Open-Sampling: Exploring Out-of-Distribution Data for Re-balancing Long-tailed Datasets

Hongxin Wei [1]   Lue Tao [2]   Renchunzi Xie [1]   Lei Feng [3]   Bo An [1]

Abstract

Deep neural networks usually perform poorly when the training dataset suffers from extreme *class imbalance*. Recent studies found that directly training with out-of-distribution data (i.e., open-set samples) in a semi-supervised manner would harm the generalization performance. In this work, we theoretically show that out-of-distribution data can still be leveraged to augment the minority classes from a Bayesian perspective. Based on this motivation, we propose a novel method called *Open-sampling*, which utilizes open-set noisy labels to re-balance the class priors of the training dataset. For each open-set instance, the label is sampled from our pre-defined distribution that is complementary to the distribution of original class priors. We empirically show that Open-sampling not only re-balances the class priors but also encourages the neural network to learn separable representations. Extensive experiments demonstrate that our proposed method significantly outperforms existing data re-balancing methods and can boost the performance of existing state-of-the-art methods.

1. Introduction

The success of deep neural networks (DNNs) heavily relies on large-scale datasets with balanced distribution (Krizhevsky et al., 2009; Russakovsky et al., 2015). However, in real-world applications like autonomous driving and medical diagnosis, large-scale datasets naturally exhibit imbalanced and long-tailed distributions, i.e., a few classes (majority classes) occupy most of the data while most classes (minority classes) are under-represented (Zhou et al., 2017; Van Horn et al., 2018; Lin et al., 2014). It has been shown that training on long-tailed datasets leads to poor generalization performance, especially on the minority classes (Zhou et al., 2020; Liu et al., 2019a; Kang et al., 2020). Thus, designing effective algorithms to handle class imbalance is of great practical importance.

In the literature, a popular direction in long-tailed learning is to re-balance the data distribution by data re-sampling (Japkowicz & Stephen, 2002; He & Garcia, 2009). For example, Over-sampling (Byrd & Lipton, 2019; Shen et al., 2016) repeats samples from under-presented classes, but it usually causes over-fitting to the minority classes. To alleviate the over-fitting issue, synthesized novel samples are introduced to augment the minority classes without repetition (Chawla et al., 2002). As a result, the model is still error-prone due to noise in the synthesized samples (Cui et al., 2019) A recent work (Yang & Xu, 2020) introduced unlabeled-in-distribution data to compensate for the lack of training samples, and showed that directly adding unlabeled data from mismatched classes (*i.e.*, out-of-distribution data) by semi-supervised learning hurts the generalization performance. These data augmentation methods normally require in-distribution data with precise labels for selected classes. However, such kind of data would be extremely hard to collect in real-world scenarios, due to the expensive labeling cost. This fatal weakness of previous methods motivates us to explore the possibility of using *out-of-distribution* (OOD) data for long-tailed imbalanced learning.

In this paper, we theoretically show that out-of-distribution data (i.e., open-set samples) could be leveraged to augment the minority classes from a Bayesian perspective. Based on this motivation, we propose a simple yet effective method called Open-sampling, which uses open-set noisy labels to re-balance the label priors of the training dataset. For each OOD instance, the label is sampled from our pre-defined distribution that is complementary to the original class priors. To alleviate the over-fitting issue on the minority classes, a class-dependent weight is used in the training objective to provide stronger regularization on the minority classes than the majority classes. In this way, the open-set noisy labels could be used to re-balance the class priors while retaining their non-toxicity.

*Equal contribution [1]Nanyang Technological University, Singapore [2]Nanjing University of Aeronautics and Astronautics, Nanjing, Jiangsu, China [3]Chongqing University, Chongqing, China. Correspondence to: Renchunzi Xie <XIER0002@e.ntu.edu.sg>.

To provide a comprehensive understanding, we conduct a series of analyses to illustrate the properties of the proposed Open-sampling method. From these empirical analyses, we show that: 1) the Complementary Distribution is superior to the commonly used Class Balanced distribution (CB) as the former is closer to the uniform distribution, which reduces the harmfulness of the open-set noisy labels; 2) real-world datasets with large sample size are the best choices for the open-set auxiliary dataset in Open-sampling and the diversity (i.e., number of classes) is not an important factor in the method; 3) the Open-sampling method not only re-balances the class prior, but also promotes the neural network to learn more separable representations.

To the best of our knowledge, we are the first to explore the benefits of OOD instances in learning from long-tailed datasets. To verify the effectiveness of our method, we conduct experiments on four long-tailed image classification benchmark datasets, including long-tailed CIFAR10/100 (Krizhevsky et al., 2009), CelebA-5 (Liu et al., 2015b; Kim et al., 2020), and Places-LT (Zhou et al., 2017). Empirical results show that our method can be easily incorporated into existing state-of-the art methods to enhance their performance on long-tailed imbalanced classification tasks. Furthermore, experimental results validate that our method could also achieve impressive performance for detecting OOD examples under class-imbalanced setting. Code and data are publicly available at https://github.com/hongxin001/logitnorm_ood.

2. Imbalanced Learning with OOD instances

2.1. Background

In this work, we consider a multi-class classification problem, where the input space is denoted by $\mathcal{X} \in \mathbb{R}^d$ and the label space $\mathcal{Y}$ is $\{1, \ldots, K\}$. We denote by $\mathcal{D}_{\text{train}} = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^N \in \mathcal{X} \times \mathcal{Y}$ the training dataset with N samples. Let n_j be the number of samples in class j, then $N = \sum_{j=1}^K n_j$. Let $P_{\text{s}}(X, Y)$ define the underlying training (source) distribution and $P_{\text{t}}(X, Y)$ define the test (target) distribution. Generally, the class imbalance problem assumes that the test data has the same class conditional probability as the training data, i.e., $P_{\text{s}}(X|Y) = P_{\text{t}}(X|Y)$, while their class priors are different, i.e., $P_{\text{s}}(Y) \neq P_{\text{t}}(Y)$.

Besides, we consider an unlabelled auxiliary dataset $\mathcal{D}_{\text{out}}^{(\boldsymbol{x})} = \{\tilde{\boldsymbol{x}}_i\}_{i=1}^M \in \mathcal{X}$ consisting of M open-set instances, and we have $M \gg N$. These open-set instances are also known as OOD data as they are sampled from $P_{\text{out}}(X)$, which is disjoint from $P_{\text{s}}(X)$. In real-world scenarios, it is easy to obtain such auxiliary datasets, which are commonly used in OOD detection tasks. In what follows, we may assign each open-set instance $\tilde{\boldsymbol{x}}_i$ with a random label $\tilde{y}_i \in \mathcal{Y}$, which is independently sampled from an appropriate label distribu-

tion $P_{\text{out}}(Y)$ over $\mathcal{Y}$. We denote by $\mathcal{D}_{\text{out}} = \{(\tilde{\boldsymbol{x}}_i, y_i)\}_{i=1}^M$ the auxiliary dataset with randomly sampled noisy labels.

2.2. Theoretical Motivation

From a Bayesian perspective, the prediction of a Bayes classifier is generally made as follows:

$$y^* = \arg\max_{y \in \mathcal{Y}} P(y|\boldsymbol{x}) = \arg\max_{y \in \mathcal{Y}} P(\boldsymbol{x}|y)P(y), \quad (1)$$

where $P(\boldsymbol{x})$ and $P(y)$ represent $P(X = \boldsymbol{x})$ and $P(Y = y)$, respectively. Unfortunately, the predicted posterior probability in Eq. (1) becomes unreliable when there is a large discrepancy between the class priors of training and test distributions. Specifically, the class prior of the test distribution is usually balanced (i.e., a uniform distribution over labels), while the training dataset exhibits a long-tailed class distribution. Formally, we have $P_{\text{s}}(Y = i) \neq P_{\text{t}}(Y = i)$ for any class $i \in \mathcal{Y}$, and $P_{\text{s}}(Y = j) \ll P_{\text{t}}(Y = j)$ for a minority class $j \in \mathcal{Y}$.

To re-balance the class priors of the training dataset, existing re-sampling methods mostly augment the minority classes with extra *in-distribution* samples with precise labels, e.g., duplicated examples (Buda et al., 2018), synthetic generation (He et al., 2008), and interpolation (Chawla et al., 2002). However, due to their strict in-distribution constraints on data distribution and label quality, generating and collecting those samples are challenging and expensive, especially for minority classes. In this paper, we instead try to break through the constraints by demonstrating that OOD instances with noisy labels are actually useful for re-balancing the training dataset.

We start by presenting an intriguing fact in the following theorem (proof in Appendix A), which demonstrates that augmenting the training data with open-set instances and uniformly random labels can be non-toxic.

Theorem 2.1. *Assume that $P_{\text{out}}(Y)$ is the discrete uniform distribution over the label space $\mathcal{Y}$. Let the augmented dataset be $\mathcal{D}_{\text{mix}} = \mathcal{D}_{\text{train}} \cup \mathcal{D}_{\text{out}}$, and $P_{\text{mix}}(X, Y)$ be the underlying data distribution of $\mathcal{D}_{\text{mix}}$, then we have*

$$\arg\max_{y \in \mathcal{Y}} P_{\text{mix}}(\boldsymbol{x}|y)P_{\text{mix}}(y) = \arg\max_{y \in \mathcal{Y}} P_{\text{s}}(\boldsymbol{x}|y)P_{\text{s}}(y).$$

Theorem 2.1 indicates that, when the labels are uniformly sampled from the in-distribution label space, the prediction of the Bayes classifier is unchanged after augmenting the training dataset with the OOD instances[1]. In this way, the theoretical result motivates us to exploit the potential value of OOD instances to repair class imbalance.

---

[1]The discovery of the non-toxicity of OOD instances is not new and has been studied in ODNL (Wei et al., 2021). Our unique contribution here is to theoretically prove this property from a Bayesian perspective. We provide a detailed comparison for these two works in Subsection 2.3.
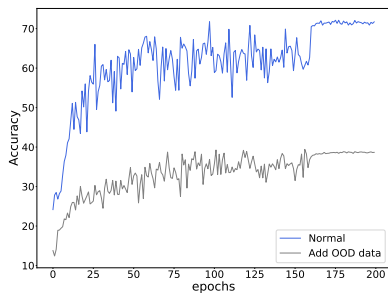
Figure 1. The test accuracy of models under different training epochs. Blue color denotes the model trained on long-tailed CIFAR-10 with ResNet-34, and gray color denotes the model trained on long-tailed CIFAR-10 and OOD data. We use instances from 300K Random Images (Hendrycks et al., 2019) as OOD data and label as a minority class—"9". The comparison implies that simply adding OOD data into training may downgrade the generalization performance.

Although using the uniform distribution as $P_{\text{out}}(Y)$ will not downgrade the Bayes classifier as shown above, the resulting classifier is not yet optimal, and its corresponding partition is still far away from the optimal decision boundary on the test distribution, since the label distribution $P_{\text{mix}}(Y)$ still remains largely imbalanced. In the following, we will show that a better classifier can be obtained by exploring the trade-off between re-balancing the label distribution and keeping non-toxicity.

2.3. Open-sampling

Motivated by the previous analysis, we propose to exploit the open-set auxiliary dataset to improve the generalization under class-imbalanced settings. With a proper label distribution, OOD instances with dynamic labels can be used to re-balance the class priors while retaining their non-toxicity. We start by giving the following definition.

Definition 2.2 (Complementary Distribution). *Complementary Distribution (CD) is a label distribution for the auxiliary dataset to re-balance the class priors of the original dataset. In particular, Minimum Complementary Distribution (MCD) is the complementary distribution that requires the smallest number of auxiliary instances to re-balance the original training set.*

Designing a proper Complementary Distribution to mitigate the class imbalance problem is a difficult problem that depends on the trade-off between re-balancing the class priors and keeping the non-toxicity of the added noisy labels. Intuitively, to re-balance the class priors, more OOD instances should be allocated into the minority classes than the majority classes. Meanwhile, the unequal number of OOD instances in different classes may shift the Bayes classifier.

The above conflict can be understood naturally through a simple case. We consider a binary classification task with
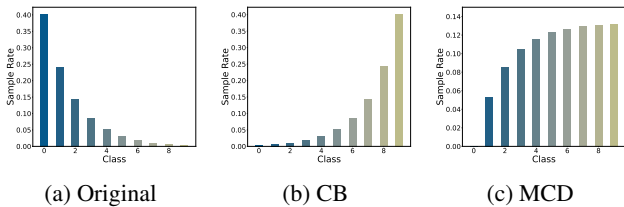

(a) Original   (b) CB   (c) MCD

Figure 2. An illustration of label distributions for long-tailed CIFAR-10 dataset with imbalance ratio 100. (2a) Original label prior, (2b) Class Balanced distribution, (2c) Minimum Complementary Distribution.

$K = 2$, and let the sample numbers of the two classes be n_1 and n_2, respectively. Without loss of generality, let $n_1 > n_2$. It is straightforward to verify that one of the optimal allocation to re-balance the class priors is simply appending $n_1 - n_2$ extra OOD instances into the minority class. However, in such a way, all OOD instances are assigned the same label, thereby downgrading the resulting classifier.

To provide a straightforward view, we show in Figure 1 the performance of model trained with the extra OOD data. Indeed, in the extreme case where all OOD instances are assigned the same label, their toxicity would be enlarged during training, leading to poor generalization performance.

Complementary Sampling Rate. To find a "sweet spot", we propose the following sampling rate that allows us to achieve a smooth transition from the MCD to the uniform distribution. We denote CD as Γ, MCD as Γ^m, and the complementary sampling rate for class j as Γ_j, then we have the following proposition.

Proposition 2.3 (Complementary Sampling Rate). $\Gamma_j = (\alpha - \beta_j)/(K \cdot \alpha - 1)$, where $\beta_j = \frac{n_j}{\sum_{i=1}^{K} n_i}$. Then, (i) $\sum_{i=1}^{K} \Gamma_i = 1$; (ii) $\Gamma = \Gamma^m$ if $\alpha = \max_j(\beta_j)$; (iii) $\Gamma_j \to 1/K$ as $\alpha \to \infty$.

The hyperparameter $\alpha \in \mathbb{R}^+ \geq \max_j(\beta_j)$ controls the trade-off bewteen the MCD and uniform distribution. As shown in Proposition 2.3, when $\alpha = \max_j(\beta_j)$, it recovers the MCD. With a larger value for the α, the label distribution of the auxiliary dataset would be closer to a uniform distribution. The proof of Proposition 2.3 is provided in Appendix B.

As shown in Figure 2, both the CB and MCD distributions exhibit an inverse relationship with the original label prior, i.e., the minority classes possess larger sampling rate than the majority classes. Compared with the CB distribution, our MCD distribution is flatter so that the instances would not be concentrated in a class. In such a manner, the harmfulness of open-set noisy labels would be alleviated to a large extent. The advantage of the MCD distribution is empirically supported in Section 4.

With the Complementary Sampling Rate, we can then build

an auxiliary dataset with OOD instances to re-balance the training dataset while retaining their non-toxicity.

Training Objective. To involve the sampled open-set noisy labels into the training, a natural idea is directly combining them with the original training dataset, and using the standard cross entropy as the training objective function:

$$\mathcal{L} = \mathbb{E}_{P_{\text{mix}}(X,Y)}[\ell(f(\boldsymbol{x};\boldsymbol{\theta}, y)]$$
$$= \mathbb{E}_{P_{\text{mix}}(X,Y)}[-\mathbf{e}^y \log f(\boldsymbol{x};\boldsymbol{\theta})],$$

where $\boldsymbol{e}^y \in \{0,1\}^K$ denotes the one-hot vector whose y-th entry of $\boldsymbol{e}^y$ is 1. In each epoch, the labels of samples from the auxiliary dataset could be updated following the Complementary Distribution Γ.

However, the naive combination would consume too much capacity of the network on fitting the open-set noisy labels, making it hard to converge, especially when the sample size of the auxiliary dataset is much larger than that of the original training dataset. To handle this issue, we propose to use the loss on the auxiliary dataset as a regularization term as shown below:

$$\mathcal{L}_{\text{reg}} = \mathbb{E}_{\widetilde{\boldsymbol{x}} \sim P_{\text{out}}(X)}\left[\ell\left(f(\widetilde{\boldsymbol{x}};\boldsymbol{\theta}), \widetilde{y}\right)\right],$$

where $\widetilde{y}$ is drawn from the Complementary Distribution Γ.

To further alleviate the over-fitting issue on the minority classes without sacrificing the performance on the majority classes, we explicitly introduce a class-dependent weighting factor ω_j based on the pre-defined Complementary Distribution. To make the total loss roughly in the same scale after applying ω_j, we normalize ω so that $\sum_{j=1}^{K} \omega_j = K$. Then, the regularization item becomes:

$$\mathcal{L}_{\text{reg}} = \mathbb{E}_{\widetilde{\boldsymbol{x}} \sim P_{\text{out}}(X)}\left[\omega_{\widetilde{y}} \cdot \ell\left(f(\widetilde{\boldsymbol{x}};\boldsymbol{\theta}), \widetilde{y}\right)\right],$$

where $\widetilde{y} \sim \Gamma$ and $\omega_{\widetilde{y}} = \Gamma_{\widetilde{y}} \cdot K$. Now, the final training objective function is given as follows:

$$\mathcal{L}_{\text{total}} = \mathbb{E}_{((\boldsymbol{x},y)\sim P_s(X,Y))}\left[\ell\left(f(\boldsymbol{x};\boldsymbol{\theta}), y\right)\right] \\ + \eta \cdot \mathbb{E}_{(\widetilde{\boldsymbol{x}})\sim P_{\text{out}}(X)}\left[\omega_{\widetilde{y}} \cdot \ell\left(f(\widetilde{\boldsymbol{x}};\boldsymbol{\theta}), \widetilde{y}\right)\right], \quad (2)$$

where η controls the strength of the regularization term. The corresponding algorithm is provided in Appendix C.

As a data re-balancing technique, Open-sampling is orthogonal to existing methods, including the training objective based methods (e.g., LDAM (Cao et al., 2019), Balanced Softmax (Ren et al., 2020)), Decoupled training methods (Kang et al., 2020), and self-supervised pre-trained methods (Yang & Xu, 2020). Our method can be easily incorporated into these algorithms to further improve their generalization performance. Given the original learning objective $\mathcal{L}_{\text{imb}}$ of the existing methods, we can formalize the final objective as:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{imb}} + \eta \cdot \mathcal{L}_{\text{reg}}. \quad (3)$$

Relation to ODNL. Recent work (Wei et al., 2021) show that open-set noisy labels could be applied to enhance the robustness against inherent noisy labels, which has some conceptual similarities to the proposed method in this work. Here, we summarize the main differences between ODNL (Wei et al., 2021) and our work.

1. Problem setting: ODNL focuses on improving the robustness against noisy labels while our work considers the problem of learning from long-tailed imbalanced datasets.

2. Technique: we proposed to sample the labels of OOD instances from the Complementary Distribution and add a weight factor to their losses, while they treats all the OOD instances equally by simply using a uniform distribution. In particular, ODNL can be seen as a special case of our method with a large value of α. As analyzed in Figures 4a, 4b, and 4c, our Open-sampling consistently outperforms the variant with a uniform distribution or a large value of α, which demonstrates the advantage of the proposed method.

3. Insight: ODNL aims to consume the extra representation capacity of neural networks to prevent over-fitting inherent noisy labels and show that their method helps the network converge to a flat minimum as SGD noises. In our work, OOD instances are applied to re-balance the label prior of the training dataset and the proposed method are shown to encourage the network to learn more separable representations.

3. Experiments

In this section, we evaluate our proposed method on long-tailed image classification datasets, including long-tailed CIFAR10/100 (Krizhevsky et al., 2009), CelebA-5 (Liu et al., 2015b; Kim et al., 2020), Places-LT (Zhou et al., 2017). Then we analyze the impact of η by sensitivity analysis. The ablation study for the class-dependent weight is provided in Appendix F. Finally, we conduct experiments to evaluate the performance of Open-sampling on OOD detection task under class-imbalanced setting. The datasets and implementation details are introduced in Appendix E.

3.1. Comparison methods

In this section, we verify that Open-sampling can boost the standard training and several state-of-the-art techniques by integrating Open-sampling with the following methods: 1) Standard: all the examples have the same weights; by default, we use standard cross-entropy loss. 2) SMOTE (Chawla et al., 2002): a variant of re-sampling with data augmentation. 3) CB-RW (Cui et al., 2019): training examples are re-weighted according to the inverse of the effective number of samples in each class, defined as $(1 - \beta^{n_i})/(1 - \beta)$.

Table 1. Test accuracy (%) of ResNet-32 on long-tailed CIFAR-10 and CIFAR-100 with various imbalance ratios. "†" indicates the reported results from (Kim et al., 2020). The bold indicates the improved results by integrating our regularization.

Dataset	Long-tailed CIFAR-10			Long-tailed CIFAR-100		
Imbalance Ratio	100	50	10	100	50	10
Standard	71.61 ± 0.21	77.30 ± 0.13	86.74 ± 0.41	37.59 ± 0.19	43.20 ± 0.30	56.44 ± 0.12
SMOTE †	71.50 ± 0.57	-	85.70 ± 0.25	34.00 ± 0.33	-	53.80 ± 0.93
CB-RW	72.57 ± 1.30	78.19 ± 1.79	87.18 ± 0.95	38.11 ± 0.78	43.26 ± 0.87	56.40 ± 0.40
CB-Focal	70.91 ± 0.39	77.71 ± 0.57	86.89 ± 0.21	37.84 ± 0.80	42.96 ± 0.77	56.09 ± 0.15
Ours	$\mathbf{77.62 \pm 0.28}$	$\mathbf{81.76 \pm 0.51}$	$\mathbf{89.38 \pm 0.46}$	$\mathbf{40.26 \pm 0.65}$	$\mathbf{44.77 \pm 0.25}$	$\mathbf{58.09 \pm 0.29}$
LDAM-RW	74.21 ± 0.61	78.86 ± 0.65	86.44 ± 0.78	29.02 ± 0.34	36.41 ± 0.84	54.23 ± 0.72
+ Ours	$\mathbf{75.19 \pm 0.34}$	$\mathbf{79.76 \pm 0.44}$	$\mathbf{87.28 \pm 0.61}$	$\mathbf{35.85 \pm 0.62}$	$\mathbf{42.18 \pm 0.82}$	$\mathbf{55.48 \pm 0.59}$
LDAM-DRW	78.08 ± 0.38	81.88 ± 0.44	87.49 ± 0.18	42.84 ± 0.25	47.13 ± 0.28	57.18 ± 0.47
+ Ours	$\mathbf{79.82 \pm 0.31}$	$\mathbf{82.22 \pm 0.45}$	$\mathbf{87.83 \pm 0.38}$	$\mathbf{44.07 \pm 0.75}$	$\mathbf{47.5 \pm 0.24}$	$\mathbf{57.43 \pm 0.31}$
Balanced Softmax	78.03 ± 0.28	81.63 ± 0.39	88.10 ± 0.32	42.11 ± 0.70	46.79 ± 0.24	58.06 ± 0.40
+ Ours	$\mathbf{79.05 \pm 0.20}$	$\mathbf{82.76 \pm 0.52}$	$\mathbf{88.89 \pm 0.21}$	$\mathbf{42.86 \pm 0.27}$	$\mathbf{47.28 \pm 0.58}$	$\mathbf{58.80 \pm 0.72}$
SSP	74.58 ± 0.16	79.20 ± 0.43	88.50 ± 0.24	43.00 ± 0.51	47.04 ± 0.60	59.08 ± 0.46
+ Ours	$\mathbf{79.38 \pm 0.65}$	$\mathbf{82.18 \pm 0.33}$	$\mathbf{88.80 \pm 0.43}$	$\mathbf{43.57 \pm 0.29}$	$\mathbf{48.66 \pm 0.57}$	$\mathbf{59.78 \pm 0.91}$

Table 2. Classification accuracy (%) on CIFAR-10 under the setting of Balance Softmax (Ren et al., 2020). The bold indicates the improved results by integrating our regularization. Baseline results are taken from Balance Softmax(Ren et al., 2020). " ⋆ " indicates decoupled training.

Imbalance Factor	200	100	10
Standard	71.2	77.4	90.0
CB-RW	72.5	78.6	90.1
LDAM-RW	73.6	78.9	90.3
Equalization Loss (Tan et al., 2020)	74.6	78.5	90.2
Balanced Softmax	79.0	83.1	90.9
Ours	80.6	83.6	90.6
cRT⋆ (Kang et al., 2020)	76.6	82.0	91.0
LWS⋆ (Kang et al., 2020)	78.1	83.7	**91.1**
Ours⋆	**81.1**	**84.9**	90.9

4) CB-Focal (Cui et al., 2019): the CB method is combined with Focal loss. 5) M2m (Kim et al., 2020): an over-sampling method with adversarial examples. 6) LDAM-RW (Cao et al., 2019): the method derives a generalization error bound for the imbalanced training and uses a margin-aware multi-class weighted cross entropy loss. 7) LDAM-DRW (Cao et al., 2019): the network is trained with LDAM loss and deferred re-balancing training. 8) Balanced Softmax (Ren et al., 2020): the method derives a Balanced Softmax function from the probabilistic perspective that explicitly models the test-time label distribution shift. 9) SSP (Yang & Xu, 2020): the method uses self-supervised learning to pre-train the network on the auxiliary dataset before standard training. Here, we do not expect vanilla Open-sampling to achieve state-of-the-art results compared with many complicated methods, our method can be still a promising option in the family of class-imbalanced learning methods, because it can outperform existing data re-balancing methods and improve existing state-of-the-art methods.

3.2. Main results

Results on long-tailed CIFAR. Extensive experiments are conducted on long-tailed CIFAR datasets with three different imbalance ratios: 10, 50, and 100. We use 300K Random Images[2] (Hendrycks et al., 2019) as the open-set auxiliary dataset. The test accuracy of ResNet-32 (He et al., 2016) on long-tailed CIFAR datasets are reported in Table 1. The results show that our method can achieve impressive improvements on the standard training method. Especially for long-tailed CIFAR-10 with imbalance ratio 100, an extreme imbalance case, the vanilla Open-sampling can significantly outperform the standard baseline by 8.39%. Besides, incorporating our method into existing state-of-the-art methods can consistently improve their performance under various imbalance ratios.

We also conduct experiments on long-tailed CIFAR-10 following the training setting of Balanced Softmax (Ren et al., 2020) and present the results in Table 2. Under this setting, our proposed method achieves the best performance among the end-to-end methods and could also improve the performance of decoupled training (Kang et al., 2020).

To further clarify the influence of η, we present a sensitivity analysis on long-tailed CIFAR-10 dataset with imbalance ratio 100 in Figure 3. We highlight the differences in the trend of test accuracy after the decay of learning rate at the 160th epoch. From the figure, we can observe that with a proper value of η like 1.5, the generalization performance can be largely improved by our proposed regularization. The result verifies that our regularization is an effective method to improve the generalization performance in long-tailed imbalanced learning.
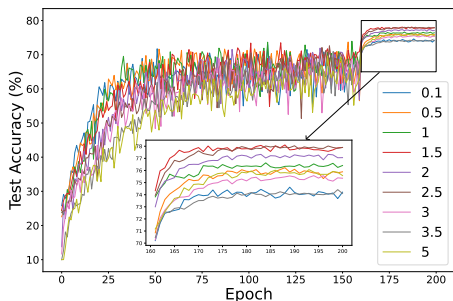
---

[2]The dataset is published on https://github.com/hendrycks/outlier-exposure.

Figure 3. Results of sensitivity analysis on long-tailed CIFAR-10 with various values for η.

Results on CelebA-5. We further verify the effectiveness of our method on real-world class-imbalanced datasets. The CelebA dataset has a long-tailed label distribution and the test set has a balanced label distribution. We use 300K Random Images (Hendrycks et al., 2019) as the open-set auxiliary dataset. Table 3 summarizes test accuracy on the CelebA-5 dataset. In particular, the proposed method outperforms existing data-rebalancing methods and is able to consistently improve the existing state-of-the-art methods in test accuracy.

Results on Places-LT. For Places-LT, we follow the protocol of Decoupled training (Kang et al., 2020) and start from a ResNet-152 (He et al., 2016) backbone pre-trained on the ImageNet dataset (Russakovsky et al., 2015). We fine-tune the backbone with Instance-balanced sampling for representation learning and then re-train the classifier with our proposed algorithm as decoupled training. Here, we use Places-extra69 (Zhou et al., 2017) as the open-set auxiliary dataset. Table 4 summarizes test accuracy on the Places-LT dataset. The proposed method achieves the best overall performance, show that our method and decoupled training scheme are two orthogonal components and Open-sampling is applicable for large-scale datasets.

OOD detection under class imbalance. Out-of-distribution (OOD) detection is an essential problem for the deployment of deep learning especially in safety-critical applications (Hendrycks et al., 2019; Liu et al., 2020; Liang et al., 2018; Yang et al., 2021) and it would be more difficult when the training dataset exhibits long-tailed distributions. Outlier Exposure (OE) (Hendrycks et al., 2019) is a commonly-used method to improve anomaly detection performance with auxiliary dataset. Specifically, the training objective of OE under class imbalance is $\mathbb{E}_{(x,y)\sim\mathcal{D}_{\text{train}}}[-\log f_y(x)]+\lambda\mathbb{E}_{x\sim\mathcal{D}_{\text{out}}}[H(P(Y);f(x))]$, where H is the cross entropy and $P(Y)$ is the label prior of the original training dataset. Following OE (Hendrycks et al., 2019), we consider the maximum softmax probability (MSP) baseline (Hendrycks & Gimpel, 2016). Here, we conduct experiments on long-tailed CIFAR10 with imbalance rate 100 to verify the advantage of Open-sampling on OOD

Table 3. Classification accuracy (%) on CelebA-5 with ResNet-32. "†" indicates the reported results are obtained from (Kim et al., 2020). The shadow indicates the improved results.

Method	Accuracy	Method	Accuracy	Method	Accuracy
Standard	72.7	M2m †	75.6	LDAM-DRW	74.5
SMOTE †	72.8	**Ours**	**76.8**	**LDAM-DRW + Ours**	**76.9**
CB-RW	73.6	LDAM-RW	73.1	Balanced Softmax	76.4
CB-Focal	74.2	**LDAM + Ours**	**75.8**	**Balanced Softmax + Ours**	**78.6**

Table 4. Top-1 accuracy (%) on Places-LT with an ImageNet pre-trained ResNet-152. Baseline results are taken from original papers. "DT" indicates decoupled training.

Method	Many	Medium	Few	Overall
Lifted Loss (Oh Song et al., 2016)	41.1	35.4	24.0	35.2
Focal Loss	41.1	34.8	22.4	34.6
Range Loss (Zhang et al., 2017)	41.1	35.4	23.2	35.1
OLTR (Liu et al., 2019b)	44.7	37.0	25.3	35.9
cRT⋆ (Kang et al., 2020)	42.1	37.6	24.9	36.7
LWS⋆ (Kang et al., 2020)	40.6	39.1	28.6	37.6
Ours⋆	42.9	**39.3**	26.8	**38.2**

Table 5. OOD detection performance comparison on long-tailed CIFAR-10. All values are percentages and are averaged over the ten test datasets described in Appendix E. "↑" indicates larger values are better, and "↓" indicates smaller values are better. Bold numbers are superior results. Detailed results for each OOD test dataset can be found in Appendix F.

Method	Test Accuracy ↑	FPR95 ↓	AUROC ↑	AUPR ↑
MSP	71.83	56.1	75.2	32.71
OE	66.74	32.38	84.15	36.86
Ours	**77.62**	**20.68**	92.40	58.38
Ours ($\alpha = 5$)	75.16	22.13	**94.26**	**75.91**

detection under class imbalance. Table 5 presents the test accuracy and the average performance over the three types of noises and seven OOD test datasets. We can observe that our method achieves impressive improvement on both the test accuracy and the detection performance.

4. A Closer Look at Open-sampling under Class Imbalance

To provide a comprehensive understanding of the proposed method, we conduct a set of analyses in this section. Firstly, we compare our defined distribution with several alternative distributions to show the advantage of the Complementary Distribution in our method. Secondly, the effect of α in our method is thoroughly analyzed by empirical studies. Then, we present a guideline about how to choose or collect a suitable open-set auxiliary dataset for long-tailed imbalanced learning. Finally, we analyze the effect of the proposed method through the lens of decision boundaries.

The advantage of Complementary Distribution. In the proposed method, the labels of OOD instances from the aux-

(a) Label distribution.

(b) Label distribution.

(c) Alpha.

(d) Auxiliary dataset.

(e) Sample size.
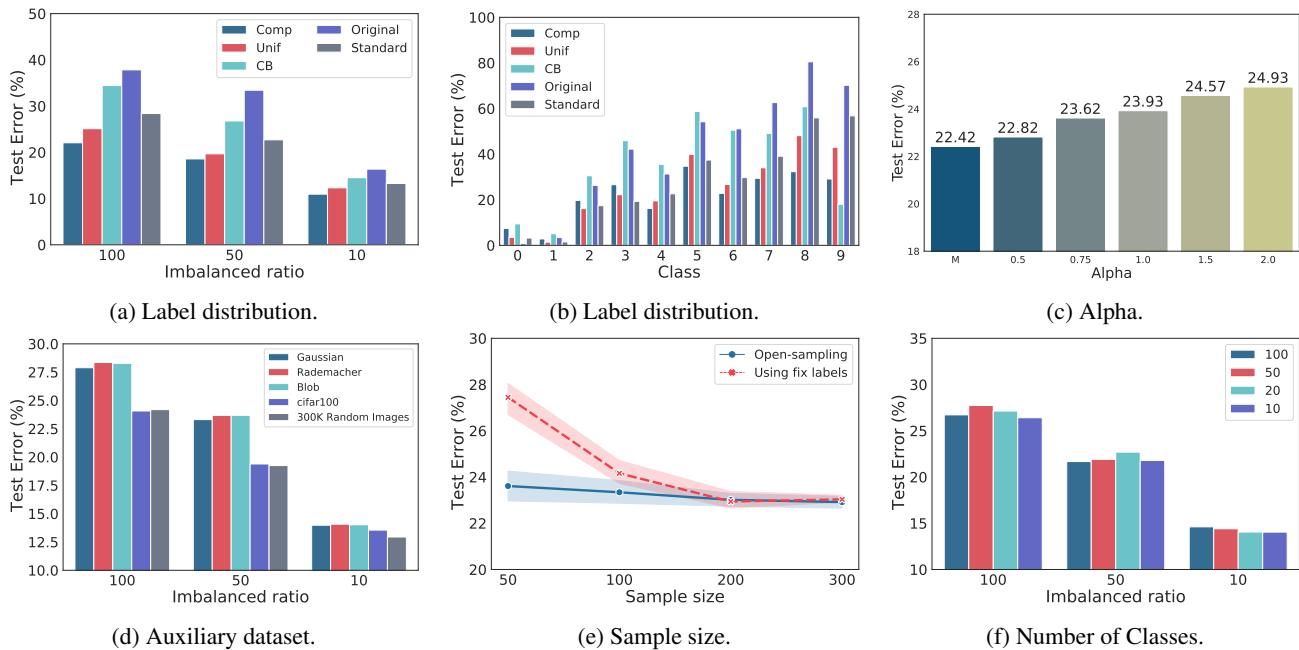
(f) Number of Classes.

Figure 4. Analytical experiments of Open-sampling on long-tailed CIFAR-10: (4a)(4b) with various label distributions, (4c) with various values of the α, (4d) with various open-set auxiliary datasets, including simulated noise datasets and real-world datasets. All the datasets contain 50,000 instances. (4e) with various sample sizes (K) of the auxiliary dataset, (4f) with various number of classes in the auxiliary datasets that are randomly sampled from CIFAR-100. Experiments in (4b), (4c), and (4e) are conducted under the imbalance ratio 100. The y-axis represents the test error in all the figures. "Standard" denotes the baseline with standard cross-entropy loss.

iliary dataset are sampled from a random label distribution. For the random label distribution, we defined a Complementary Distribution in Proposition 2.3 and also presented several commonly used distributions, including uniform distribution (Unif), class balanced distribution (CB) (Cui et al., 2019), and the original class priors of the training dataset (Original). Here, we conduct experiments to compare the performance of the Open-sampling variants with different label distributions. The results in Figure 4a show that using the Complementary Distribution consistently achieves the best performance on the test set. In particular, using the uniform distribution can also improve the generalization performance while both CB and the original class priors deteriorate the performance of the neural networks.

To further understand why CB is not a good choice in our method, we present the per-class top-1 error on long-tailed CIFAR-10 in Figure 4b. Although using the CB distribution can achieve better performance on the smallest class, it downgrades the generalization performance on the other classes. The reason is that the CB distribution is far away from the uniform distribution, thereby introducing too much noise to the Bayes classifier. Different from CB, the Complementary Distribution is closer to a uniform distribution, making it achievable to re-balance the class priors while almost keeping non-toxicity of the open-set noisy labels.

Here, we also show the effect of α in Figure 4c. As

analyzed in Section 2, the larger the value of α is, the Complementary Distribution tends to be closer to a uniform distribution. Here, "M" denotes the default value of $\alpha = (\max_j \beta_j + \min_j \beta_j)$. Additionally, the MCD variant achieves 22.51% on the test error. From Figure 4c, the test error presented a slightly upward trend with the increasing of the value of α. The results verified that it is necessary to adopt a complementary distribution, instead of simply using a uniform distribution.

The choices of open-set auxiliary dataset. With proper label distribution, can any OOD dataset be used to improve Long-tail imbalanced learning? In Figure 4d, we show the test error on long-tailed CIFAR-10 with imbalance ratio 100 using Open-sampling with different auxiliary datasets. We can observe that using simple noises like Gaussian noise could not achieve similar performance than those of real-world datasets, *e.g.*, CIFAR100 and 300K Random Images (Hendrycks et al., 2019).

The sample size is another important factor for the open-set auxiliary dataset. In Figure 4e, we show that the performance of Open-sampling would be slightly better with a larger sample size of the auxiliary dataset. It is worth noting that the phenomenon is not consistent with that of (Wei et al., 2021), where using larger open-set auxiliary dataset could not improve the performance on learning with noisy labels. Here, we also conduct experiments using a variant
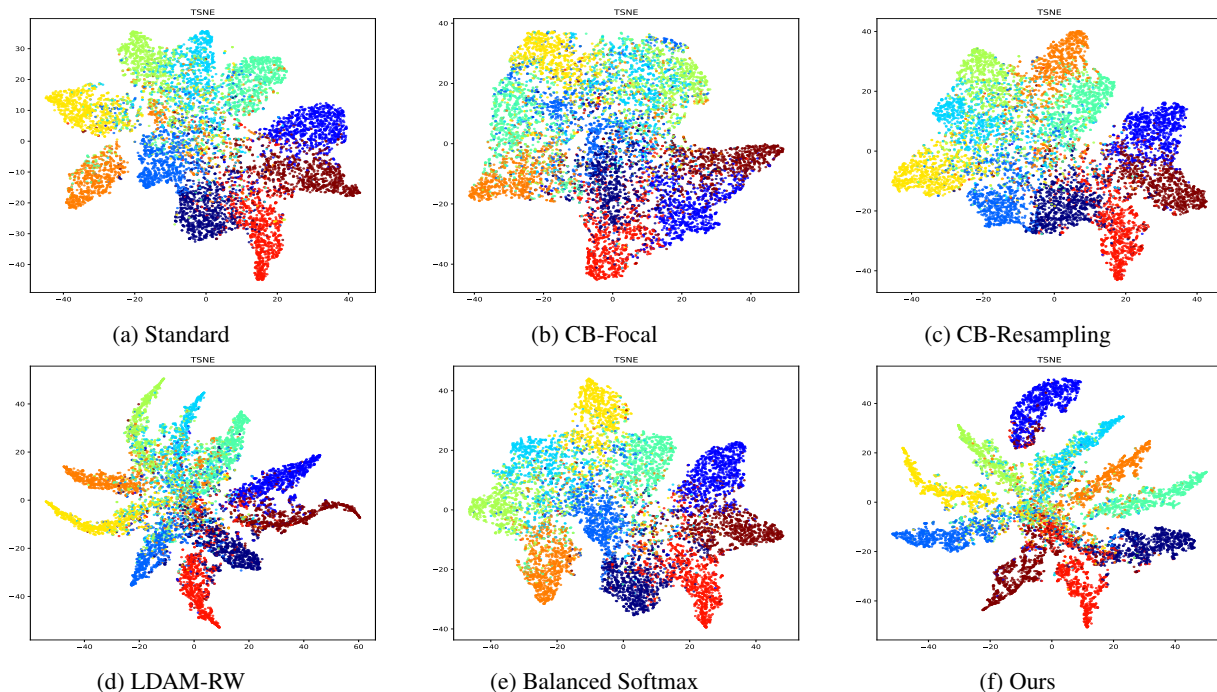
Figure 5. t-SNE visualization of test set on long-tailed CIFAR-10 with imbalance ratio 100. We can observe that LDAM and our method appear to learn more separable representations than Standard training and the other algorithms.

of Open-sampling with fixed labels for OOD instances. We can observe that the variant performs poorly with a small auxiliary dataset, but the performance can be significantly improved by increasing the sample size of auxiliary dataset. In particular, the variant could achieve nearly the same improvement as the proposed Open-sampling method, with a large enough sample size.

In addition, we also find that the diversity of the open-set auxiliary dataset is unimportant for the effectiveness of Open-sampling. We conduct experiments on long-tailed CIFAR-10 with imbalance ratio 100 and use the subset of CIFAR100 with different number of classes as the open-set auxiliary dataset. The sample size of these subsets are fixed as 5,000. The results of using different subsets are almost the same, achieving 73.28% test accuracy. The results are consistent with that in Figure 4d, where using CIFAR100 as open-set auxiliary dataset can achieve almost the same improvements as 300K Random Images (Hendrycks et al., 2019).

The effect of Open-sampling as regularization. In addition to re-balancing the class prior, what is the effect of Open-sampling as a regularization method? To gain additional insight, we look at the t-SNE projection of the learnt representations for different algorithms in Figure 5. For each method, the projection is performed over test data. The figures show that the decision boundaries of Open-sampling and LDAM (Cao et al., 2019) are much clearer than those of

the other methods. The phenomenon illustrates that Open-sampling can encourage the minority classes to have a larger margin, which is similar to the effect of the LDAM loss (Cao et al., 2019). As shown in Section 3, our method could still boost the performance of the LDAM method, which demonstrates the differences between our works.

5. Related Work

Re-sampling. Re-sampling methods aims to re-balance the class priors of the training dataset. Under-sampling methods remove examples from the majority classes, which is infeasible under extremely data imbalanced settings (He & Garcia, 2009; Japkowicz & Stephen, 2002). The over-sampling method adds repeated samples for the minority classes, usually causing over-fitting to the minority classes (Buda et al., 2018; Byrd & Lipton, 2019; Shen et al., 2016). Some methods utilize synthesized in-distribution samples to alleviate the over-fitting issue but introduce extra noise (Chawla et al., 2002; He et al., 2008; Kim et al., 2020). In contrast to in-distribution samples used in existing methods, our approach exploits OOD instances to re-balance the class priors of the training dataset.

Re-weighting. Re-weighting methods propose to assign adaptive weights for different classes or samples. Generally, the vanilla scheme re-weights classes proportionally to the inverse of their frequency (Huang et al., 2016). Focal loss (Lin et al., 2017) assigns low weights to the well-classified

examples. Class-balanced loss (Cui et al., 2019) proposes to re-weight by the inverse effective number of samples. However, these re-weighting methods tend to make the optimization of DNNs difficult under extremely data imbalanced settings (Wang et al., 2017; Mikolov et al., 2013).

Other methods for long-tailed datasets. In addition to the data re-balancing approaches, some other solutions are also applied for class-imbalanced learning, including transfer-learning based methods (Yin et al., 2019; Jamal et al., 2020), two-stage training methods(Kang et al., 2020; Zhong et al., 2021; Zhang et al., 2021; Liu et al., 2019b), training objective based methods (Cao et al., 2019; Ren et al., 2020; Hong et al., 2021), expert methods (Wang et al., 2021b) and self-supervised learning methods (Yang & Xu, 2020; Cui et al., 2021). For example, transfer-learning based methods addressed the class-imbalanced issue by transferring features learned from head classes with abundant training instances to under-represented tail classes (Yin et al., 2019; Jamal et al., 2020). Two-stage training methods proposed to apply decoupled training where the classifier is re-balanced during the fine-tuning stage (Kang et al., 2020; Zhong et al., 2021; Zhang et al., 2021; Liu et al., 2019b). Generally, our proposed method is complementary to these existing methods and our method can further improve their performance, which are explicitly shown in our experiments.

Utilizing auxiliary dataset. In the deep learning community, auxiliary dataset is utilized in various contexts, e.g., adversarial machine learning (Madry et al., 2018; Lee et al., 2021) and weakly supervised learning (Wei et al., 2020a;b; Zhu et al., 2021a; Cheng et al., 2021; Wang et al., 2021a; Zhu et al., 2021b; Wei et al., 2022b; Zhu et al., 2022a;b). For example, OE (Hendrycks et al., 2019) regularized the network to give conservative predictions on OOD instances. OAT (Lee et al., 2021) utilized OOD instances to improve generalization in adversarial training. ODNL (Wei et al., 2021) proposed to use open-set auxiliary dataset to prevent the model from over-fitting inherent noisy labels. In long-tailed imbalanced learning, recent work (Yang & Xu, 2020) introduced unlabeled-in-distribution data to compensate for the lack of training samples. Another work (Su et al., 2021) adopted semi-supervised learning techniques to improve performance on long-tailed datasets with out-of-class images from related classes, which are different from out-of-distribution data in our method. To the best of our knowledge, we are the first to explore the benefits of *out-of-distribution data* in learning with long-tailed datasets.

OOD detection. OOD detection is an essential building block for safely deploying machine learning models in the open world (Hendrycks & Gimpel, 2016; Hendrycks et al., 2019; Liu et al., 2020; Wei et al., 2022a). A common baseline for OOD detection is the softmax confidence score (Hendrycks & Gimpel, 2016). It has been theoretically shown that neural networks with ReLU activation can produce arbitrarily high softmax confidence for OOD inputs (Hein et al., 2019). To improve the performance, previous methods have explored using artificially synthesized data from GANs (Goodfellow et al., 2014) or unlabeled data (Hendrycks et al., 2019; Lee et al., 2017) as auxiliary OOD training data. Energy scores are shown to be better for distinguishing in- and out-of-distribution samples than softmax scores (Liu et al., 2020). As a side effect, our method could also achieve superior performance in OOD detection tasks even if the labels of training dataset are noisy.

6. Conclusion

In this paper, we propose a simple yet effective method termed Open-sampling, by introducing OOD instances to re-balance the class priors of the training dataset. To the best of our knowledge, our method is the first to utilize OOD instances in long-tailed learning. We show that our method not only re-balances the training dataset, but also promotes the neural network to learn more separable representations. Besides, we also present a guideline about the open-set auxiliary datasets: the realism and sample size are more important than the diversity. Extensive experiments demonstrate that our proposed method can enhance the performance of existing state-of-the-art methods, and also achieves strong OOD detection performance under class-imbalanced setting. Overall, our method enables to exploit OOD data for long-tail learning and we expect that our insights inspire future research to further explore the benefits of auxiliary datasets.

Acknowledgements

References

Buda, M., Maki, A., and Mazurowski, M. A. A systematic study of the class imbalance problem in convolutional neural networks. *Neural Networks*, 106:249–259, 2018.

Byrd, J. and Lipton, Z. What is the effect of importance weighting in deep learning? In *International Conference on Machine Learning*, pp. 872–881. PMLR, 2019.

Cao, K., Wei, C., Gaidon, A., Arechiga, N., and Ma, T. Learning imbalanced datasets with label-distribution-aware margin loss. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

Chawla, N. V., Bowyer, K. W., Hall, L. O., and Kegelmeyer, W. P. Smote: Synthetic minority over-sampling technique.

Journal of Artificial Intelligence Research, 16:321–357, 2002.

Cheng, H., Zhu, Z., Li, X., Gong, Y., Sun, X., and Liu, Y. Learning with instance-dependent label noise: a sample sieve approach. In *International Conference on Learning Representations*, 2021.

Cimpoi, M., Maji, S., Kokkinos, I., Mohamed, S., and Vedaldi, A. Describing textures in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3606–3613, 2014.

Cui, J., Zhong, Z., Liu, S., Yu, B., and Jia, J. Parametric contrastive learning. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 715–724, 2021.

Cui, Y., Jia, M., Lin, T.-Y., Song, Y., and Belongie, S. Class-balanced loss based on effective number of samples. In *Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition*, pp. 9268–9277, 2019.

Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. Generative adversarial networks. *arXiv preprint arXiv:1406.2661*, 2014.

Goyal, P., Dollár, P., Girshick, R., Noordhuis, P., Wesolowski, L., Kyrola, A., Tulloch, A., Jia, Y., and He, K. Accurate, large minibatch sgd: Training imagenet in 1 hour. *arXiv preprint arXiv:1706.02677*, 2017.

He, H. and Garcia, E. A. Learning from imbalanced data. *IEEE Transactions on Knowledge and Data Engineering*, 21(9):1263–1284, 2009.

He, H., Bai, Y., Garcia, E. A., and Li, S. Adasyn: Adaptive synthetic sampling approach for imbalanced learning. In *2008 IEEE International Joint Conference on Neural Networks*, pp. 1322–1328, 2008.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pp. 770–778, 2016.

Hein, M., Andriushchenko, M., and Bitterwolf, J. Why relu networks yield high-confidence predictions far away from the training data and how to mitigate the problem. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 41–50, 2019.

Hendrycks, D. and Gimpel, K. A baseline for detecting misclassified and out-of-distribution examples in neural networks. *arXiv preprint arXiv:1610.02136*, 2016.

Hendrycks, D., Mazeika, M., and Dietterich, T. Deep anomaly detection with outlier exposure. *Proceedings of the International Conference on Learning Representations*, 2019.

Hong, Y., Han, S., Choi, K., Seo, S., Kim, B., and Chang, B. Disentangling label distribution for long-tailed visual recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6626–6636, 2021.

Huang, C., Li, Y., Loy, C. C., and Tang, X. Learning deep representation for imbalanced classification. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pp. 5375–5384, 2016.

Jamal, M. A., Brown, M., Yang, M.-H., Wang, L., and Gong, B. Rethinking class-balanced methods for long-tailed visual recognition from a domain adaptation perspective. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7610–7619, 2020.

Japkowicz, N. and Stephen, S. The class imbalance problem: A systematic study. *Intelligent Data Analysis*, 6(5):429–449, 2002.

Kang, B., Xie, S., Rohrbach, M., Yan, Z., Gordo, A., Feng, J., and Kalantidis, Y. Decoupling representation and classifier for long-tailed recognition. In *Eighth International Conference on Learning Representations (ICLR)*, 2020.

Kim, J., Jeong, J., and Shin, J. M2m: Imbalanced classification via major-to-minor translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13896–13905, 2020.

Komodakis, N. and Gidaris, S. Unsupervised representation learning by predicting image rotations. In *International Conference on Learning Representations*, 2018.

Krizhevsky, A., Hinton, G., et al. Learning multiple layers of features from tiny images. 2009.

Lee, K., Lee, H., Lee, K., and Shin, J. Training confidence-calibrated classifiers for detecting out-of-distribution samples. *arXiv preprint arXiv:1711.09325*, 2017.

Lee, S., Park, C., Lee, H., Yi, J., Lee, J., and Yoon, S. Removing undesirable feature contributions using out-of-distribution data. *arXiv preprint arXiv:2101.06639*, 2021.

Liang, S., Li, Y., and Srikant, R. Enhancing the reliability of out-of-distribution image detection in neural networks. *International Conference on Learning Representations (ICLR)*, 2018.

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., and Zitnick, C. L. Microsoft coco: Common objects in context. In *European Conference on Computer Vision*, pp. 740–755. Springer, 2014.

Lin, T.-Y., Goyal, P., Girshick, R., He, K., and Dollár, P. Focal loss for dense object detection. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 2980–2988, 2017.

Liu, W., Wang, X., Owens, J., and Li, Y. Energy-based out-of-distribution detection. *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

Liu, Z., Luo, P., Wang, X., and Tang, X. Deep learning face attributes in the wild. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 3730–3738, 2015a.

Liu, Z., Luo, P., Wang, X., and Tang, X. Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision*, December 2015b.

Liu, Z., Miao, Z., Zhan, X., Wang, J., Gong, B., and Yu, S. X. Large-scale long-tailed recognition in an open world. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2537–2546, 2019a.

Liu, Z., Miao, Z., Zhan, X., Wang, J., Gong, B., and Yu, S. X. Large-scale long-tailed recognition in an open world. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2019b.

Loshchilov, I. and Hutter, F. Sgdr: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983*, 2016.

Madry, A., Makelov, A., Schmidt, L., Tsipras, D., and Vladu, A. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations*, 2018.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 3111–3119, 2013.

Netzer, Y., Wang, T., Coates, A., Bissacco, A., Wu, B., and Ng, A. Y. Reading digits in natural images with unsupervised feature learning. 2011.

Oh Song, H., Xiang, Y., Jegelka, S., and Savarese, S. Deep metric learning via lifted structured feature embedding. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pp. 4004–4012, 2016.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in Neural Information Processing Systems (NeurIPS)*, 32:8026–8037, 2019.

Ren, J., Yu, C., Sheng, S., Ma, X., Zhao, H., Yi, S., and Li, H. Balanced meta-softmax for long-tailed visual recognition. In *Advances in Neural Information Processing Systems (NeurIPS)*, Dec 2020.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A. C., and Fei-Fei, L. ImageNet large scale visual recognition challenge. *International Journal of Computer Vision*, 115(3):211–252, 2015.

Shen, L., Lin, Z., and Huang, Q. Relay backpropagation for effective learning of deep convolutional neural networks. In *European Conference on Computer Vision*, pp. 467–482. Springer, 2016.

Su, J.-C., Cheng, Z., and Maji, S. A realistic evaluation of semi-supervised learning for fine-grained classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12966–12975, 2021.

Tan, J., Wang, C., Li, B., Li, Q., Ouyang, W., Yin, C., and Yan, J. Equalization loss for long-tailed object recognition. In *Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition*, pp. 11662–11671, 2020.

Van Horn, G., Mac Aodha, O., Song, Y., Cui, Y., Sun, C., Shepard, A., Adam, H., Perona, P., and Belongie, S. The inaturalist species classification and detection dataset. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pp. 8769–8778, 2018.

Wang, J., Guo, H., Zhu, Z., and Liu, Y. Policy learning using weak supervision. *Advances in Neural Information Processing Systems*, 34, 2021a.

Wang, X., Lian, L., Miao, Z., Liu, Z., and Yu, S. Long-tailed recognition by routing diverse distribution-aware experts. In *International Conference on Learning Representations*, 2021b.

Wang, Y.-X., Ramanan, D., and Hebert, M. Learning to model the tail. In *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 7032–7042, 2017.

Wei, H., Feng, L., Chen, X., and An, B. Combating noisy labels by agreement: A joint training method with co-regularization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13726–13735, 2020a.

Wei, H., Feng, L., Wang, R., and An, B. Metainfonet: Learning task-guided information for sample reweighting. *arXiv preprint arXiv:2012.05273*, 2020b.

Wei, H., Tao, L., Xie, R., and An, B. Open-set label noise can improve robustness against inherent label noise. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

Wei, H., Renchunzi, X., Hao, C., Feng, L., An, B., and Yixuan, L. Mitigating neural network overconfidence with logit normalization. In *International Conference on Machine Learning (ICML)*. PMLR, 2022a.

Wei, J., Zhu, Z., Cheng, H., Liu, T., Niu, G., and Liu, Y. Learning with noisy labels revisited: A study using real-world human annotations. In *International Conference on Learning Representations*, 2022b.

Xiao, J., Hays, J., Ehinger, K. A., Oliva, A., and Torralba, A. Sun database: Large-scale scene recognition from abbey to zoo. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3485–3492. IEEE, 2010.

Xu, P., Ehinger, K. A., Zhang, Y., Finkelstein, A., Kulkarni, S. R., and Xiao, J. Turkergaze: Crowdsourcing saliency with webcam based eye tracking. *arXiv preprint arXiv:1504.06755*, 2015.

Yang, J., Zhou, K., Li, Y., and Liu, Z. Generalized out-of-distribution detection: A survey. *arXiv preprint arXiv:2110.11334*, 2021.

Yang, Y. and Xu, Z. Rethinking the value of labels for improving class-imbalanced learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

Yin, X., Yu, X., Sohn, K., Liu, X., and Chandraker, M. Feature transfer learning for face recognition with under-represented data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5704–5713, 2019.

Yu, F., Seff, A., Zhang, Y., Song, S., Funkhouser, T., and Xiao, J. Lsun: Construction of a large-scale image dataset using deep learning with humans in the loop. *arXiv preprint arXiv:1506.03365*, 2015.

Zhang, S., Li, Z., Yan, S., He, X., and Sun, J. Distribution alignment: A unified framework for long-tail visual recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2361–2370, 2021.

Zhang, X., Fang, Z., Wen, Y., Li, Z., and Qiao, Y. Range loss for deep face recognition with long-tailed training data. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 5409–5418, 2017.

Zhong, Z., Cui, J., Liu, S., and Jia, J. Improving calibration for long-tailed recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2021.

Zhou, B., Lapedriza, A., Khosla, A., Oliva, A., and Torralba, A. Places: A 10 million image database for scene recognition. *IEEE transactions on Pattern Analysis and Machine Intelligence*, 40(6):1452–1464, 2017.

Zhou, B., Cui, Q., Wei, X.-S., and Chen, Z.-M. 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*, pp. 9719–9728, 2020.

Zhu, Z., Liu, T., and Liu, Y. A second-order approach to learning with instance-dependent label noise. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2021a.

Zhu, Z., Song, Y., and Liu, Y. Clusterability as an alternative to anchor points when learning with noisy labels. In *Proceedings of the 38th International Conference on Machine Learning*, 2021b.

Zhu, Z., Dong, Z., and Liu, Y. Detecting corrupted labels without training a model to predict. In *International Conference on Machine Learning (ICML)*, 2022a.

Zhu, Z., Wang, J., and Liu, Y. Beyond images: Label noise transition matrix estimation for tasks with lower-quality features. In *International Conference on Machine Learning (ICML)*, 2022b.

A. Proof of Theorem 2.1

Since $\mathcal{D}_{\mathrm{mix}} = \mathcal{D}_{\mathrm{train}} \cup \mathcal{D}_{\mathrm{out}}$, the underlying data distribution of $\mathcal{D}_{\mathrm{mix}}$ will be a linear combination of the training distribution $P_{\mathrm{s}}(X, Y)$ and the OOD distribution $P_{\mathrm{out}}(X, Y)$:

$$P_{\mathrm{mix}}(\boldsymbol{x}, y) = \frac{N}{M + N} P_{\mathrm{s}}(\boldsymbol{x}, y) + \frac{M}{M + N} P_{\mathrm{out}}(\boldsymbol{x}, y), \tag{4}$$

where N is the size of $\mathcal{D}_{\mathrm{train}}$ and M is the size of $\mathcal{D}_{\mathrm{out}}$.

By the virtue of Bayes' theorem, we have

$$
\begin{aligned}
&P_{\mathrm{mix}}(\boldsymbol{x}|y) P_{\mathrm{mix}}(y) \\
=&\frac{N}{M + N} P_{\mathrm{s}}(\boldsymbol{x}|y) P_{\mathrm{s}}(y) + \frac{M}{M + N} P_{\mathrm{out}}(\boldsymbol{x}|y) P_{\mathrm{out}}(y) \\
=&\frac{N}{M + N} P_{\mathrm{s}}(\boldsymbol{x}|y) P_{\mathrm{s}}(y) + \frac{M}{M + N} P_{\mathrm{out}}(\boldsymbol{x}) P_{\mathrm{out}}(y) \\
=&\frac{N}{M + N} P_{\mathrm{s}}(\boldsymbol{x}|y) P_{\mathrm{s}}(\boldsymbol{x}, y) + \frac{1}{K} \cdot \frac{M}{M + N} P_{\mathrm{out}}(\boldsymbol{x}),
\end{aligned} \tag{5}
$$

where the second equality follows the fact that $P_{\mathrm{mix}}(\boldsymbol{x}|y) = P_{\mathrm{mix}}(\boldsymbol{x})$ since the label y is independent of the instance x for the OOD data, and the third equality is simply the fact that $P_{\mathrm{out}}(y) = 1/K$.

Then, by taking the maximum of both sides, we have

$$
\begin{aligned}
&\arg\max_{y \in \mathcal{Y}} P_{\mathrm{mix}}(\boldsymbol{x}|y) P_{\mathrm{mix}}(y) \\
=&\arg\max_{y \in \mathcal{Y}} \left\{ \frac{N}{M + N} P_{\mathrm{s}}(\boldsymbol{x}|y) P_{\mathrm{s}}(\boldsymbol{x}, y) + \frac{M/K}{M + N} P_{\mathrm{out}}(\boldsymbol{x}) \right\} \\
=&\arg\max_{y \in \mathcal{Y}} \frac{N}{M + N} P_{\mathrm{s}}(\boldsymbol{x}|y) P_{\mathrm{s}}(\boldsymbol{x}, y) \\
=&\arg\max_{y \in \mathcal{Y}} P_{\mathrm{s}}(\boldsymbol{x}|y) P_{\mathrm{s}}(\boldsymbol{x}, y).
\end{aligned} \tag{6}
$$

Thus Theorem 2.1 is proved. $\square$

B. Proof of Proposition 2.3

By definition, $\sum_{i=1}^{K} \Gamma_i = 1$ naturally holds for any α.

If $\alpha = \max_j(\beta_j)$, then $\Gamma_j = (\max_j(\beta_j) - \beta_j)/(K \cdot \max_j(\beta_j) - 1)$. In particular, for $k = \arg\max_i(\beta_i)$, we have $\Gamma_k = 0$.

In the case of $\alpha \to \infty$, let us denote $f(\alpha) = \alpha - \beta_j$ and $g(\alpha) = K \cdot \alpha - 1$. Since $\lim_{\alpha \to \infty} f(\alpha) = \lim_{\alpha \to \infty} g(\alpha) = \infty$, $g'(\alpha) = K \neq 0$, and $\lim_{\alpha \to \infty} f'(\alpha)/g'(\alpha) = 1/K$ exists, using L'Hôpital's rule, we have:

$$\lim_{\alpha \to \infty} \Gamma_j = \lim_{\alpha \to \infty} \frac{f(\alpha)}{g(\alpha)} = \lim_{\alpha \to \infty} \frac{f'(\alpha)}{g'(\alpha)} = 1/K$$

Thus Proposition 2.3 is proved. $\square$

C. Algorithm Details

The details of Open-sampling are provided below:

D. Datasets and empirical settings

Long-Tailed CIFAR. The original version of CIFAR-10 and CIFAR-100 contains 50,000 training images and 10,000 validation images of size 32×32 with 10 and 100 classes, respectively. To create their long-tailed version, we reduce the

---

Algorithm 1 Open-sampling

Require: Training dataset $\mathcal{D}_{\text{train}}$. Open-set auxiliary dataset $\mathcal{D}_{\text{out}}^{(x)}$;
1: **for** each iteration **do**
2:     Sample a mini-batch of original training samples $\{(\boldsymbol{x}_i, y_i)\}_{i=0}^{n}$ from $\mathcal{D}_{\text{train}}$;
3:     Sample a mini-batch of open-set instances $\{\widetilde{\boldsymbol{x}}_i\}_{i=0}^{m}$ from $\mathcal{D}_{\text{out}}^{(x)}$;
4:     Generate random noisy label $\widetilde{y}_i \sim \Gamma$ for each open-set instance $\widetilde{\boldsymbol{x}}_i$;
5:     Perform gradient descent on f with $\mathcal{L}_{\text{total}}$ from Equation (2);
6: **end for**

---

number of training examples per class according to an exponential function $n = n_j \mu^j$, where j is the class index, n_j is the original number of training images, and $\mu \in (0, 1)$. Besides, the validation set and the test set are kept unchanged. The imbalance ratio of a dataset is defined as the number of training samples in the largest class divided by that of the smallest.

CelebA-5. CelebFaces Attributes (CelebA) dataset is a real-world long-tailed dataset. It is originally composed of 202,599 number of RGB face images with 40 binary attributes annotations per image. Note that CelebA is originally a multi-labeled dataset, we port it to a 5-way classification task by filtering only the samples with five non-overlapping labels about hair colors. We also subsampled the full dataset by 1/20 while maintaining the imbalance ratio as 10.7, following (Kim et al., 2020). In particular, We pick out 50 and 100 samples in each class for validation and testing. We denote the resulting dataset by CelebA-5.

Places-LT. Places-LT is a long-tailed version of the large-scale scene classification dataset Places (Zhou et al., 2017). It consists of 184.5K images from 365 categories with class cardinality ranging from 5 to 4,980.

Auxiliary datasets. 300K Random Images (Hendrycks et al., 2019) is a cleaned and debiased dataset with 300K natural images. In this dataset, Images that belong to CIFAR classes from it, images that belong to Places or LSUN classes, and images with divisive metadata are removed so that $\mathcal{D}_{\text{train}}$ and $\mathcal{D}_{\text{out}}$ are disjoint. We use the dataset as the open-set auxiliary dataset for experiments with CIFAR-10/100 and CelebA-5. For Experiments on Places-LT, we use the training set of Places-Extra69 (Zhou et al., 2017) as the open-set auxiliary dataset.

OOD test datasets. Following OE (Hendrycks et al., 2019), we comprehensively evaluate OOD detectors on artificial and real anomalous distributions, including: Gaussian, Rademacher, Blobs, Textures (Cimpoi et al., 2014), SVHN (Netzer et al., 2011), Places365 (Zhou et al., 2017), LSUN-Crop (Yu et al., 2015), LSUN-Resize (Yu et al., 2015), iSUN (Xu et al., 2015). For experiments on CIFAR-10, we also use CIFAR-100 as OOD test dataset. *Gaussian* noises have each dimension i.i.d. sampled from an isotropic Gaussian distribution. *Rademacher* noises are images where each dimension is -1 or 1 with equal probability, so each dimension is sampled from a symmetric Rademacher distribution. *Blobs* noises consist in algorithmically generated amorphous shapes with definite edges. *Textures* (Cimpoi et al., 2014) is a dataset of describable textural images. *SVHN* dataset (Netzer et al., 2011) contains 32×32 color images of house numbers. There are ten classes comprised of the digits 0-9. *Places365* (Zhou et al., 2017) consists in images for scene recognition rather than object recognition. *LSUN* (Yu et al., 2015) is another scene understanding dataset with fewer classes than Places365. Here we use *LSUN-Crop* and *LSUN-Resize* to denote the cropped and resized version of the LSUN dataset respectively. *iSUN* (Xu et al., 2015) is a large-scale eye tracking dataset, selected from natural scene images of the SUN database (Xiao et al., 2010).

E. Implementation details

For experiments on Long-Tailed CIFAR-10/100 (Krizhevsky et al., 2009) and CelebA-5 (Liu et al., 2015a), we perform training with ResNet-32 (He et al., 2016) for 200 epochs, using SGD with a momentum of 0.9, and a weight decay of 0.0002. We set the initial learning rate as 0.1, then decay by 0.01 at the 160th epoch and again at the 180th epoch. For fair comparison, We also use linear warm-up learning rate schedule (Goyal et al., 2017) for the first 5 epochs. For data augmentation in training, we use the commonly used version: 4 pixels are padded on each side, and a 32×32 crop is randomly sampled from the padded image or its horizontal flip. For experiments under the setting of Balanced Softmax (Ren et al., 2020), we use Nesterov SGD with momentum 0.9 and weight-decay 0.0005 for training. We use a total mini-batch size of 512 images on a single GPU. The learning rate increased from 0.05 to 0.1 in the first 800 iterations. Cosine scheduler (Loshchilov & Hutter, 2016) is applied afterward, with a minimum learning rate of 0. Our augmentation follows Balanced Softmax and Equalization Loss (Tan et al., 2020). In testing, the image size is 32×32. In end-to-end training, the model is

trained for 13K iterations. In decoupled training experiments, we fix the Softmax model, i.e., the instance-balanced baseline model obtained from the standard end-to-end training, as the feature extractor. And the classifier is trained for 2K iterations.

For Places-LT, we choose ResNet-152 as the backbone network and pretrain it on the full ImageNet-2012 dataset, following Decoupled training (Kang et al., 2020). We use SGD optimizer with momentum 0.9, batch size 512, cosine learning rate schedule (Loshchilov & Hutter, 2016). Similar to Decoupled training, we fine-tune the backbone with Instance-balanced sampling for representation learning and then re-train the classifier with our proposed algorithm with class-balanced sampling. Through the paper, we refer to decoupled training as training the last linear classifier on a fixed feature extractor obtained from instance-balanced training. For experiments with the SSP method, we use rotation prediction (Komodakis & Gidaris, 2018) as self-supervised learning method, where an image is rotated by a random multiple of 90 degrees and a model is trained to predict the rotation degree as a 4-way classification problem.

We conduct all the experiments on NVIDIA GeForce RTX 3090, and implement all methods with default parameters by PyTorch (Paszke et al., 2019). All experiments are repeated five times with different seeds and we report the average test accuracy. We tune the hyperparameter η on the validation set, then train the model on the full training set. The α is fixed as $(\max_j \beta_j + \min_j \beta_j)$ by default and we find this value performs well overall. For the η in the training objective, the best value depends on the dataset, imbalance ratio, network architecture, and the integrated method.

For OOD detection task, we measure the following metrics: (1) the false positive rate (FPR95) of OOD examples when the true positive rate of in-distribution examples is at 95%; (2) the area under the receiver operating characteristic curve (AUROC); and (3) the area under the precision-recall curve (AUPR).

F. More empirical results

F.1. Ablation study

To show the effect of the class-dependent weighting factor w_j in Open-sampling, we conduct ablation study on long-tailed CIFAR-10 with imbalanced rate 100. As shown in Table 6, the class-dependent weighting factor generally improves the performance, when compared with the variant without the weight.

Table 6. Results of ablation study on long-tailed CIFAR-10 for the class-dependent weighting factor w_j.

Imbalance Factor	100	50	10
Standard	71.61	77.30	86.74
Ours w/o w_j	76.57	81.18	88.57
Ours	**77.62**	**81.76**	**89.38**

F.2. Detailed results for OOD detection.

Table 7 presents the detailed results of OOD detection performance on long-tailed CIFAR-10 with imbalanced rate 100 with various OOD test datasets. From the results, we show that our method can outperform OE (Hendrycks et al., 2019) in OOD detection tasks when training dataset is class-imbalanced.

Table 7. OOD detection performance comparison on long-tailed CIFAR-10 with imbalanced rate 100. ↑ indicates larger values are better and ↓ indicates smaller values are better.

OOD test dataset	Method	FPR95 ↓	AUROC ↑	AUPR ↑
Gaussian	MSP	43.02	68.62	21.54
	OE	13.71	87.94	40.61
	Ours	7.52	93.37	53.82
	Ours ($\alpha = 5$)	1.20	99.55	95.72
Rademacher	MSP	38.33	81.44	32.97
	OE	13.88	87.17	39.50
	Ours	7.93	92.91	52.39
	Ours ($\alpha = 5$)	3.25	97.90	78.00
Blob	MSP	49.03	80.88	39.51
	OE	44.29	76.55	27.28
	Ours	9.42	93.33	55.45
	Ours ($\alpha = 5$)	33.65	91.50	66.51
Textures	MSP	66.53	72.19	29.27
	OE	38.00	84.92	38.14
	Ours	23.29	92.53	57.99
	Ours ($\alpha = 5$)	24.64	94.91	78.43
SVHN	MSP	63.74	69.11	24.06
	OE	43.46	84.17	37.06
	Ours	22.09	93.53	63.92
	Ours ($\alpha = 5$)	28.89	92.12	67.60
CIFAR-100	MSP	70.50	70.24	28.27
	OE	61.15	77.88	32.77
	Ours	55.32	84.09	45.81
	Ours ($\alpha = 5$)	54.73	84.75	54.57
LSUN-C	MSP	40.68	85.52	50.07
	OE	22.60	86.80	39.06
	Ours	14.73	96.70	80.81
	Ours ($\alpha = 5$)	13.66	96.25	82.77
LSUN-R	MSP	59.27	76.86	36.88
	OE	17.91	87.36	39.86
	Ours	15.29	93.87	60.49
	Ours ($\alpha = 5$)	10.68	97.10	82.44
iSUN	MSP	61.86	74.66	32.68
	OE	20.02	87.44	40.37
	Ours	15.33	93.66	58.85
	Ours ($\alpha = 5$)	11.47	96.92	82.40
Places365	MSP	67.98	72.47	31.82
	OE	48.77	81.24	33.90
	Ours	35.91	89.98	54.23
	Ours ($\alpha = 5$)	39.16	91.60	70.69