017. 1, 6, 8

[60] Yifan Zhang, Bingyi Kang, Bryan Hooi, Shuicheng Yan, and Jiashi Feng. Deep long-tailed learning: A survey. *arXiv preprint arXiv:2110.04596*, 2021. 1, 8

[61] Yang Zhao, Hao Zhang, and Xiuyuan Hu. Penalizing gradient norm for efficiently improving generalization in deep learning. *arXiv preprint arXiv:2202.03599*, 2022. 3

[62] Zhisheng Zhong, Jiequan Cui, Shu Liu, and Jiaya Jia. Improving calibration for long-tailed recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16489–16498, 2021. 1, 2, 6, 7, 8

[63] 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/CVF conference on computer vision and pattern recognition*, pages 9719–9728, 2020. 1, 5, 6

[64] Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10 million image database for scene recognition. *IEEE transactions on pattern analysis and machine intelligence*, 40(6):1452–1464, 2017. 1

[65] Linchao Zhu and Yi Yang. Inflated episodic memory with region self-attention for long-tailed visual recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4344–4353, 2020. 1, 6, 8