OPEN-SAMPLING: RE-BALANCING LONG-TAILED DATASETS WITH OUT-OF-DISTRIBUTION DATA

Anonymous authors
Paper under double-blind review

ABSTRACT

Deep neural networks usually perform poorly when the training dataset suffers from extreme *class imbalance*. To handle this issue, popular re-sampling methods generally require in-distribution data to balance the class priors. However, obtaining suitable in-distribution data with precise labels for selected classes is challenging. 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 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 original class priors. Furthermore, class-dependent weights are generated to provide stronger regularization on the minority classes than on the majority classes. We empirically show that Open-sampling not only re-balances the class prior, but also encourages the neural network to learn separable representations. Extensive experiments on benchmark datasets demonstrate that our proposed method significantly outperforms existing data re-balancing methods and can be easily incorporated into existing state-of-the-art methods to enhance their performance.

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., 2019; Kang et al., 2020). Thus, designing effective algorithms to handle class imbalance is of great practical importance.

In the literature, a popular direction in learning from long-tailed datasets is to re-balance the data distribution by data re-sampling. The simplest methods are under-sampling that discards data from the majority classes (Buda et al., 2018; He & Garcia, 2009; Japkowicz & Stephen, 2002) and over-sampling that repeats samples from the minority classes (Byrd & Lipton, 2019; Shen et al., 2016). The former is infeasible if the data imbalance is extreme, while the latter usually causes over-fitting to the minority classes (Cui et al., 2019). To alleviate the over-fitting issue, novel samples are introduced to augment the minority classes without repetition (Chawla et al., 2002). For example, SMOTE (Chawla et al., 2002) produces artificial minority samples by interpolating from neighboring samples, and ADASYN (He et al., 2008) uses synthesized data. However, the model is still error-prone due to noise in the novel 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 adding unlabeled data from mismatched classes would hurt 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.

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 since the uniformity of the former 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 three long-tailed image classification benchmark datasets, including long-tailed CIFAR-10/100 dataset and a real-world long-tailed dataset, CelebA-5. Despite the simplicity of our method, it achieves significant improvements over existing re-sampling or re-weighting methods across all the datasets. Furthermore, experimental results also validate that our method can be easily incorporated into existing state-of-the art methods (e.g., LDAM (Cao et al., 2019) and Balanced Softmax (Ren et al., 2020)) to enhance their performance on long-tailed imbalanced classification tasks.

2 Related Literature

In this section, we introduce the related studies of data re-balancing methods (including re-sampling and re-weighting) and the utilization of auxiliary dataset in the deep learning community.

Re-sampling. Re-sampling methods aims to re-balance the class priors of the training dataset. Under-sampling methods achieve the goal by removing examples from the majority classes, which is infeasible under extreme data imbalanced settings (Buda et al., 2018; He & Garcia, 2009; Japkowicz & Stephen, 2002). The vanilla 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). To alleviate the over-fitting issue, some methods introduce novel in-distribution samples, e.g., interpolated from neighboring samples (Chawla et al., 2002) or synthesized samples (He et al., 2008; Kim et al., 2020). However, the novel samples are either challenging to collect or introduce extra noise. In contrast to in-distribution samples used in existing over-sampling methods, our approach exploits OOD instances to re-balance the class priors of the training dataset.

Re-weighting. In re-weighting methods, adaptive weights are assigned for different classes or even different samples. Generally, the vanilla scheme re-weights classes proportionally to the inverse of their frequency in the training dataset (Huang et al., 2016; Wang et al., 2017). Focal loss (Lin et al., 2017) assigns low weights to the well-classified examples. Cui et al. (2019) showed that re-weighting by inverse class frequency yields poor performance on frequent classes, and thus propose re-weighting by the inverse effective number of samples. However, these re-weighting methods tend to make the optimization of deep neural networks difficult under extreme data imbalanced settings and large-scale scenarios (Huang et al., 2016; Wang et al., 2017; Mikolov et al., 2013).

Utilizing auxiliary dataset. To the best of our knowledge, we are the first to explore the benefits of open-set auxiliary dataset in learning with long-tailed imbalanced data. For learning with long-tailed imbalanced data, Yang & Xu (2020) introduced unlabeled-in-distribution (UID) data to compensate for the lack of training samples. In the deep learning community, auxiliary dataset is also utilized in other problem settings. For example, pre-training a network on the large ImageNet dataset (Russakovsky et al., 2015) can produce general representations that are useful in many fine-tuning applications (Zeiler & Fergus, 2014). OE (Hendrycks et al., 2019) uses an auxiliary dataset to teach the network better representations for OOD detection. To improve robustness against adver-

sarial attacks, a popular method is to train on adversarial examples which can be seen as a generated auxiliary dataset (Goodfellow et al., 2014). Unlabelled data is also shown to be beneficial for the adversarial robustness (Carmon et al., 2019; Uesato et al., 2019). OAT (Lee et al., 2021) utilizes OOD instances to improve generalization in adversarial training. Wei et al. (2021) proposed to use open-set auxiliary dataset to prevent the model from over-fitting inherent noisy labels.

3 IMBALANCED LEARNING WITH OOD INSTANCES

3.1 BACKGROUND

In this work, we consider a multi-class classification problem, where the input space $\mathcal{X}$ is $\mathbb{R}^d$ and the label space $\mathcal{Y}$ is $\{1, \dots, K\}$. 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 points 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 distribution $P_{\text{out}}(Y)$ 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.

3.2 THEORETICAL MOTIVATION

From a Bayesian perspective, the final prediction is generally made as follows:

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

where $P(\boldsymbol{x})$ and $P(y)$ represent $P(X = \boldsymbol{x})$ and $P(Y = y)$, respectively. However, the discrepancy between the class priors of training and test distributions makes the prediction in Eq. (1) unreliable, thereby downgrading the generalization performance. Specifically, the class prior of the test distribution is usually class-balanced (i.e., a uniform distribution over labels), while the training dataset exhibits a long-tailed class distribution. Thus, $P_{\text{s}}(Y = i) \neq P_{\text{t}}(Y = i)$ for any class $i \in \mathcal{Y}$, and $P_{\text{s}}(Y = j) \ll P_{\text{t}}(Y = j)$ for a minority class $j \in \mathcal{Y}$.

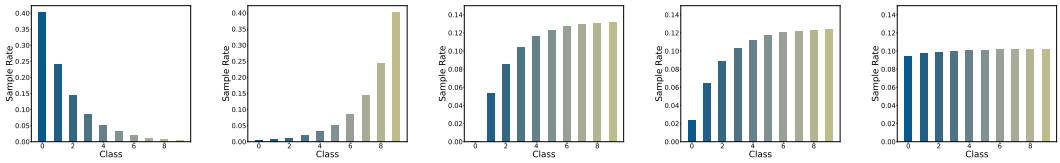
To re-balance the class priors of the training dataset, existing re-sampling methods mostly augment the minority classes with extra *in-distribution* samples with precise labels, e.g., duplicated examples (Buda et al., 2018), synthetic generation (He et al., 2008), and interpolation (Chawla et al., 2002).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

To safely apply OOD instances, we first provide the following theoretical evidence to prove the non-toxicity of open-set instances with uniformly sampled labels.

Theorem 1. *Assume that $P_{\text{out}}(Y)$ is the discrete uniform distribution over the label space $\mathcal{Y}$. Let $\mathcal{D}_{\text{mix}} = \mathcal{D}_{\text{train}} \cup \mathcal{D}_{\text{out}}$, and $P_{\text{mix}}(X, Y)$ be the underlying data distribution of $\mathcal{D}_{\text{mix}}$, then we have*

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

The proof is provided in Appendix A. From Theorem 1, we can see that the prediction of the Bayes classifier is unchanged after augmenting the training dataset with the auxiliary dataset, if the labels of the OOD instances are uniformly sampled from the in-distribution label space. We notice that the non-toxicity of OOD instances with uniformly sampled labels is not new and has been studied before



(a) Original (b) CB (c) MCD (d) CD (α=0.5) (e) CD (α=5)

Figure 1: An illustration of label distributions for long-tailed CIFAR-10 dataset with imbalance ratio 100. (1a) Original label prior, (1b) Class Balanced distribution, (1c) Minimum Complementary Distribution, (1d) and (1e) are Complementary Distributions with different values of α.

in Wei et al. (2021). Nevertheless, our contribution here is to theoretically prove its correctness from a Bayesian perspective. We provide a detailed comparison for these two works in Appendix. C.

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 we can obtain a better classifier by exploring the trade-off between re-balancing the label distribution and keeping non-toxicity.

3.3 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 1 (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, which 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. On the other hand, 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 $K = 2$, 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. As a remedy, we make the following hypothesis.

Hypothesis 1. *There exists a "sweet spot" between the uniform label distribution and the Minimum Complementary Distribution, where a proper Complementary Distribution can be found to obtain a better classifier.*

To find the "sweet spot", we propose the following sampling rate, which allows us to achieve a smooth transition from the Minimum Complementary Distribution to the uniform distribution. Let us denote Complementary Distribution as Γ, Minimum Complementary Distribution as Γ^m, and the complementary sampling rate for class j as Γ_j, then we have:

Proposition 1 (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 bwteen the Minimum Complementary Distribution and uniform distribution. As shown in Proposition 1, when $\alpha = \max_j(\beta_j)$, it recovers a Minimum Complementary Distribution. With a larger value for the α, the label distribu-

tion of the auxiliary dataset would be closer to a uniform distribution. The proof of Proposition 1 is provided in Appendix A.

Figure 1 presents an example of label distribution of the original long-tailed training dataset, Class Balanced distribution (CB) (Cui et al., 2019), and Complementary Distribution (CD) with various α. From Figure 1, we can observe that both the CB distribution and the Complementary Distribution 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 Complementary 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.

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 as far as possible.

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:

$$\mathcal{L} = \mathbb{E}_{\mathcal{D}_{\text{mix}}}[\ell(f(\boldsymbol{x};\boldsymbol{\theta}), y)] = \mathbb{E}_{\mathcal{D}_{\text{mix}}}[-\mathbf{e}^y \log f(\boldsymbol{x};\boldsymbol{\theta})], \tag{2}$$

where $\mathcal{D}_{\text{mix}} = \mathcal{D}_{\text{train}} \cup \mathcal{D}_{\text{out}}$, and $\boldsymbol{e}^y = (0, \ldots, 1, \ldots, 0)^\top \in \{0,1\}^K$ is a one-hot vector in which only the 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:

$$\mathcal{L}_{\text{reg}} = \mathbb{E}_{\mathcal{D}_{\text{out}}}\left[\ell\left(f(\widetilde{\boldsymbol{x}};\boldsymbol{\theta}), \widetilde{y}\right)\right] = \mathbb{E}_{\mathcal{D}_{\text{out}}}[-\mathbf{e}^{\widetilde{y}} \log f(\widetilde{\boldsymbol{x}};\boldsymbol{\theta})], \text{where } \widetilde{y} \sim \Gamma. \tag{3}$$

To 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:

$$\mathcal{L}_{\text{reg}} = \mathbb{E}_{\mathcal{D}_{\text{out}}}\left[\omega_{\widetilde{y}} \cdot \ell\left(f(\widetilde{\boldsymbol{x}};\boldsymbol{\theta}), \widetilde{y}\right)\right], \text{where } \widetilde{y} \sim \Gamma,\ \omega_{\widetilde{y}} = \Gamma_{\widetilde{y}} \cdot K. \tag{4}$$

Now, the final training objective function is given as follows:

$$\mathcal{L}_{\text{total}} = \mathbb{E}_{\mathcal{D}_{\text{train}}}\left[\ell\left(f(\boldsymbol{x};\boldsymbol{\theta}), y\right)\right] + \eta \cdot \mathbb{E}_{\mathcal{D}_{\text{out}}}\left[\omega_{\widetilde{y}} \cdot \ell\left(f(\widetilde{\boldsymbol{x}};\boldsymbol{\theta}), \widetilde{y}\right)\right], \text{where } \widetilde{y} \sim \Gamma,\ \omega_{\widetilde{y}} = \Gamma_{\widetilde{y}} \cdot K, \tag{5}$$

where η controls the strength of the regularization term. The details of the proposed algorithm are provided in Appendix B.

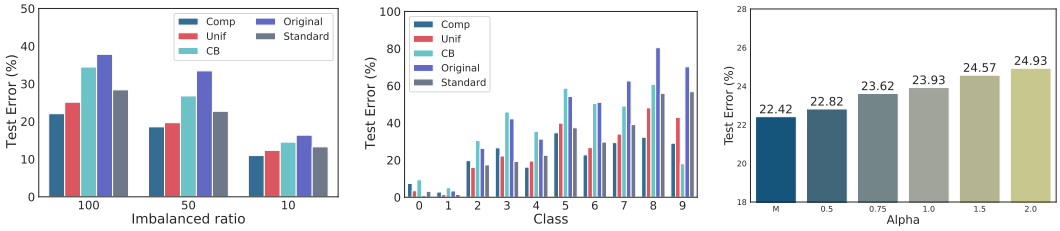
As a data re-balancing technique, Open-sampling is orthogonal to the training objective based methods (e.g., LDAM (Cao et al., 2019), Balanced Softmax (Ren et al., 2020)), and can be easily incorporated into these algorithms to further improve their generalization performance. Given the original learning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

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{imb}} + \eta \cdot \mathcal{L}_{\text{reg}}. \tag{6}$$

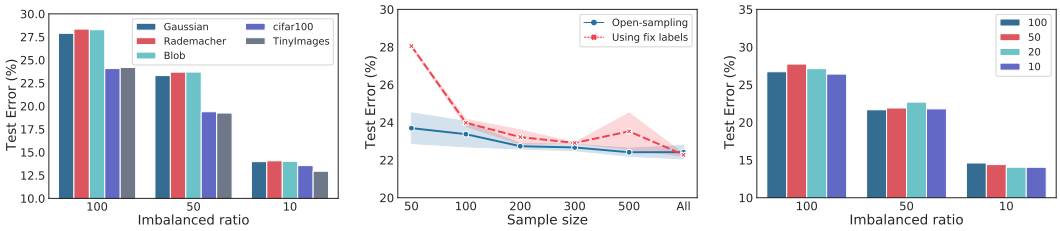
4 A Closer Look at Open-sampling under Class Imbalance

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. Furthermore, 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 auxiliary dataset are sampled from a random label distribution. For the random



(a) Label distribution. (b) Label distribution. (c) Alpha.



(d) Auxiliary dataset. (e) Sample size. (f) Number of Classes.

Figure 2: Analytical experiments of Open-sampling on long-tailed CIFAR-10: (2a)(2b) with various label distributions, (2c) with various values of the α, (2d) with various open-set auxiliary datasets, including simulated noise datasets and real-world datasets. All the datasets contain 50,000 instances. (2e) with various sample sizes (K) of the auxiliary dataset, (2f) with various number of classes in the auxiliary datasets that are randomly sampled from CIFAR-100. Experiments in (2b), (2c), and (2e) are conducted under the imbalance ratio 100. The y-axis represents the test error in all the figures.

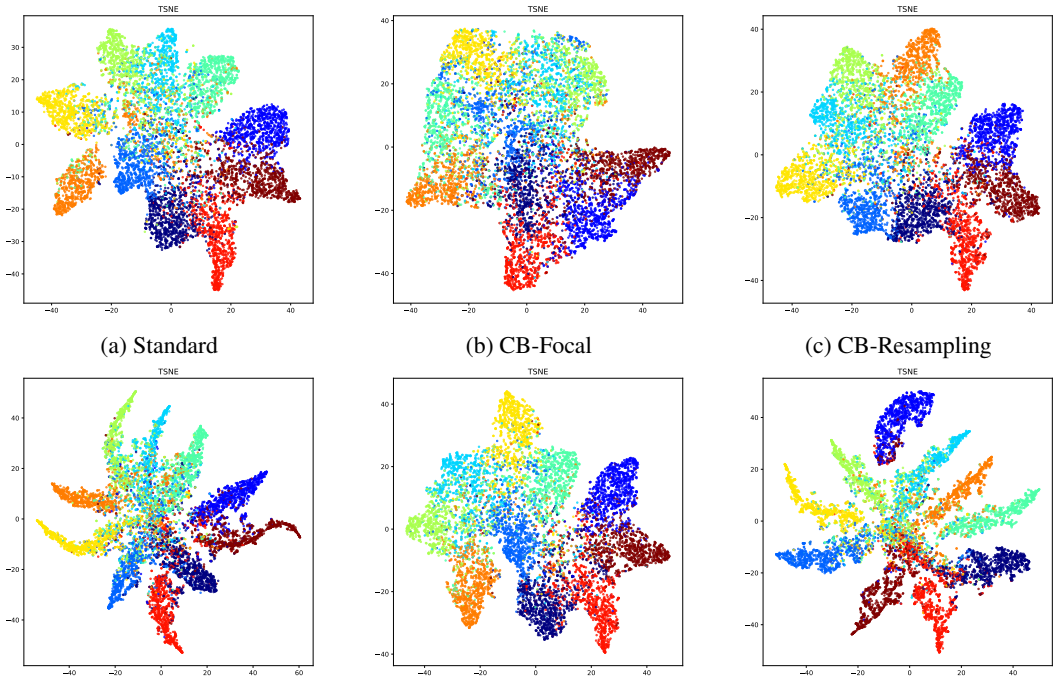
label distribution, we defined a Complementary Distribution in Proposition 1 and also presented several commonly used distributions, including uniform distribution (Unif), class balanced distribution (CB) (Cui et al., 2019), and the original class priors of the training dataset (Original). Here, we conduct experiments to compare the performance of the Open-sampling variants with different label distributions. The results in Figure 2a 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 2b. 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 proposed 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 2c. As analyzed in Section 3, 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 $\alpha = (\max_j \beta_j + \min_j \beta_j)$. From Figure 2c, the test error presented a slightly upward trend with the increasing of the value of α. The results verified that it is necessary to find the "sweet spot" in Hypothesis 1 for a better classifier, 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 2d, 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 noise, e.g., Gaussian noise, Rademacher noise, and Blob noise would almost not change the performance on the test set, while using real-world datasets like CIFAR100 and 80M Tinyimages (Torralba et al., 2008) could achieve impressive improvements.

The sample size is another important factor for the open-set auxiliary dataset. In Fig. 2e, 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



(a) Standard (b) CB-Focal (c) CB-Resampling

(d) LDAM (e) Balanced Softmax (f) Ours

Figure 3: 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.

with noisy labels. Here, we also conduct experiments using a variant 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 2d, where using CIFAR100 as open-set auxiliary dataset can achieve almost the same improvements as 80M Tinyimages.

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 3. 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 5, our method could still boost the performance of the LDAM method, which demonstrates the differences between our works.

5 Experiments

In this section, we evaluate our proposed method on artificially simulated long-tailed CIFAR datasets with controllable degrees of data imbalance and a real-world long-tailed datasets, Celeba-A. Besides, we also analyze the impact of η by sensitivity analysis.

Table 1: Test accuracy (%) of ResNet-32 on long-tailed CIFAR-10 and CIFAR-100 with various imbalance ratios. The † 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-RS	72.31 ± 0.50	76.91 ± 0.83	86.48 ± 0.49	38.41 ± 0.29	42.97 ± 0.57	56.28 ± 0.73
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	**78.11 ± 0.33**	**81.76 ± 0.51**	**89.38 ± 0.46**	**40.26 ± 0.65**	**44.77 ± 0.25**	**58.09 ± 0.29**
LDAM	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	**75.19 ± 0.34**	**79.76 ± 0.44**	**87.28 ± 0.61**	**35.85 ± 0.62**	**42.18 ± 0.82**	**55.48 ± 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	**79.82 ± 0.31**	**82.22 ± 0.45**	**87.83 ± 0.38**	**44.07 ± 0.75**	**47.5 ± 0.24**	**57.43 ± 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	**79.05 ± 0.20**	**82.76 ± 0.52**	**88.89 ± 0.21**	**42.86 ± 0.27**	**47.28 ± 0.58**	**58.80 ± 0.72**

5.1 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 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.

5.2 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 $(1 - \beta^{n_i})/(1 - \beta)$. 4) CB-RS (Cui et al., 2019): balancing the objective from different sampling probability for each sample using class-balanced distribution. 5) CB-Focal (Cui et al., 2019): the CB method is combined with Focal loss. 6) M2m (Kim et al., 2020): an over-sampling method with adversarial examples. 7) LDAM (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. 8) LDAM-DRW (Cao et al., 2019): the network is trained with LDAM loss and deferred re-balancing training. 9) 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. Roughly, these comparison methods can be classified into three categories: (i) re-sampling based methods - (2, 4, 6), (ii) re-weighting based methods - (3, 5), and (iii) different loss functions - (7, 8, 9). 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.

Table 2: Classification accuracy (%) on CelebA-5. The † indicates the reported results 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	73.1	Balanced Softmax	76.4
CB-Focal	74.2	**LDAM + Ours**	**75.8**	**Balanced Softmax + Ours**	**78.6**

5.3 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. The average 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 9.06%. Besides, incorporating our method into existing state-of-the-art methods is shown to consistently improve their performance under different imbalance ratios. In particular, we can observe that our method can boost the performance of existing state-of-the-art methods, such as LDAM-DRW and Balanced Softmax.

To further clarify the influence of η, we present a sensitivity analysis on long-tailed CIFAR-10 dataset with imbalance ratio 100 in Fig. 4. Here, the η is fixed as 1.0. Specifically, 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. It is worth noting that η with too large value would downgrade the performance of the neural network. In other word, the test performance would be a function of η that first increases and then decreases. The phenomenon inspires us to quickly search the best value for η using the validation accuracy throughout training.

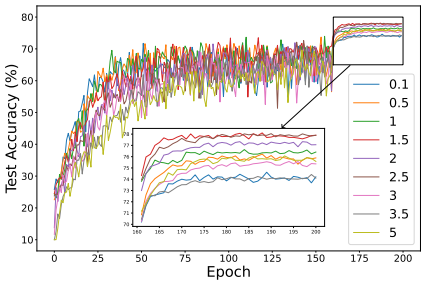


Figure 4: Results of sensitivity analysis on long-tailed CIFAR-10 with various values for η.

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 is designed to have a balanced label distribution. Table 2 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. The results show that our method is applicable for real-world scenarios.

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 the problem of long-tailed imbalanced 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 on benchmark datasets demonstrate that our proposed method significantly outperforms existing data re-balancing methods and can be easily incorporated into existing state-of-the-art methods to enhance their performance.

REFERENCES

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.

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

Kaidi Cao, Colin Wei, Adrien Gaidon, Nikos Arechiga, and Tengyu Ma. Learning imbalanced datasets with label-distribution-aware margin loss. In *Advances in Neural Information Processing Systems*, 2019.

Yair Carmon, Aditi Raghunathan, Ludwig Schmidt, Percy Liang, and John C Duchi. Unlabeled data improves adversarial robustness. *arXiv preprint arXiv:1905.13736*, 2019.

Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. Smote: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, 16:321–357, 2002.

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*, pp. 9268–9277, 2019.

Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014.

Priya Goyal, Piotr Dollár, Ross Girshick, Pieter Noordhuis, Lukasz Wesolowski, Aapo Kyrola, Andrew Tulloch, Yangqing Jia, and Kaiming He. Accurate, large minibatch sgd: Training imagenet in 1 hour. *arXiv preprint arXiv:1706.02677*, 2017.

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

Haibo He, Yang Bai, Edwardo A Garcia, and Shutao Li. Adasyn: Adaptive synthetic sampling approach for imbalanced learning. In *2008 IEEE International Joint Conference on Neural Networks*, pp. 1322–1328, 2008.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pp. 770–778, 2016.

Dan Hendrycks, Mantas Mazeika, and Thomas Dietterich. Deep anomaly detection with outlier exposure. *Proceedings of the International Conference on Learning Representations*, 2019.

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

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

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

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

Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

Saehyung Lee, Changhwa Park, Hyungyu Lee, Jihun Yi, Jonghyun Lee, and Sungroh Yoon. Removing undesirable feature contributions using out-of-distribution data. *arXiv preprint arXiv:2101.06639*, 2021.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European Conference on Computer Vision*, pp. 740–755. Springer, 2014.

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*, pp. 2980–2988, 2017.

Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of the IEEE international conference on computer vision*, pp. 3730–3738, 2015.

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*, pp. 2537–2546, 2019.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems*, pp. 3111–3119, 2013.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in Neural Information Processing Systems*, 32: 8026–8037, 2019.

Jiawei Ren, Cunjun Yu, Shunan Sheng, Xiao Ma, Haiyu Zhao, Shuai Yi, and Hongsheng Li. Balanced meta-softmax for long-tailed visual recognition. In *Proceedings of Neural Information Processing Systems*, Dec 2020.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet large scale visual recognition challenge. *International Journal of Computer Vision*, 115 (3):211–252, 2015.

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

Antonio Torralba, Rob Fergus, and William T Freeman. 80 million tiny images: A large data set for nonparametric object and scene recognition. *IEEE transactions on Pattern Analysis and Machine Intelligence*, 30(11):1958–1970, 2008.

Jonathan Uesato, Jean-Baptiste Alayrac, Po-Sen Huang, Robert Stanforth, Alhussein Fawzi, and Pushmeet Kohli. Are labels required for improving adversarial robustness? *arXiv preprint arXiv:1905.13725*, 2019.

Grant Van Horn, Oisin Mac Aodha, Yang Song, Yin Cui, Chen Sun, Alex Shepard, Hartwig Adam, Pietro Perona, and Serge Belongie. The inaturalist species classification and detection dataset. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pp. 8769–8778, 2018.

Yu-Xiong Wang, Deva Ramanan, and Martial Hebert. Learning to model the tail. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pp. 7032–7042, 2017.

Hongxin Wei, Lue Tao, Renchunzi Xie, and Bo An. Open-set label noise can improve robustness against inherent label noise. *arXiv preprint arXiv:2106.10891*, 2021.

Yuzhe Yang and Zhi Xu. Rethinking the value of labels for improving class-imbalanced learning. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2020.

Matthew D Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In *European Conference on Computer Vision*, pp. 818–833. Springer, 2014.

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

Boyan Zhou, Quan Cui, Xiu-Shen Wei, and Zhao-Min Chen. Bbn: Bilateral-branch network with cumulative learning for long-tailed visual recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9719–9728, 2020.

A PROOFS

Theorem 1 (restated). *Assume that $P_{\text{out}}(Y)$ is the discrete uniform distribution over the label space $\mathcal{Y}$. Let $\mathcal{D}_{\text{mix}} = \mathcal{D}_{\text{train}} \cup \mathcal{D}_{\text{out}}$, and $P_{\text{mix}}(X, Y)$ be the underlying data distribution of $\mathcal{D}_{\text{mix}}$, then we have*

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

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

$$P_{\text{mix}}(\boldsymbol{x}, y) = \frac{N}{M+N} P_{\text{s}}(\boldsymbol{x}, y) + \frac{M}{M+N} P_{\text{out}}(\boldsymbol{x}, y), \tag{7}$$

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

By the virtue of Bayes' theorem, we have

$$\begin{aligned} P_{\text{mix}}(\boldsymbol{x}|y)P_{\text{mix}}(y) &= \frac{N}{M+N} P_{\text{s}}(\boldsymbol{x}|y)P_{\text{s}}(y) + \frac{M}{M+N} P_{\text{out}}(\boldsymbol{x}|y)P_{\text{out}}(y) \\ &= \frac{N}{M+N} P_{\text{s}}(\boldsymbol{x}|y)P_{\text{s}}(y) + \frac{M}{M+N} P_{\text{out}}(\boldsymbol{x})P_{\text{out}}(y) \\ &= \frac{N}{M+N} P_{\text{s}}(\boldsymbol{x}|y)P_{\text{s}}(\boldsymbol{x}, y) + \frac{1}{K} \cdot \frac{M}{M+N} P_{\text{out}}(\boldsymbol{x}), \end{aligned} \tag{8}$$

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

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

$$\begin{aligned} \arg\max_{y \in \mathcal{Y}} P_{\text{mix}}(\boldsymbol{x}|y)P_{\text{mix}}(y) &= \arg\max_{y \in \mathcal{Y}} \left\{ \frac{N}{M+N} P_{\text{s}}(\boldsymbol{x}|y)P_{\text{s}}(\boldsymbol{x}, y) + \frac{1}{K} \cdot \frac{M}{M+N} P_{\text{out}}(\boldsymbol{x}) \right\} \\ &= \arg\max_{y \in \mathcal{Y}} \frac{N}{M+N} P_{\text{s}}(\boldsymbol{x}|y)P_{\text{s}}(\boldsymbol{x}, y). \\ &= \arg\max_{y \in \mathcal{Y}} P_{\text{s}}(\boldsymbol{x}|y)P_{\text{s}}(\boldsymbol{x}, y). \end{aligned} \tag{9}$$

□

Proposition 1 (restated) (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 $\alpha \to \infty$.*

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

If $\beta = \max_j(\beta_j)$, 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:

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

□

B ALGORITHM

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 (5);
6: **end for**

C DISCUSSIONS

Relation to Wei et al. (2021). 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 Wei et al. (2021) and our work.

1. Problem setting: they 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, Wei et al. (2021) can be seen as a special case of our method with a large value of α. As analyzed in Figures 2a and 2c, 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: Wei et al. (2021) 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.

D IMPLEMENTATION DETAILS

For experiments on Long-Tailed CIFAR10 and CIFAR-100 (Krizhevsky et al., 2009), 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 on CelebA-5 (Liu et al., 2015), we use the same training setting as that of CIFAR datasets. For all the experiments, we use 80 Million Tiny Images (Torralba et al., 2008) as the open-set auxiliary dataset. 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 example, for training ResNet-32 network on the long-tailed CIFAR-10 dataset using vanilla Open-sampling, we set $\eta = 1.0$ for experiments with imbalance ratio 100, $\eta = 2.0$ for experiments with imbalance ratio 50 and $\eta = 1.0$ for experiments with imbalance ratio 10.