Revive Re-weighting in Imbalanced Learning by Density Ratio Estimation

Jiaan Luo[1,3†] **Feng Hong**[1†] **Jiangchao Yao**[1,3‡] **Bo Han**[4] **Ya Zhang**[2,3] **Yanfeng Wang**[2,3]
[1]Cooperative Medianet Innovation Center, Shanghai Jiao Tong University
[2]School of Artificial Intelligence, Shanghai Jiao Tong University
[3]Shanghai Artificial Intelligence Laboratory
[4]Hong Kong Baptist University
{luojiaan, feng.hong, Sunarker, ya_zhang, wangyanfeng}@sjtu.edu.cn
bhanml@comp.hkbu.edu.hk

Abstract

In deep learning, model performance often deteriorates when trained on highly imbalanced datasets, especially when evaluation metrics require robust generalization across underrepresented classes. To address the challenges posed by imbalanced data distributions, this study introduces a novel method utilizing density ratio estimation for dynamic class weight adjustment, termed as Re-weighting with Density Ratio (RDR). Our method adaptively adjusts the importance of each class during training, mitigates overfitting on dominant classes and enhances model adaptability across diverse datasets. Extensive experiments conducted on various large scale benchmark datasets validate the effectiveness of our method. Results demonstrate substantial improvements in generalization capabilities, particularly under severely imbalanced conditions. The code is available here.

1 Introduction

In recent years, deep learning has made significant strides across various domains by utilizing complex architectures and large-scale datasets, setting new benchmarks for performance. However, these advancements often rely on well-curated datasets that ensure balanced class distributions [Russakovsky et al., 2015]. In contrast, real-world datasets typically exhibit a long-tailed distribution, where few classes dominate the majority of samples, while many others are underrepresented [Krizhevsky et al., 2009]. This imbalance leads to model biases favoring frequent classes, thereby reducing performance on the less common ones. Yet, in many applications—such as medical diagnostics and financial analysis—greater emphasis is placed on ensuring strong generalization for underrepresented classes. Addressing this challenge not only reduces data collection costs but also improves the robustness and fairness of the models.

Many excellent methods, such as re-sampling [Bowyer et al., 2011], re-weighting [Morik et al., 1999], decoupled learning [Kang et al., 2020], margin-based learning [Cao et al., 2019, Menon et al., 2020], transfer learning [Yin et al., 2019] and contrastive learning [Tian et al., 2021], have been proposed to tackle the issue of imbalanced data. Despite the simplicity of re-weighting, it falls behind in performance significantly compared with other directions of methods due to the inappropriate weighting coefficients during training. Cui et al. [2019] proposes a method for re-weighting by effective number, which accounts for potential overlaps among data samples and adjusts the weights for each category based on the actual effective number of samples. Chen et al. [2023b] leverages the effective area to re-weight, considering the actual spanned space of each class. However, such subsequent improvements can alleviate but still cannot effectively push that forward. Wang et al.

[†]The first two authors contribute equally.
[‡]The corresponding author is Jiangchao Yao (Sunarker@sjtu.edu.cn).

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

https://doi.org/10.52202/079017-2539

[2023] obtains a fine-grained generalization bound for re-weighting in imbalanced learning through the data-dependent contraction technique. Limited research has focused on the intrinsic limitations of the commonly employed re-weighting-based loss functions and the corresponding balancing mechanisms designed to enhance parity in class representation.

This study rethinks the characteristics of re-weighted loss and explores the question *"Why is re-weighting necessary under conditions of sample imbalance?"* Under conditions of sample imbalance, the variation in weights of samples arises due to discrepancies between the distribution of collected data and a balanced data distribution. In scenarios where class balance exists, such discrepancies are absent, thus obviating the need for re-weighting. Conversely, in imbalanced settings, re-weighting becomes essential to bridge the gap between these distributions. The weights must therefore represent a suitable compromise between balanced and imbalanced distributions and necessarily reflect accurately on each sample. Additionally, as model training progresses dynamically, optimizing the fit to feature distributions, the weights applied to each sample should be continuously updated to maintain robust performance.

This research introduces a novel method, Reweighting with Density Ratio (RDR), designed to mitigate learning disparities in imbalanced distributions. In this method, a feature extractor is employed to discern the features from the training data. A more balanced feature distribution is approximated by continuously updating the momentum on the feature level. This enables real-time density ratio estimation with features learned under imbalanced distributions, thereby obtaining the sample-wise weights. Notably, as the learned features evolve, our method dynamically adjusts weights in response to observed shifts in class density throughout the training cycle, ensuring that the model remains adaptive and effective. This method significantly enhances the robustness and adaptability of the training process. By integrating density ratio estimation to evaluate the difference between the balanced and real data distributions, our approach more accurately reflects the underlying class distribution and improves the model's generalization capabilities across diverse datasets. The contributions are summarized as follows:

- We explore the existing re-weighting techniques, and model the performance of various algorithms during training under different data distributions. This approach offers a novel perspective on understanding re-weighting methods in the scenarios of sample imbalance.
- We introduce a novel methodology, Re-weighting with Density Ratio (RDR), which leverages the method of density ratio estimation to dynamically adjust class weights during model training. This approach not only addresses the limitations of prior re-weighting methods but also introduces a mechanism to continuously adapt to the changing importance of classes as learning progresses, thereby enhancing model robustness and adaptability.
- We conduct extensive experiments to validate the effectiveness of our proposed RDR method. These experiments are conducted across various large-scale, long-tailed datasets, demonstrating substantial improvements in handling class imbalance. Our results illustrate significant enhancements in generalization capabilities, particularly under severely imbalanced scenarios.

2 Related Work

2.1 Re-weighting Based Methods

Re-weighting methods for addressing class imbalance have evolved significantly over the years. Early techniques, such as [Zadrozny et al., 2003], employed inverse frequency techniques to address class imbalances but failed to consider deeper data distribution traits, leading to sub-optimal outcomes. Addressing these shortcomings, Huang et al. [2016] introduced a cost-sensitive learning framework that, beyond simple frequency adjustments, incorporated misclassification costs to achieve a more nuanced balance. However, this approach still struggled with complexities like class overlap and label noise. To further refine this approach, Lin et al. [2017] developed Focal Loss, which employs a modulation factor based on the prediction probability to adjust the loss function, thereby amplifying the impact of hard-to-classify samples while reducing the loss contribution of easy-to-classify samples. Cui et al. [2019] introduced Class-Balanced Loss, which adjusts loss by data overlap, calculating the effective number of each class. Advancements continued with methods based on training gradients, such as [Ren et al., 2020b, Wang et al., 2021a]. Chen et al. [2023b] proposed Adaptive Re-weighting via effective area, which enhances model accuracy by considering the spatial distribution and density

of data points within classes. Ma et al. [2023] introduced a re-weighting method that adjusts based on semantic richness and visual variability. However, no prior work has tackled the issue of sample imbalance by dynamically re-weighting based on the model's performance across training and test sets with differing distributions during the training process.

2.2 Non-re-weighting Based Methods

In addition to re-weighting, many other methods are available to address the issue of sample imbalance. Re-sampling techniques [Kubat and Matwin, 1997, Wallace et al., 2011, Han et al., 2005, Hong et al., 2024a] mitigate category imbalances by under-sampling dominant classes [Buda et al., 2018] or over-sampling minority classes [Bowyer et al., 2011]. However, under-sampling may degrade feature representation by discarding valuable majority class data, whereas over-sampling could cause overfitting by duplicating minority class samples. Decoupled training approaches, such as [Kang et al., 2020], challenge the traditional joint training model by separating representation learning from classification. Margin-based methods such as, LADM [Cao et al., 2019], LA [Menon et al., 2020] and VS [Kini et al., 2021], adjust training processes to increase minority class margins, therefore to obtain a more balanced decision boundary. More flexible and robust methods are proposed, including Transfer Learning [Yin et al., 2019, Liu et al., 2019], Contrastive Learning [Li et al., 2022, Chen et al., 2023a], Ensemble Learning [Wang et al., 2021b, Cai et al., 2021] and Self-supervised Learning [Liu et al., 2022, Zhou et al., 2023b]. Please refer to Appendix B for more discussions.

3 Method

3.1 Problem Setup

For a typical classification task in imbalanced learning, suppose given a training dataset $\mathcal{S} = \bigcup_{i=1}^{n}\{(\boldsymbol{x}_i, \boldsymbol{y}_i)\}$, where n is the total number of samples. Denote $\{n_1, n_2, ..., n_d\}$as the sample number of each class. We assume, without loss of generality, that $n_i < n_j$, when $i < j$, with n_d typically much larger than n_1, reflecting a pronounced imbalance in class distribution. We use a typical loss function like $l\colon \mathcal{W} \times \mathcal{X} \times \mathcal{Y} \to \mathbb{R}_+$. Denote $\{\pi_1, \pi_2, ..., \pi_d\}$ as the proportions of each class, such that $\sum_{i=1}^{d} \pi_i = 1$. Define a family of deep learning models parameterized by $\omega \in \mathcal{W} \subseteq \mathbb{R}^k$. Typically, a model consists of a feature extractor $f(x;\phi)$ and a classifier $h(z;\theta)$, with $\omega = \bigcup\{\phi, \theta\}$. The notations used in this paper are summarized in Appendix A.

3.2 Motivation

Inspired by our review of prior methods, we observe a gap in the adaptation of dynamic class-weight adjustments during training phases. Building on the groundwork of static re-weighting strategies, we introduce a novel approach utilizing the method of density ratio estimation to dynamically recalibrate class weights. This innovation aims to provide a more refined adjustment by estimating real-time class density, thereby promoting an equitable influence of all classes throughout the training.

3.3 Dynamic Re-weighting with Density Ratio

In a typical training optimization problem, our objective is to minimize the empirical risk of the loss function, *i.e.*, $\overline{R} = \frac{1}{n}\sum_{i=1}^{n} l(x_i, y_i; \omega)$. However, in imbalanced datasets, where the frequency of samples across different classes varies, it is necessary to adjust for these gaps by applying different weights for the samples. We assume that the weight of each sample is denoted by $\alpha(x, y; \omega)$, then the empirical risk can be formulated like $\overline{R} = \frac{1}{n}\sum_{i=1}^{n} \alpha(x_i, y_i; \omega) l(x_i, y_i; \omega)$.

In naive re-weighting approaches, the weight α of class y is often set to $\frac{1}{\pi_y}$. This setting is based on the assumption that the distribution of the training set P and the distribution balanced data set P_{bal} satisfy the equation $\overline{P}(x|y;\omega) = P_{bal}(x|y;\omega)$. However, in practical training scenarios, both training and test sets are subsets drawn from the actual distribution, leading to potential missing of feature patterns. Furthermore, classes with more complex features and lower sample frequencies tend to exhibit more pronounced missing of patterns. Therefore, in the training process, there exists a discrepancy between $\overline{P}(x|y;\omega)$ and $P_{bal}(x|y;\omega)$. We measure the extent of this discrepancy using the ratio $r(x|y;w) = \overline{P}(x|y;\omega)/P_{bal}(x|y;\omega)$, incorporating it as a correction term into our weighting

https://doi.org/10.52202/079017-2539

scheme. Consequently, the empirical risk can be reformulated as follows

$$\overline{R} = \frac{1}{n}\sum_{i=1}^{n}\frac{r(x_i|y_i;\omega)}{\pi_{y_i}}l(x_i,y_i;\omega) \tag{1}$$

We can explain the rationality of this formula as follows. Considering each class i, where $P_{bal}(y_i;\omega) \propto \pi_i^{-1}P(y_i;\omega)$ and $P_{bal}(x_i|y_i;\omega) = r(x_i|y_i;\omega)P(x_i|y_i;\omega)$, with conditional probability formula, we can derive:

$$\begin{aligned} R &= \mathbb{E}_P\frac{1}{\pi_i}(r(x_i|y_i;\omega)l(x_i,y_i;\omega)) = \mathbb{E}_P\frac{P_{bal}(y;\omega)}{P(y;\omega)}\frac{P_{bal}(x|y;\omega)}{P(x|y;\omega)}l(x,y;\omega) \\ &= \mathbb{E}_P\left(\frac{P_{bal}(x,y;\omega)}{P(x,y;\omega)}l(x,y;\omega)\right) = \mathbb{E}_{P_{bal}}l(x,y;\omega) \end{aligned} \tag{2}$$

Eq. (2) demonstrates that our approach aligns with the balanced risk of the loss function. Consequently, minimizing Eq. (1) also serves to minimize the balanced risk.

Let's take a closer look at $r(x|y;w) = \overline{P}(x|y;\omega)/P_{bal}(x|y;\omega)$. The variable r represents the ratio of two different distributions. We approximate this ratio using methods of density ratio estimation. This problem can be solved by first-order moment matching approach. Our goal is to minimize

$$\underset{r}{\operatorname{argmin}}\left\|\int xr(x|y;\omega)P_{bal}(x|y;\omega)\mathrm{d}x - \int xP(x|y;\omega)\mathrm{d}x\right\|^2 \tag{3}$$

where $\|\cdot\|$ denotes the Euclidean norm. Recall that we capture the features of the input samples by the feature extractor $f(x;\phi)$ in our model, and these features are a good reflection of what our model learned from the distribution of the input samples. Therefore, in order to capture more complex structures and patterns in raw data, we use $f(x;\phi)$ to obtain a variant of Eq. (3). Our goal can be achieved by obtaining $\operatorname{argmin}_r \mathrm{MM}'(r)$, where $\mathrm{MM}'(r)$ denotes

$$\left\|\int f(x;\phi)r(x|y;\omega)P_{bal}(x|y;\omega)\mathrm{d}x - \int f(x;\phi)P(x|y;\omega)\mathrm{d}x\right\|^2 \tag{4}$$

where MM stands for 'moment matching'. Let us ignore the irrelevant constant in $\mathrm{MM}'(r)$, and define the rest as $\mathrm{MM}(r)$:

$$\left\|\int f(x;\phi)r(x|y;\omega)P_{bal}(x|y;\omega)\mathrm{d}x\right\|^2 - 2\left\langle\int f(x;\phi)r(x|y;\omega)P(x|y;\omega)\mathrm{d}x, \int f(x;\phi)P(x|y;\omega)\mathrm{d}x\right\rangle \tag{5}$$

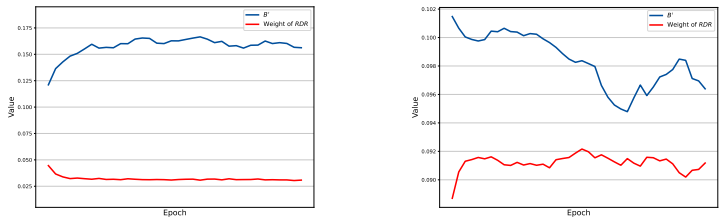
where $\langle\cdot,\cdot\rangle$ denotes the inner product. In practice, as for the real-world imbalanced data distribution P, we denote $\mathbf{\Phi}_P$ to dynamically reflect the knowledge learned from the distribution P, that is $\mathbf{\Phi}_P = (f(x_1;\phi),\ldots,f(x_n;\phi))$. Remember that the output of feature extractor is z, *i.e.*, $z = f(x;\phi)$ is a Z-dimensional vector, then $\mathbf{\Phi}_P$ would be a $[Z,n]$-dimensional vector. Similarly, we denote $\mathbf{\Phi}_P^i$ for class i, which is a $[Z,n_i]$-dimensional vector. As for the balanced data distribution P_{bal}, we design a momentum mechanism to accumulatively estimate the expectation of features learned from balanced data distribution along with the training. Concretely, for each class, we maintain a prototype feature F for the entire training progress, using each batch's feature expectation for momentum updates. Therefore, we define $F_{P_{bal}}$ as follows $F_{P_{bal}} = (F_1,\ldots,F_d)$.

Since the total number of classes is d, $F_{P_{bal}}$ would be a $[Z,d]$-dimensional vector. For each batch, we can obtain $\overline{z} = (\overline{z}_1,\ldots,\overline{z}_d)$, where $\overline{z}_i$ the mean of z of all samples in class i. Then, the momentum updates works as follows

$$F_{P_{bal}} \leftarrow mF_{P_{bal}} + (1-m)\overline{z} \tag{6}$$

where $m \in [0,1)$ is a momentum coefficient. Back to Eq. (5), replace the expectations over P_{bal} and P by $\mathbf{\Phi}_P$ and $F_{P_{bal}}$, respectively. Then, take the derivative of $\mathrm{MM}(r)$ with respect to r and set it to zero. Detailed derivations are provided in Appendix C.1. For each class i, we can obtain the estimation of density ratio in imbalanced learning as follows

$$\widehat{r}_i = n_i\left(\mathbf{\Phi}_P^{i\top}\mathbf{\Phi}_P^i\right)^{-1}\mathbf{\Phi}_P^{i\top}F_i \tag{7}$$



(a) *Many* Category (b) *Few* Category

Figure 1: Dynamic trend of the RDR weights well inversely aligns with B' throughout the training process in different categories. B'_y denotes $\sqrt{\pi_y}\left[1-\text{softmax}\left(B_y(m)\right)\right]$, where $B_y(m)$ denotes the minimal prediction on the ground-truth class y, i.e., $\min_{\boldsymbol{x}\in\mathcal{S}_y} m(\boldsymbol{x})_y$. Experiments were conducted on CIFAR-10-LT dataset with an imbalance factor of 10.

Substitute Eq. (7) into Eq. (1), we can obtain our object to optimize

$$\begin{aligned}\overline{R} &= \sum_{i=1}^{d}\frac{1}{\pi_i}\sum_{y_j=i} n_i \left(\boldsymbol{\Phi}_P^{i\,\top}\boldsymbol{\Phi}_P^i\right)^{-1}\boldsymbol{\Phi}_P^{i\,\top} F_i\cdot l(x_j,y_j;\omega)\\ &\propto \sum_{i=1}^{d}\sum_{y_j=i}\left(\boldsymbol{\Phi}_P^{i\,\top}\boldsymbol{\Phi}_P^i\right)^{-1}\boldsymbol{\Phi}_P^{i\,\top} F_i\cdot l(x_j,y_j;\omega)\end{aligned} \tag{8}$$

In our implementation, we introduced a warm-up phase to pre-adapt the feature distribution in $F_{P_{bal}}$, thereby mitigating excessive oscillations during the initial stages of training. Additionally, we employed a temperature coefficient γ to modulate the influence of weights, which is typically set to 1. When integrating with logit adjustment (LA) [Menon et al., 2020], we adhere to the same procedures outlined in [Wang et al., 2023] to ensure the fisher consistency. The framework and pseudo-code of our method are shown in Appendix C.2 and Appendix C.3.

3.4 Generalization Bound Analysis

Here, we use a formal generalization analysis to characterize the interesting point of our method.

Theorem 1. *Given a model $m\in\mathcal{M}$ and the loss function l, for any $\delta\in(0,1)$, with probability at least $1-\delta$ over the training set $\mathcal{S}$, according to [Wang et al., 2023], the following generalization bound holds for the risk on the balanced distribution*

$$R_{bal}^{l}(m) \precsim \Phi\left(l,\delta\right) + \frac{\mathfrak{S}_{\mathcal{S}}(\mathcal{M})}{d\pi_1}\sum_{y=1}^{d} w_y\sqrt{\pi_y}\left[1-\text{softmax}\left(B_y(m)\right)\right] \tag{9}$$

where $\Phi\left(l,\delta\right)$ is positively correlated with the empirical re-weighting risk of the training set. $\mathfrak{C}_{\mathcal{S}}(\mathcal{M})$ denotes the empirical complexity of the function set $\mathcal{M}$. $B_y(f)$ denotes the minimal prediction on the ground-truth class y in the training set. w_y refers to the weight of class y of the re-weighting loss.

Specifically, from the above generalization bound, we can find two inherent requirements for re-weighting methods. 1) *Why re-weighting is necessary*: w_y helps to re-balance the imbalanced term $\sqrt{\pi_y}\left[1-\text{softmax}\left(B_y(m)\right)\right]$ to get a sharper bound. 2) *Why dynamic re-weighting is necessary*: The term $B_y(m)$ changes dynamically with model training. Therefore, we need a w_y that can adapt dynamically to the changes of $B_y(m)$. 3) *Why RDR works*: From Fig. 1, we can observe that the dynamic trend of the RDR weight aligns well with $\sqrt{\pi_y}\left[1-\text{softmax}\left(B_y(m)\right)\right]$, denoted as B'_y. This shows that our RDR can adapt to the dynamics in B'_y, maintaining a sharp bound during training.

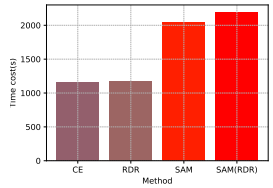
3.5 Implementation and Complexity Analysis

At the end of each epoch, a global variable is maintained and updated using momentum, as described by Eq. (6). Within each minibatch, the weight of each sample is computed dynamically. Typical

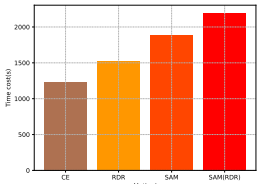
https://doi.org/10.52202/079017-2539

optimization procedures for deep neural networks entail both forward and backward passes per mini-batch, characterized by a computational complexity of $\mathcal{O}(B\Lambda)$, where B represents the batch size and Λ denotes the overall parameter size. Within the RDR framework, suppose the feature dimension used as input to the classifier is K, and the sample weights are computed according to Eq. (8). This computation for all d classes aggregates to a complexity of $\mathcal{O}(\sum_{i=1}^{d}(n_i \times K^2 + K^3))$ where n_i is the sample count of class i in a minibatch. Given that K generally exceeds n_i, the complexity predominantly stems from the matrix inversion, approximating to $\mathcal{O}(dK^3)$. The complexity for momentum updates is $\mathcal{O}(BK)$. Notice that K and B are considerably minor relative to the scale of the model parameters, rendering the time overhead of this method manageable.

On the storage front, the memory cost of the RDR primarily arises from the matrix inversion step in Eq. (8), resulting in a space complexity of $\mathcal{O}(K^2)$. Given the scales of K is much lower than Λ, the extra memory usage is negligible when compared with the memory utilization of the model parameters. To this end, RDR imposes a relatively small computational or space cost, enabling its integration with existing approaches at a reduced cost. An empirical evaluation of the computational expense is presented in Fig. 2. For more discussions about limitations of RDR , please refer to Appendix E.



(a) Time Cost on CIFAR-10-LT



(b) Time Cost on CIFAR-100-LT

Figure 2: Visualization of the time cost for training 200 epochs using four methods: CE, RDR, SAM and RDR(SAM) on CIFAR-10-LT and CIFAR-100-LT datasets.

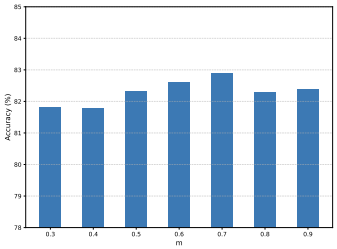


Figure 3: The impact of momentum coefficient m in RDR under the measure of top-1 accuracy.

4 Expriments

4.1 Experimental Setup

Datasets. We conduct experiments on four major long-tailed datasets, CIFAR-10-LT, CIFAR-100-LT, ImageNet-LT [Liu et al., 2019] and Places-LT [Liu et al., 2019]. CIFAR-10-LT and CIFAR-100-LT are two datasets sampled from the original CIFAR [Krizhevsky et al., 2009] dataset with a total of 10 and 100 classes, respectively. We conduct experiments with different imbalance factors $IF = \frac{n_{max}}{n_{min}}$, where n_{max} and n_{min} denotes the number of the most and least frequent classes [Kang et al., 2020, Hong et al., 2023, 2024b]. Following the mainstream protocol [Wang et al., 2023], we set the imbalance factor as 100 and 10 for evaluation. ImageNet-LT has 115.8K training images covering 1000 classes, with imbalance factor being 256. The number of samples per class ranges from 1280 to 5 images. Places-LT contains 62.5K training images covering 365 categories, with imbalance factor being 996. The number of samples per class ranges from 4980 to 5 images.

Evaluation Protocol. In the task of long-tailed classification, all classes are treated equally during testing. Following [Rangwani et al., 2022, Zhou et al., 2023c], we also report accuracy on three splits of classes according to the number of training data. Since the number of samples per class increases by its class index, for CIFAR-10-LT dataset, class[0, 3), class[3, 7) and class[7, 10) are reported as *Many*, *Medium* and *Few* classes, respectively. Similarly, CIFAR-100-LT is splited as class[0, 35), class[35, 69) and class[69, 100). ImageNet-LT is splited as class[0, 390), class[390, 835) and class[835, 1000), while Places-LT is splited as class[0, 131), class[131, 288) and class[288, 365).

Baselines. Our method is combined with existing long-tailed classification methods to demonstrate the efficacy, including the baseline trained by cross-entropy loss (CE), focal loss (Focal) [Lin et al., 2017], class-balanced loss (CB) [Cui et al., 2019] and logit adjustment (LA) [Menon et al., 2020]. Recently, Sharpness-Aware minimization (SAM) [Foret et al., 2021] has been proved to be a powerful method in imbalanced learning, therefore we also adopt baseline including SAM [Foret et al., 2021],

Table 1: Top-1 accuracy (%) (↑) results for overall classes on CIFAR-10-LT, CIFAR-100-LT, ImageNet-LT and Places-LT, CIFAR-10-LT and CIFAR-100-LT are employed with imbalance factors of 10 and 100, respectively.

Dataset	CIFAR-10-LT		CIFAR-100-LT		ImageNet-LT	Places-LT
	IF=10	IF=100	IF=10	IF=100		
CE	88.9±0.4	75.6±0.8	59.3±0.7	42.7±0.3	43.2±0.1	29.3±0.2
Focal	89.0±0.3	76.0±0.1	59.7±0.5	43.0±0.6	43.8±0.8	29.5±0.2
CB	89.0±0.4	76.7±0.8	60.4±0.6	43.5±1.2	43.8±0.1	32.5±0.3
LA	89.2±0.3	82.2±0.7	62.3±0.5	48.2±0.4	47.9±0.4	37.5±0.2
RDR+CE	89.9±0.1	81.9±0.1	62.3±0.4	48.5±0.4	45.2±0.1	39.4±0.2
RDR+LA	**90.2**±0.4	**83.4**±0.3	**62.9**±0.2	**49.4**±0.3	**48.1**±0.3	**39.5**±0.1

ImbSAM [Zhou et al., 2023a] and CCSAM [Zhou et al., 2023c], the latter two are also SAM-based methods. The strategies above have been demonstrated superior performance in imbalanced learning.

Implementation details. Our code is implemented with Pytorch 1.12.1. Experiments based on CIFAR-10-LT and CIFAR-100-LT are carried out on NVIDIA GeForce RTX 3090 GPUs, while experiments based on ImageNet-LT and Places-LT are carried out on NVIDIA A100 GPUs. For a fair comparison, we use ResNet32 on CIFAR-10-LT and CIFAR-100-LT, ResNet50 on ImageNet-LT and pre-trained ResNet-152 on Places-LT. We train each model with batch size of 128 (for CIFAR-10-LT and CIFAR-100-LT) / 256 (for Places-LT and ImageNet-LT), SGD optimizer with momentum of 0.9, weight decay of 0.0002. The initial learning rate is set to 0.1, with cosine learning-rate scheduling along training. The results of ImbSAM and CCSAM are obtained by implementing the official codes.

4.2 Comparison Results

Comparative analyses have been performed to evaluate the effectiveness of the proposed RDR. The results are presented in Table 1, Table 2 and Table 4. The metric employed to measure performance is the top-1 accuracy on the test sets.

Results on CIFAR-10-LT and CIFAR-100-LT. We first evaluate RDR on CIFAR-10-LT and CIFAR-100-LT. We report the final accuracy of different methods with imbalance factor ratio {10, 100} in Table 1 and Table 2. We can observe that RDR significantly outperforms all baselines under different imbalance factor ratios across the two datasets. Our observations highlight that RDR consistently outperforms baselines across various class distributions—*Many*, *Medium*, and *Few*—particularly under severe imbalance (*IF*=100).

In CIFAR-10-LT, combined with CE and LA, our method shows substantial improvement in the categories with fewer samples, increasing the accuracy by 19.8% and 9.5% respectively in the Few category under *IF*=100. This improvement is notable as it effectively addresses the challenge of learning from scarce data. With the inclusion of SAM, the performance of RDR is further enhanced. Under a less severe imbalance (*IF*=10), where the results show less performance drop-off between categories, RDR combined methods still maintain high performance across all categories, suggesting scalability and reliability of our approach in different imbalance contexts.

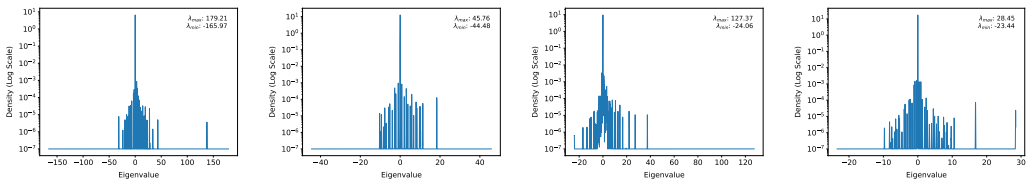
In CIFAR-100-LT, where the data distributions are more diverse and challenging, RDR also enhances the overall performance, particularly for the *Medium* and *Few* categories. Under the imbalance factor of 10 and 100, RDR increases the accuracy in *Few* classes by 11.8% and 20.0% respectively, compared to the original CE loss. Furthermore, it is notable that techniques like ImbSAM and CCSAM, which especially focus on the *Few* categories, may heavily sacrifice the performance on *Many* classes. The results in both datasets show that RDR generally outperforms in the *Many* classes compared to the other two variants of SAM, indicating that RDR can efficiently address the overfitting issues for *Few* classes. For more experimental details, please refer to Appendix D.2 and Appendix D.1.

Flat minima of loss landscape. Key metrics associated with Eigen Spectral Density, such as the maximum and minimum eigenvalues (λ_{max} and λ_{min}) and the trace of the Hessian matrix ($Tr(H)$), effectively reflect the smoothness of the loss landscape. Lower values of λ_{max} and $Tr(H)$ indicate a smoother loss landscape. Rangwani et al. [2022] have demonstrated that smoother loss landscapes correlate with stronger model generalization, which is particularly crucial when dealing

https://doi.org/10.52202/079017-2539

Table 2: Top-1 accuracy (%) (↑) results for *Many*, *Medium*, *Few* and overall classes on CIFAR-10-LT, categorized by imbalance factors (*IF*) of 100 and 10. The experiments are employed with the integration of Sharpness-Aware-Minimization-based methods.

Dataset	Loss	Method	IF=100				IF=10			
			Many	Med.	Few	All	Many	Med.	Few	All
CIFAR-10-LT	CE	SAM	94.8	74.5	60.6	76.4	95.8	85.9	86.5	89.3
		ImbSAM	94.6	73.9	67.7	78.3	94.2	84.1	90.1	89.0
		CCSAM	85.6	79.2	80.3	81.4	90.6	85.3	90.8	88.5
		RDR+SAM	91.8	78.6	81.9	**83.6**	94.2	86.3	92.5	**90.5**
	LA	SAM	90.8	78.1	82.9	83.3	92.6	86.5	91.8	89.9
		ImbSAM	85.2	75.1	89.6	82.5	90.8	83.6	94.8	89.1
		CCSAM	85.8	77.8	80.5	81.0	90.7	82.4	90.8	87.4
		RDR+SAM	88.5	80.0	85.0	**84.1**	92.5	86.5	93.6	**90.4**
CIFAR-100-LT	CE	SAM	72.7	40.3	7.5	41.5	75.3	60.2	45.1	60.8
		ImbSAM	71.5	40.6	17.7	44.3	72.5	57.9	53.5	61.7
		CCSAM	61.5	50.8	29.7	48.0	62.8	59.3	54.5	59.0
		RDR+SAM	63.4	51.9	30.5	**49.3**	68.0	61.6	58.3	**62.8**
	LA	SAM	64.3	50.5	30.8	49.2	66.1	60.9	59.7	62.3
		ImbSAM	58.8	45.4	40.1	48.4	62.9	56.6	64.2	61.2
		CCSAM	57.7	48.9	29.0	45.8	61.0	57.8	51.9	57.1
		RDR+SAM	63.9	52.4	30.5	**49.6**	67.6	61.7	60.1	**63.2**



(a) CE (b) RDR (c) SAM (d) RDR(SAM)

Figure 4: Eigen spectral density for the class with the fewest samples across different methods. Experiments conduct on the CIFAR-10-LT, under an imbalance factor of 100. Maximum eigenvalue λ_{max} (↓) minimum eigenvalue λ_{min} (↑) in the top right corner of each panel. A lower λ_{max} indicates a smoother loss landscape, while a higher λ_{min} suggests conditions more favorable for escaping from saddle points, thereby enhancing the model's generalization capabilities.

with imbalanced data. λ_{min} also serves as a significant indicator of the loss landscape characteristics. A preponderance of negative eigenvalues from the Hessian spectrum, resulting in smaller λ_{min} values, empirically suggests convergence to saddle points. Saddle points typically represent regions in the loss landscape characterized by a plateau with some negative curvature. In non-convex settings, it has been shown that an exponential number of saddle points exist, and convergence to these points is indicative of poor generalization.

Fig. 4 illustrates the Eigen Spectral Density under different loss function training regimes. It is evident that combining our method with the CE technique significantly improves the loss landscape. On one hand, λ_{max} is substantially reduced, indicating a flatter loss landscape. On the other hand, there is an increase in λ_{min}, suggesting our method's effectiveness in escaping from saddle points.

Table 3 delves deeper into the changes in the loss landscape for all *Few* classes. It reveals that combining our method with CE and SAM results in average reductions in λ_{max} by 58.5% and 66.1%, respectively, and increases in λ_{min} by 49.2% and 7.6%, respectively. Furthermore, our method significantly reduces $Tr(H)$ for minority classes (class7, class8 and class9). The average value of $Tr(H)$ decreases by 55.7% and 66.0% when combined with CE and SAM, respectively. These findings underscore the efficacy of our method in improving the loss landscape and enhancing the generalization capability of the model [Dauphin et al., 2014].

Results on ImageNet-LT and Places-LT. Our experiments conducted on ImageNet-LT and Places-LT, two large-scale datasets characterized by irregular and complex data distributions, demonstrate notable accuracy improvements through the application of RDR, as shown in Table 1 and Table 4.

https://doi.org/10.52202/079017-2539

Table 3: Loss landscape metrics across different methods on CIFAR-10-LT, with imbalance factor 100. Average minimum eigenvalues $\overline{\lambda_{\text{min}}}$ (↑), average maximum eigenvalues $\overline{\lambda_{\text{max}}}$ (↓), and the trace Tr (↓) of the Hessian matrix for classes with few samples. Tr_6, Tr_7, Tr_8, and Tr_9 represent the traces of the Hessian matrix for class 6, 7, 8 and 9, respectively, with descending sample quantities. $\overline{Tr}_{Few}$ denotes average trace of Hessian matrix over *Few* classes. Lower λ_{max} and Tr values indicate a flatter loss landscape, while a higher λ_{min} suggests a landscape more conducive to escaping from saddle points, thereby potentially enhancing model generalization.

Method	$\overline{\lambda_{\text{min}}}$	$\overline{\lambda_{\text{max}}}$	Tr_6	Tr_7	Tr_8	Tr_9	$\overline{Tr}_{Few}$
SGD	-76.24	110.78	335.64	301.67	234.97	323.87	286.84
RDR	-38.73	45.95	290.06	293.04	46.15	**41.71**	126.96
SAM	-14.42	74.82	300.13	314.48	123.10	291.00	242.86
RDR+SAM	**-13.33**	**25.36**	**179.50**	**147.11**	**36.28**	64.32	**82.57**

Specifically, when combined with CE and LA on ImageNet-LT, our method achieves accuracy enhancements of 18.7% and 2% in *Few* classes, respectively. While the SAM technique combined with CE and LA offers limited accuracy improvements, its integration with our approach still results in an overall accuracy increase of approximately 2.3% compared to other methods. Notably, the LA method tends to suppress accuracy in *Many* classes more than the CE method; however, this side effect is effectively mitigated when LA is combined with RDR.

Table 4: Top-1 accuracy (%) (↑) results for *Many*, *Medium*, *Few* and overall classes on ImageNet-LT and Places-LT. The experiments are employed with the integration of Sharpness-Aware-Minimization-based methods.

Loss	Method	ImageNet-LT				Places-LT			
		Many	Med.	Few	All	Many	Med.	Few	All
CE	SAM	64.6	35.8	10.1	42.8	45.2	27.6	12.1	30.6
	ImbSAM	62.5	37.3	13.9	43.3	43.1	26.2	16.1	30.1
	CCSAM	54.1	44.1	30.8	45.8	40.4	39.7	31.7	38.2
	RDR+SAM	59.1	46.5	26.2	**48.1**	41.5	41.3	39.0	**40.9**
LA	SAM	39.6	45.9	39.1	42.3	42.5	41.9	35.5	40.8
	ImbSAM	56.1	45.1	37.9	48.2	38.6	35.8	40.4	37.8
	CCSAM	52.2	43.9	32.8	45.3	32.1	41.4	40.9	37.9
	RDR+SAM	57.5	49.7	35.9	**50.5**	41.5	42.9	37.8	**41.3**

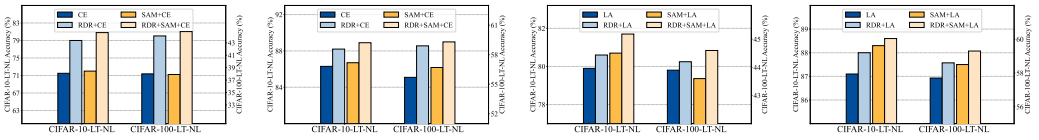
In the Places-LT dataset, the performance of RDR is even more pronounced. Combinations of our method with CE and LA result in accuracy gains of 10.1% and 2% in overall classes, respectively. Additionally, integrating SAM with our method also yields incremental improvements of 10.3% and 0.5% under CE and LA conditions, respectively. Our approach not only enhances accuracy in *Few* classes but also surpasses other methods in *Medium* classes, indicating its comprehensive efficacy across different categories. This broad applicability is particularly crucial for addressing the challenges of imbalanced learning.

Results on data with label noise. We further investigate the performance of our approach on datasets with label noise. Specifically, we evaluate two datasets: CIFAR-10-LT-NL and CIFAR-100-LT-NL, both of which exhibit class imbalance and label noise. Experiments are conducted with a noise ratio of 5%, and the results are presented in Fig. 5. As shown, our method demonstrates consistent and significant improvements on more datasets with label noise.

4.3 Ablation Study

In our study, we perform an ablation experiment to validate the efficacy of the multiple components that comprise our method. The outcomes from experiments across four datasets are delineated in Table 5. It is crucial to note that our weighting definition for each category i follows the formula $w = r/n_i$. When $r = 1$, our method simplifies to the traditional inverse frequency weighting $w = 1/n_i$. We explored the differences between our approach with and without the integration of SAM compared to this conventional weighting method.

https://doi.org/10.52202/079017-2539



(a) *IF*=10 with CE (b) *IF*=100 with CE (c) *IF*=10 with LA (d) *IF*=100 with LA

Figure 5: Top-1 accuracy (%) (↑) results for overall classes on CIFAR-10-LT-NL and CIFAR-100-LT-NL with 5% noise ratio, categorized by imbalance factors (*IF*) of 100 and 10.

Table 5: Top-1 accuracy (%) (↑) results from ablation studies across diverse datasets. Experiments conduct on CIFAR-10-LT, CIFAR-100-LT, ImageNet-LT, and Places-LT, comparing different method combinations. $1/n$ denote classic inverse frequency weighting method, assigning weights of $1/n_i$ for class i.

Loss	Method			CIFAR-10-LT		CIFAR-100-LT		ImageNet-LT	Places-LT
	$1/n$	RDR	SAM	IF=100	IF=10	IF=100	IF=10		
CE				75.61	88.86	42.66	60.17	43.18	29.27
	✓			77.78	89.50	44.21	61.89	44.62	34.27
		✓		81.87	89.94	48.54	62.28	45.19	35.45
			✓	76.44	89.33	41.46	60.82	42.78	30.64
	✓		✓	80.98	90.22	44.82	61.43	47.47	38.73
		✓	✓	**82.88**	**90.52**	**49.30**	**62.80**	**48.06**	**48.06**
LA				82.15	89.17	48.34	62.27	47.86	37.51
	✓			82.27	89.27	48.55	62.26	45.86	37.48
		✓		83.44	90.21	49.35	62.88	48.12	37.79
			✓	83.37	89.89	49.21	62.34	42.34	40.77
	✓		✓	82.71	90.25	47.36	61.88	43.40	39.48
		✓	✓	**83.56**	**90.41**	**49.63**	**63.06**	**50.45**	**41.33**

The results depicted in Table 5 reveal that our dynamic weighting approach consistently outperforms the classic method under various scenarios. Without SAM, when combined with CE and LA, our method achieves accuracy improvements ranging from 0.4% to 4.3% and 0.3% to 2.3%, respectively. When integrated with SAM, the improvement in accuracy is particularly notable on large datasets. Specifically, the accuracy enhancements on ImageNet-LT and Places-LT reach 7.1% and 1.9% respectively when combined with LA. These results underscore the tangible benefits of our dynamic weighting strategy in enhancing model performance. The impact of momentum coefficient m in RDR is shown in Fig. 3. For more experimental details, please refer to Appendix D.3.

5 Conclusion

In this work, we have introduced RDR, a novel approach for mitigating model degradation in imbalanced learning scenarios by dynamically adjusting class weights using density ratio estimation. Our method dynamically adjusts class weights during training based on density ratio estimation, enhancing both model robustness and adaptability. Extensive experiments on diverse large-scale datasets demonstrate the effectiveness of RDR, particularly in severely imbalanced settings. Future work will focus on refining the dynamic adjustment mechanisms and exploring broader applicability across various domains and dataset complexities.

Acknowledgement

Jiaan Luo, Feng Hong, Jiangchao Yao, Ya Zhang and Yanfeng Wang are supported by the National Key R&D Program of China (No. 2022ZD0160702), STCSM (No. 22511106101, No. 22DZ2229005), 111 plan (No. BP0719010) and National Natural Science Foundation of China (No. 62306178). Bo Han is supported by NSFC General Program No. 62376235, Guangdong Basic and Applied Basic Research Foundation No. 2022A1515011652 and No. 2024A1515012399, HKBU Faculty Niche Research Areas No. RC-FNRA-IG/22-23/SCI/04 and HKBU CSD Departmental Incentive Scheme.

References

Kevin W. Bowyer, Nitesh V. Chawla, Lawrence O. Hall, and W. Philip Kegelmeyer. SMOTE: synthetic minority over-sampling technique. *CoRR*, abs/1106.1813, 2011. URL http://arxiv.org/abs/1106.1813.

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. doi: 10.1016/J.NEUNET.2018.07.011. URL https://doi.org/10.1016/j.neunet.2018.07.011.

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.

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.

Mengxi Chen, Jiangchao Yao, Linyu Xing, Yu Wang, Ya Zhang, and Yanfeng Wang. Redundancy adaptive multimodal learning for imperfect data. *arXiv preprint arXiv:2310.14496*, 2023a.

Xiaohua Chen, Yucan Zhou, Dayan Wu, Chule Yang, Bo Li, Qinghua Hu, and Weiping Wang. Area: adaptive reweighting via effective area for long-tailed classification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 19277–19287, 2023b.

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.

Yann N Dauphin, Razvan Pascanu, Caglar Gulcehre, Kyunghyun Cho, Surya Ganguli, and Yoshua Bengio. Identifying and attacking the saddle point problem in high-dimensional non-convex optimization. *Advances in neural information processing systems*, 27, 2014.

Pierre Foret, Ariel Kleiner, Hossein Mobahi, and Behnam Neyshabur. Sharpness-aware minimization for efficiently improving generalization. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021. URL https://openreview.net/forum?id=6Tm1mposlrM.

Yu Gong, Greg Mori, and Frederick Tung. Ranksim: Ranking similarity regularization for deep imbalanced regression. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvári, Gang Niu, and Sivan Sabato, editors, *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, volume 162 of *Proceedings of Machine Learning Research*, pages 7634–7649. PMLR, 2022. URL https://proceedings.mlr.press/v162/gong22a.html.

Hui Han, Wenyuan Wang, and Binghuan Mao. Borderline-smote: A new over-sampling method in imbalanced data sets learning. In *ICIC*, volume 3644 of *Lecture Notes in Computer Science*, pages 878–887. Springer, 2005.

Feng Hong, Jiangchao Yao, Zhihan Zhou, Ya Zhang, and Yanfeng Wang. Long-tailed partial label learning via dynamic rebalancing. In *International Conference on Learning Representations*. OpenReview.net, 2023.

Feng Hong, Yueming Lyu, Jiangchao Yao, Ya Zhang, Ivor W. Tsang, and Yanfeng Wang. Diversified batch selection for training acceleration. In *International Conference on Machine Learning*. OpenReview.net, 2024a.

Feng Hong, Jiangchao Yao, Yueming Lyu, Zhihan Zhou, Ivor W. Tsang, Ya Zhang, and Yanfeng Wang. On harmonizing implicit subpopulations. In *International Conference on Learning Representations*. OpenReview.net, 2024b.

Chen Huang, Yining Li, Chen Change Loy, and Xiaoou Tang. Learning deep representation for imbalanced classification. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 5375–5384. IEEE Computer Society, 2016. doi: 10.1109/CVPR.2016.580. URL https://doi.org/10.1109/CVPR.2016.580.

 https://doi.org/10.52202/079017-2539

Bingyi Kang, Saining Xie, Marcus Rohrbach, Zhicheng Yan, Albert Gordo, Jiashi Feng, and Yannis Kalantidis. Decoupling representation and classifier for long-tailed recognition. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020. URL https://openreview.net/forum?id=r1gRTCVFvB.

Mahsa Keramati, Lili Meng, and R. David Evans. Conr: Contrastive regularizer for deep imbalanced regression. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024. URL https://openreview.net/forum?id=RIuevDSK5V.

Ganesh Ramachandra Kini, Orestis Paraskevas, Samet Oymak, and Christos Thrampoulidis. Label-imbalanced and group-sensitive classification under overparameterization. In Marc'Aurelio Ranzato, Alina Beygelzimer, Yann N. Dauphin, Percy Liang, and Jennifer Wortman Vaughan, editors, *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pages 18970–18983, 2021. URL https://proceedings.neurips.cc/paper/2021/hash/9dfcf16f0adbc5e2a55ef02db36bac7f-Abstract.html.

Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. *Handbook of Systemic Autoimmune Diseases*, 2009.

Miroslav Kubat and Stan Matwin. Addressing the curse of imbalanced training sets: One-sided selection. In *ICML*, pages 179–186. Morgan Kaufmann, 1997.

Tianhong Li, Peng Cao, Yuan Yuan, Lijie Fan, Yuzhe Yang, Rogerio S Feris, Piotr Indyk, and Dina Katabi. Targeted supervised contrastive learning for long-tailed recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6918–6928, 2022.

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.

Hong Liu, Jeff Z. HaoChen, Adrien Gaidon, and Tengyu Ma. Self-supervised learning is more robust to dataset imbalance. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net, 2022. URL https://openreview.net/forum?id=4AZz9osqrar.

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.

Yanbiao Ma, Licheng Jiao, Fang Liu, Yuxin Li, Shuyuan Yang, and Xu Liu. Delving into semantic scale imbalance. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023. URL https://openreview.net/pdf?id=07tc5kKRIo.

Aditya Krishna Menon, Sadeep Jayasumana, Ankit Singh Rawat, Himanshu Jain, Andreas Veit, and Sanjiv Kumar. Long-tail learning via logit adjustment. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*, 2020.

Katharina Morik, Peter Brockhausen, and Thorsten Joachims. Combining statistical learning with a knowledge-based approach - A case study in intensive care monitoring. In Ivan Bratko and Saso Dzeroski, editors, *Proceedings of the Sixteenth International Conference on Machine Learning (ICML 1999), Bled, Slovenia, June 27 - 30, 1999*, pages 268–277. Morgan Kaufmann, 1999.

Harsh Rangwani, Sumukh K. Aithal, Mayank Mishra, and Venkatesh Babu R. Escaping saddle points for effective generalization on class-imbalanced data. In Sanmi Koyejo, S. Mohamed, A. Agarwal, Danielle Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*, 2022. URL http://papers.nips.cc/paper_files/paper/2022/hash/8f4d70db9ecec97b6723a86f1cd9cb4b-Abstract-Conference.html.

Jiawei Ren, Cunjun Yu, Shunan Sheng, Xiao Ma, Haiyu Zhao, Shuai Yi, and Hongsheng Li. Balanced meta-softmax for long-tailed visual recognition. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin, editors, *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020a. URL https://proceedings.neurips.cc/paper/2020/hash/2ba61cc3a8f44143e1f2f13b2b729ab3-Abstract.html.

Jiawei Ren, Cunjun Yu, Shunan Sheng, Xiao Ma, Haiyu Zhao, Shuai Yi, and Hongsheng Li. Balanced meta-softmax for long-tailed visual recognition. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin, editors, *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020b. URL https://proceedings.neurips.cc/paper/2020/hash/2ba61cc3a8f44143e1f2f13b2b729ab3-Abstract.html.

Jiawei Ren, Mingyuan Zhang, Cunjun Yu, and Ziwei Liu. Balanced MSE for imbalanced visual regression. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 7916–7925. IEEE, 2022. doi: 10.1109/CVPR52688.2022.00777. URL https://doi.org/10.1109/CVPR52688.2022.00777.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael S. Bernstein, Alexander C. Berg, and Li Fei-Fei. Imagenet large scale visual recognition challenge. *Int. J. Comput. Vis.*, 115(3):211–252, 2015. doi: 10.1007/S11263-015-0816-Y. URL https://doi.org/10.1007/s11263-015-0816-y.

Yonglong Tian, Olivier J. Hénaff, and Aäron van den Oord. Divide and contrast: Self-supervised learning from uncurated data. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, pages 10043–10054. IEEE, 2021.

Byron C. Wallace, Kevin Small, Carla E. Brodley, and Thomas A. Trikalinos. Class imbalance, redux. In *ICDM*, pages 754–763. IEEE Computer Society, 2011.

Jiaqi Wang, Wenwei Zhang, Yuhang Zang, Yuhang Cao, Jiangmiao Pang, Tao Gong, Kai Chen, Ziwei Liu, Chen Change Loy, and Dahua Lin. Seesaw loss for long-tailed instance segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pages 9695–9704. Computer Vision Foundation / IEEE, 2021a. doi: 10.1109/CVPR46437.2021.00957. URL https://openaccess.thecvf.com/content/CVPR2021/html/Wang_Seesaw_Loss_for_Long-Tailed_Instance_Segmentation_CVPR_2021_paper.html.

Xudong Wang, Long Lian, Zhongqi Miao, Ziwei Liu, and Stella X. Yu. Long-tailed recognition by routing diverse distribution-aware experts. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021b. URL https://openreview.net/forum?id=D9I3drBz4UC.

Zitai Wang, Qianqian Xu, Zhiyong Yang, Yuan He, Xiaochun Cao, and Qingming Huang. A unified generalization analysis of re-weighting and logit-adjustment for imbalanced learning. In Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine, editors, *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023. URL http://papers.nips.cc/paper_files/paper/2023/hash/973a0f50d43cf99118cdab456edcacda-Abstract-Conference.html.

Ziyan Wang and Hao Wang. Variational imbalanced regression: Fair uncertainty quantification via probabilistic smoothing. In Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine, editors, *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023. URL http://papers.nips.cc/paper_files/paper/2023/hash/612a56f193d031687683445cd0001083-Abstract-Conference.html.

Yuzhe Yang, Kaiwen Zha, Ying-Cong Chen, Hao Wang, and Dina Katabi. Delving into deep imbalanced regression. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 11842–11851. PMLR, 2021. URL http://proceedings.mlr.press/v139/yang21m.html.

 https://doi.org/10.52202/079017-2539

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.

Bianca Zadrozny, John Langford, and Naoki Abe. Cost-sensitive learning by cost-proportionate example weighting. In *Proceedings of the 3rd IEEE International Conference on Data Mining (ICDM 2003), 19-22 December 2003, Melbourne, Florida, USA*, page 435. IEEE Computer Society, 2003. doi: 10.1109/ICDM.2003.1250950. URL https://doi.org/10.1109/ICDM.2003.1250950.

Yixuan Zhou, Yi Qu, Xing Xu, and Hengtao Shen. Imbsam: A closer look at sharpness-aware minimization in class-imbalanced recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11345–11355, 2023a.

Zhihan Zhou, Jiangchao Yao, Feng Hong, Ya Zhang, Bo Han, and Yanfeng Wang. Combating representation learning disparity with geometric harmonization. *Advances in Neural Information Processing Systems*, 36, 2023b.

Zhipeng Zhou, Lanqing Li, Peilin Zhao, Pheng-Ann Heng, and Wei Gong. Class-conditional sharpness-aware minimization for deep long-tailed recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3499–3509, 2023c.

A Notations

In Table 6, we summarize the notations used in this paper.

Table 6: Description of Notations

Category	Notation	Description
Data and Sets	$\mathcal{S}$	Training dataset
	π_i	Proportion of class i
	n_i	Sample number of class i
	d	Number of classes
	B	Training batch size
Model and Functions	$f(\cdot;\phi)$	Feature extractor with parameters ϕ
	$h(\cdot;\theta)$	Classifier with parameters θ
	$\omega = \bigcup\{\phi, \theta\}$	Model parameters consist of ϕ and θ
	$l(x, y; \omega)$	Loss function with input x, output y and parameters ω
	R	Empirical risk of the loss function
	P	Distribution on training set
	P_{bal}	Distribution on balanced data set
	r	Density ratio
	$\mathbf{\Phi}_P$	Matrix of knowledge learned from the distribution P
	$\mathbf{\Phi}_P^i$	$\mathbf{\Phi}_P$ for class i
	$F_{P_{bal}}$	Matrix of feature expectation for momentum update
	Z	Dimension of features extracted from feature extractor
	$B_y(\cdot)$	Minimal prediction on the ground-truth class y
	$m \in M$	Model m in function set M
Others	*IF*	Imbalance factor
	$\overline{\lambda_{max}}$, $\overline{\lambda_{min}}$	Average maximum and minimum eigenvalues
	$\boldsymbol{Tr}_i$	Trace of Hessian matrix for class i
	$\overline{\boldsymbol{Tr}}_{Few}$	Average trace of Hessian matrix over *Few* categories

B More Discussions of Related Work

Beyond imbalanced classification that has discretized label space, an noteworthy area, imbalanced regression that has a continuous label space is also very common in real applications [Yang et al., 2021, Gong et al., 2022]. In this direction, the empirical label distribution often does not accurately reflect the true label density in regression tasks, which limits the effectiveness of traditional re-weighting techniques [Yang et al., 2021, Wang and Wang, 2023]. Label Distribution Smoothing (LDS) [Yang et al., 2021] and Variational Imbalanced Regression (VIR) [Wang and Wang, 2023] propose using kernel smoothing and other techniques to estimate an accurate label density distribution. Ranking Similarity (Ranksim) [Gong et al., 2022] leverages local and global dependencies by encouraging the correspondence between the similarity order of labels and features. Balanced Mean Squared Error (Balanced MSE) [Ren et al., 2022] extends the concept of Balanced Softmax [Ren et al., 2020a] to regression tasks to achieve a balanced predictive distribution. Contrastive Regularizer (ConR) [Keramati et al., 2024] improves contrastive learning techniques to translate label similarities into the feature space.

C Algorithm Details

C.1 Detailed Derivations of Eq. (7)

Back to Eq. (5), replace the expectations over P_{bal} and P by $\mathbf{\Phi}_P$ and $F_{P_{bal}}$, respectively. For each class i, We can obtain $\widehat{r_i} = \widehat{\mathrm{MM}}(r)$, where

$$\widehat{\mathrm{MM}}(r) = \frac{1}{n_i^2} {r_i}^\top {\mathbf{\Phi}_P^i}^\top \mathbf{\Phi}_P^i r_i - \frac{2}{n_i} r_i^\top {\mathbf{\Phi}_P^i}^\top F_i \tag{10}$$

https://doi.org/10.52202/079017-2539

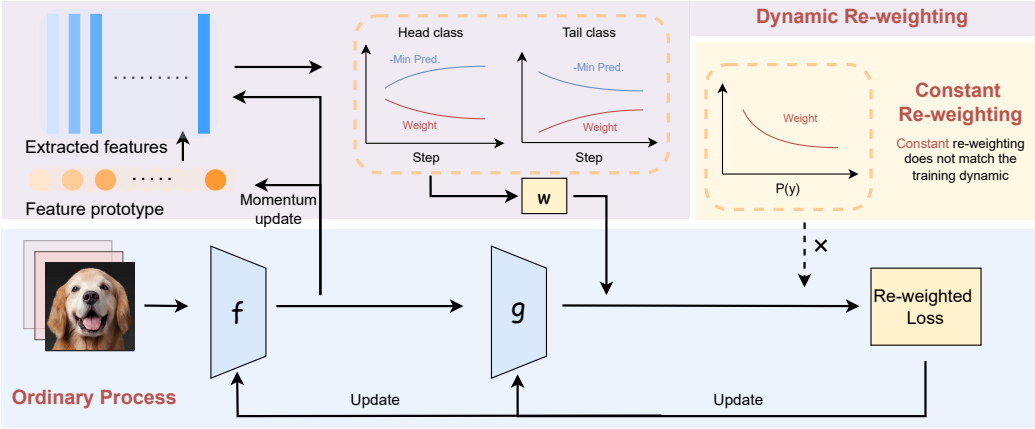


Figure 6: Framework of RDR

Then, taking the derivative of $\widehat{\mathrm{MM}}(r)$ with respect to r and setting it to zero, we can obtain the estimation of density ratio in imbalanced learning as follows

$$\frac{2}{n_i^2}{\mathbf{\Phi}_P^i}^\top \mathbf{\Phi}_P^i r_i - \frac{2}{n_i}{\mathbf{\Phi}_P^i}^\top F_i = 0 \tag{11}$$

Solving equation above with respect to r_i, we can obtain the solution as

$$\widehat{r_i} = n_i \left({\mathbf{\Phi}_P^i}^\top \mathbf{\Phi}_P^i\right)^{-1} {\mathbf{\Phi}_P^i}^\top F_i \tag{12}$$

C.2 Framework of RDR

We provide the framework of RDR, which is shown in Fig. 6.

C.3 Pseudo-code of RDR

We provide the pseudo-code of RDR to demonstrate the process of implementing our method in detail, as shown in Algorithm 1. In addition, we also provide pseudo-code that combines our method with the SAM method, as shown in Algorithm 2.

D Supplement for Experiments

D.1 Experiment with More Imbalanced Data

We conduct experiments on the more imbalanced CIFAR-10-LT and CIFAR-100-LT datasets, specifically with imbalance factors of 200 and 500. As shown in Table 7, our method consistently achieves significant improvements.

D.2 Dynamically Re-weighting Process

Our approach dynamically adjusts the weights assigned to each category throughout the training process. To gain more insights into RDR, we sampled the weights of each category during training. Fig. 7 presents the results of four samplings during the training processes at imbalance factors of 10 and 100, respectively.

The analysis of these results reveals a consistent trend in weight changes across different imbalance factors. For the *Many* classes, the weights of the categories consistently decrease during training. Specifically, under the *IF* of 10, the weights of class0, class1, and class2 (the three classes with the highest sample counts, in descending order) decrease by 5.6%, 6.7%, and 8.0%, respectively. Under the *IF* of 100, these decreases are more pronounced, with reductions of 10.3%, 15.4%, and 14.7%, respectively. For the *Medium* classes, the weight changes are less marked, with an average decrease

Algorithm 1 Training Paradigm of RDR.

Input: Training dataset $\mathcal{S} = \cup_{i=1}^{n}\{(x_i, y_i)\}$, model $\mathcal{M}_\omega$ with feature extractor f_ϕ and classifier h_θ, loss function l, momentem coefficient m, learning rate α, weight decay coefficient λ, batch size b, temperature coefficient γ
Output: Trained parameters ϕ^*, θ^*
Initialize the model parameters ϕ and θ ramdomly, $F_{P_{bal}} = (F_1, \ldots, F_d) \leftarrow 0$
for $t = 1$ to T **do**
  $\mathcal{B} \leftarrow$SampleMiniBatch$(\mathcal{S}, b)$
  $z \leftarrow f(x, \phi_t)$
  $output \leftarrow h(z, \theta_t)$
  if $t < T_0$ **then**
    // warm up
    $w \leftarrow 1$
  else
    for class i to d **do**
      $\mathbf{\Phi}_P^i \leftarrow (z_j)$ where $y_j = i$
      compute w_i via Eq. (8) and γ
    end for
    $w \leftarrow$ normalize(w^γ)
  end if
  $\mathcal{L}(\omega_t, \mathcal{B}) \leftarrow \frac{1}{b}\sum_{\mathcal{B}} w \cdot l(output, y)$
  $\omega_t = \omega_t - \alpha_t\left[\nabla\mathcal{L}(\omega_t, \mathcal{B}) + \lambda\omega_t\right]$
  $F_{P_{bal}} \leftarrow mF_{P_{bal}} + (1-m)\overline{z}$
  Optional: anneal the learning rate α_t
end for

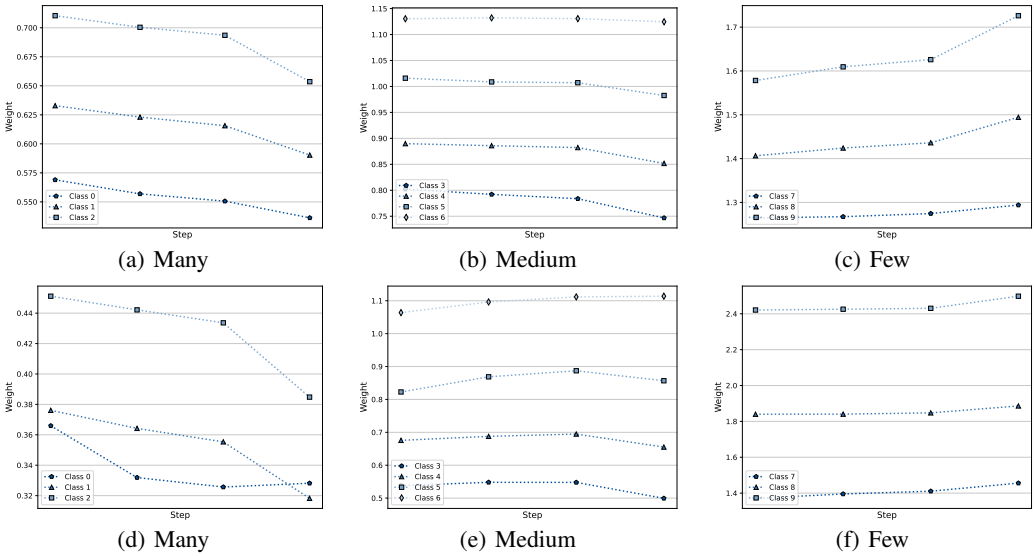
Algorithm 2 Training Paradigm of RDR combined with SAM.

Input: Training dataset $\mathcal{S} = \cup_{i=1}^{n}\{(x_i, y_i)\}$, model $\mathcal{M}_\omega$ with feature extractor f_ϕ and classifier h_θ, loss function l, momentem coefficient m, learning rate α, weight decay coefficient λ, batch size b, neighborhood size ρ, temperature coefficient γ
Output: Trained parameters ϕ^*, θ^*
Initialize the model parameters ϕ and θ ramdomly, $F_{P_{bal}} = (F_1, \ldots, F_d) \leftarrow 0$
for $t = 1$ to T **do**
  $\mathcal{B} \leftarrow$SampleMiniBatch$(\mathcal{S}, b)$
  $z \leftarrow f(x, \phi_t)$
  $output \leftarrow h(z, \theta_t)$
  if $t < T_0$ **then**
    // warm up
    $w \leftarrow 1$
  else
    for class i to d **do**
      $\mathbf{\Phi}_P^i \leftarrow (z_j)$ where $y_j = i$
      compute w_i via Equation 8 and γ
    end for
    $w \leftarrow$ normalize(w^γ)
  end if
  $\mathcal{L}_1(\omega_t, \mathcal{B}) \leftarrow \frac{1}{b}\sum_{\mathcal{B}} w \cdot l(output, y)$
  $\epsilon_t \leftarrow \rho\frac{\nabla\mathcal{L}_1(\omega_t, \mathcal{B})}{|\nabla\mathcal{L}_1(\omega_t, \mathcal{B})|}$
  $\mathcal{L}_2(\omega_t + \epsilon_t, \mathcal{B}) \leftarrow \frac{1}{b}\sum_{\mathcal{B}} w \cdot l(f_{\omega_t+\epsilon_t}(\cdot), y)$
  $\omega_t = \omega_t - \alpha_t\left[\nabla\mathcal{L}_2(\omega_t + \epsilon_t, \mathcal{B}) + \lambda\omega_t\right]$
  $F_{P_{bal}} \leftarrow mF_{P_{bal}} + (1-m)\overline{z}$
  Optional: anneal the learning rate α_t
end for

https://doi.org/10.52202/079017-2539

Table 7: Top-1 accuracy (%) (↑) results under more imbalanced conditions.

Method	IF=500		IF=200	
	CIFAR-10-LT	CIFAR-100-LT	CIFAR-10-LT	CIFAR-100-LT
CE	60.1	34.1	68.6	37.9
RDR+CE	69.7	39.0	76.7	43.1
SAM+CE	60.4	34.5	69.9	39.0
RDR+SAM+CE	**71.1**	**40.0**	**80.0**	**44.8**
LA	73.9	38.8	76.7	44.0
RDR+LA	75.8	40.1	80.8	45.0
SAM+LA	74.9	39.6	80.2	44.2
RDR+SAM+LA	**76.4**	**40.6**	**81.6**	**45.1**



(a) Many (b) Medium (c) Few

(d) Many (e) Medium (f) Few

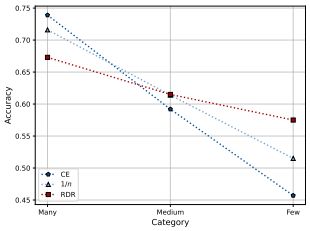
Figure 7: Visualization of weight changes across different categories throughout the training process. Experiments conducted on CIFAR-10-LT. (a), (b), and (c) illustrate the weight changes for *Many*, *Medium*, and *Few* classes, respectively, under an imbalance factor of 10. (d), (e), and (f) correspondingly show the changes for each class type under an imbalance factor of 100.

of 3.7% at *IF*=10 and 0.3% at *IF*=100. For the *Few* classes, there exits a notable increase in weights during training; at *IF*=10, the weights of class7, class8, and class9 increase by 2.3%, 6.3%, and 9.4%, respectively, while at *IF*=100, they increase by 6.2%, 6.5%, and 3.2%.

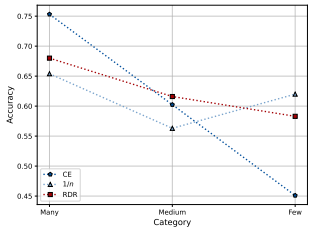
These results suggest that our method increasingly focuses on minority classes as training progresses. Initially, our method effectively learns common features across all categories, while later in training, increasing the weights helps to target learning towards minority samples, thereby enhancing the model's generalizability. Wang et al. [2023] also corroborate these findings.

D.3 Ablation Study

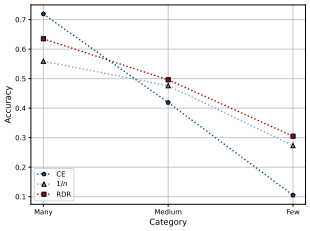
We provide more detailed experimental results for each category in ablation study, as illustrated in Fig. 8 and Fig. 9. From these figures, we can find that across various datasets and different imbalance factors, our method significantly enhances the generalizability of both *Few* classes and *Medium* classes. Moreover, our method maintains superior performance when combined with the SAM method.



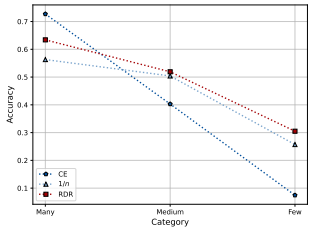
(a) Accuracy under *IF*=10



(b) Accuracy under *IF*=10 with SAM

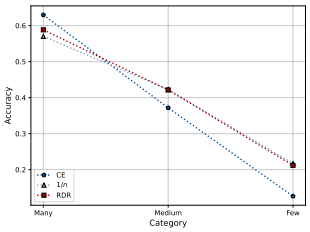


(c) Accuracy under *IF*=100

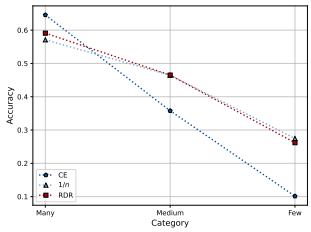


(d) Accuracy under *IF*=100 with SAM

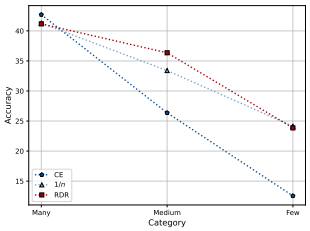
Figure 8: Visualization of top-1 accuracy (↑) across different categories on CIFAR-100-LT, under three different methods: CE, Inverse Frequency ($1/n$) and RDR. Experiments conducted under *IF*=10 (Plot (a) and Plot (b)) and 100 (Plot (c) and Plot (d)).



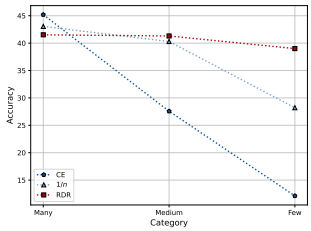
(a) Accuracy on ImageNet-LT



(b) Accuracy on ImageNet-LT with SAM



(c) Accuracy on Places-LT



(d) Accuracy on Places-LT with SAM

Figure 9: Visualization of top-1 accuracy (↑) across different categories on ImageNet-LT (Plot (a) and Plot (b)) and Places-LT (Plot (c) and Plot (d)), under three different methods: CE, Inverse Frequency ($1/n$) and RDR. Plot (b) and Plot (d) are integrated with SAM while Plot (a) and Plot (c) are not.

https://doi.org/10.52202/079017-2539

E More Discussions about Limitations

While our approach demonstrates promising results, there are potential challenges that warrant further attention. In particular, as the number of classes increases to a very large scale, especially in certain tasks such as face recognition, retail product recommendation, or landmark detection, there could be concerns regarding computational efficiency. It is important to consider lightweight techniques to ensure that scalability does not compromise practical applicability. Additionally, in addressing imbalanced learning, care must be taken to avoid excessive rebalancing toward minority groups, as this could unintentionally affect the learning performance of the majority class, which would not align with the broader goals of fairness. Rebalancing should be conducted within a reasonable framework, mindful of avoiding misuse or overcompensation that could arise from improper manipulation by any group. Ensuring fairness for all remains a critical consideration.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction have stated the claims.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: See Section 3.5 and Appendix E.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA]

https://doi.org/10.52202/079017-2539

Justification: This paper does not include theoretical results.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: See Section 4 and Appendix C.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [NA]

Justification: The code of this paper will be released after anonymized review.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: See Section 4.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: We report error bars for CIFAR-10-LT and CIFAR-100-LT dataset. For the large scale ImageNet-LT and Places-LT, we do not report error bars due to computational constraints.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).

https://doi.org/10.52202/079017-2539

- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: See Section 3 and Section 4.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics https://neurips.cc/public/EthicsGuidelines?

   Answer: [Yes]

   Justification: This paper is conducted with the NeurIPS Code of Ethics.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: Our proposed method can enhance the efficiency and robustness of imbalance learning, thereby increasing productivity.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: This paper poses no such risks.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [NA]

    Justification: This paper does not use such assets.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.

 https://doi.org/10.52202/079017-2539

- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification: This paper does not release new assets.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: This paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: This paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.