# ON THE EFFECTIVENESS OF OUT-OF-DISTRIBUTION DATA IN SELF-SUPERVISED LONG-TAIL LEARNING

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

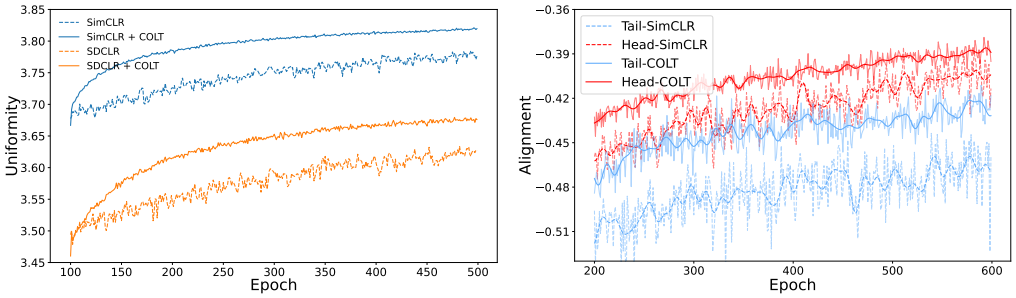
Though Self-supervised learning (SSL) has been widely studied as a promising technique for representation learning, it doesn't generalize well on long-tailed datasets due to the majority classes dominating the feature space. Recent work shows that the long-tailed learning performance could be boosted by sampling extra in-domain (ID) data for self-supervised training, however, large-scale ID data which can rebalance the minority classes are expensive to collect. In this paper, we propose an alternative but easy-to-use and effective solution, **C**ontrastive with **O**ut-of-distribution (OOD) data for **L**ong-**T**ail learning (COLT), which can effectively exploit OOD data to dynamically re-balance the feature space. We empirically identify the counter-intuitive usefulness of OOD samples in SSL long-tailed learning and principally design a novel SSL method. Concretely, we first localize the '*head*' and '*tail*' samples by assigning a tailness score to each OOD sample based on its neighborhoods in the feature space. Then, we propose an online OOD sampling strategy to dynamically re-balance the feature space. Finally, we enforce the model to be capable of distinguishing ID and OOD samples by a distribution-level supervised contrastive loss. Extensive experiments are conducted on various datasets and several state-of-the-art SSL frameworks to verify the effectiveness of the proposed method. The results show that our method significantly improves the performance of SSL on long-tailed datasets by a large margin, and even outperforms previous work which uses external ID data.

## 1 INTRODUCTION

Self-supervised learning (SSL) methods (Chen et al., 2020; He et al., 2020; Grill et al., 2020) provide distinctive and transferable representations in an unsupervised manner and achieve significant progress in various downstream tasks (He et al., 2022; Radford et al., 2021; Devlin et al., 2018). However, most SSL methods are performed on well-curated and balanced datasets (e.g., ImageNet), while many real-world datasets in practical applications, such as medical imaging and self-driving cars, usually follow a long-tailed distribution (Spain & Perona, 2007). Recent research (Liu et al., 2021; Jiang et al., 2021b) indicates that existing SSL methods exhibit severe performance degradation when exposed to imbalanced datasets.

To enhance the robustness of SSL methods under long-tailed data, several pioneering methods (Jiang et al., 2021b; Zhou et al., 2022) are proposed for a feasible migration of cost-sensitive learning, which is widely studied in supervised long-tail learning (Elkan, 2001; Sun et al., 2007; Cui et al., 2019b). The high-level intuition of these methods is to re-balance classes by adjusting loss values for different classes, i.e., forcing the model to pay more attention to tail samples. Another promising line of work explores the probability of improving the SSL methods with external data. (Jiang et al., 2021a) suggests re-balancing the class distributions by sampling external in-distribution (ID) tail instances in the wild. Nevertheless, they still require available ID samples in the sampling pool, which is hard to collect in many real-world scenarios, e.g., medical image diagnosis (Ju et al., 2021) or species classification (Miao et al., 2021).

The aforementioned findings and challenges motivate us to investigate another more practical and challenging setting: *when the ID data is not available, can we leverage the out-of-distribution (OOD) data to improve the performance of SSL in long-tailed learning?* Compare to MAK Jiang et al. (2021a) that assumes external ID samples are available, while we consider a more practical




(a) Uniformity

(b) Alignment

Figure 1: (1a): Feature space uniformity of different SSL frameworks. (1b): Visualization of the alignment property of samples in minority classes and majority classes w/ or w/o COLT. The experiment is conducted with ResNet-18 on CIFAR-100-LT.

scenario where we only have access to OOD data that can be easily collected (e.g., downloaded from the internet). In MAK, it rejects the OOD samples during training and views OOD samples as useless to performance improvement, while we provides counter-intuitive insights to MAK, i.e., OOD samples could be beneficial to self-supervised long-tail learning. A very recent work (Wei et al., 2022) proposes to re-balance the class priors by assigning labels to OOD images following a predefined distribution. However, it is performed in a supervised manner while not directly applicable to the SSL frameworks.

In this paper, we proposed a novel and principal method to exploit the unlabeled OOD data to improve SSL performance on long-tailed learning. As suggested in previous research, the standard contrastive learning would naturally put more weight on the loss of majority classes and less weight on that of minority classes, resulting in imbalanced feature spaces and poor linear separability on tail samples(Kang et al., 2020; Li et al., 2022). However, rebalancing minorities with ID samples, no matter labeled or unlabeled, is quite expensive. To alleviate these issues, we devise a framework, Contrastive Learning with OOD data for Long-Tailed learning (COLT), to dynamically augment the minorities with unlabeled OOD samples which are close to tail classes in the feature space. As illustrated in Fig. 1, our COLT can significantly improve SSL baselines in terms of the Alignment and Uniformity(Wang & Isola, 2020), two widely-used metrics to evaluate the performance of contrastive learning methods, demonstrating the effectiveness of our method.

The pipeline of our method is illustrated in Fig. 2. To augment the long-tail ID dataset, we define a tailness score to localize the head and tail samples in an unsupervised manner. Afterward, we design an online sampling strategy to dynamically re-balance the long-tail distribution by selecting OOD samples close (with a large cosine similarity in the feature space) to the head or tail classes based on a predefined budget allocation function. We follow the intuition to allocate more OOD samples to the tail classes for rebalancing. Those selected OOD samples are augmented with the ID dataset for contrastive training, where an additional distribution-level supervised contrastive loss makes the model aware of the samples from different distributions. Experimental results on four long-tail datasets demonstrate that COLT can greatly improve the performance of various SSL methods and even surpass the state-of-the-art baselines with auxiliary ID data. We also conduct comprehensive ablation studies and analyses to understand and corroborate the effectiveness of COLT. Our contributions can be summarized as:

- We raise the question of whether we can and how to improve SSL on long-tailed datasets effectively with external unlabeled OOD data, which is better aligned with the practical scenarios but counter-intuitive to most existing work and rarely investigated before.
- We design a novel yet easy-to-use SSL method, which is composed of tailness score estimation, dynamic sampling strategies, and additional contrastive losses for long-tail learning with external OOD samples, to alleviate the imbalance issues during contrastive learning.
- We conducted extensive experiments on various datasets and SSL frameworks to verify and understand the effectiveness of the proposed method. Our method consistently outperforms baselines by a large margin with the consistent agreement between the superior performance and various feature quality evaluation metrics of contrastive learning.

## 2 RELATED WORKS

**Supervised learning with imbalanced datasets** Early attempts aim to highlight the minority samples by re-balancing strategy. These methods fall into two categories: re-sampling at the data level Shen et al. (2016); Zou et al. (2018); Geifman & El-Yaniv (2017), or re-weighting at the loss (gradient) level Cui et al. (2019b); Khan et al. (2017); Cao et al. (2019); Jamal et al. (2020). Due to the usage of label-related information, the above methods can not be generalized to the unsupervised field. Kang et al. (2019) suggests that the scheme of decoupling learning representations and classifiers benefits long-tail learning. The feasibility of two-stage training promotes the exploration in unsupervised scenarios.

**Self-supervised long tail learning** (Yang & Xu, 2020) is, to our best, the first to analyze the performance of self-supervised learning (SSL) methods in long-tail learning and verify the effectiveness of self-supervised pre-training theoretically and experimentally. Empirical results show that pre-training models with Rotation (Gidaris et al., 2018) or MoCo (He et al., 2020) consistently outperform the baselines. However, (Liu et al., 2021) shows that SSL methods – although more robust than the supervised methods – are not immune to the imbalanced datasets. Follow-up studies improve the ability of SSL methods on long-tailed datasets. Motivated by the observation that deep neural networks would easily forget hard samples after pruning, (Hooker et al., 2019), (Jiang et al., 2021b) proposed a self-competitor to pay more attention to the hard (tail) samples. BCL (Zhou et al., 2022) further involved the memorization effect of deep neural networks (Zhang et al., 2021b; Arpit et al., 2017) into contrastive learning, i.e., they emphasize samples from tail by assigning more powerful augmentation based on the memorization clue. We show that our method is non-conflict with these existing methods and can further improve the balancedness and accuracy (Section 4.2).

**Learning with auxiliary data** Auxiliary data is widely used in the field of deep learning for different purposes, e.g., improving model robustness (Lee et al., 2020), combating label noise (Wei et al., 2021), OOD detection (Liang et al., 2018; Hendrycks et al., 2018a), domain generalization (Li et al., 2021; Liu et al., 2020; Long et al., 2015), neural network compression (Fang et al., 2021), training large models (Alayrac et al., 2022; Brown et al., 2020). In long-tail learning, MAK (Jiang et al., 2021a) suggests tackling the dataset imbalance problem by sampling in-distribution tail classes' data from an open-world sampling pool. On the contrary, we explore the probability of helping long-tail learning with OOD samples, i.e., ***none*** of the ID samples are included in the sampling pool. Open-Sampling (Wei et al., 2022) utilizes the OOD samples by assigning a label to each sample following a pre-defined label distribution. Their work is performed under supervised scenarios, and the OOD data is not filtered, which results in a massive computation overhead.

## 3 METHOD

### 3.1 PRELIMINARIES

Unsupervised visual representation learning methods aim to find an optimal embedding function $f$, which projects input image $X \in \mathbb{R}^{CHW}$ to the feature space $Z \in \mathbb{R}^d$ with $\boldsymbol{z} = f(\boldsymbol{x})$, such that $\boldsymbol{z}$ retains the discriminative semantic information of the input image. SimCLR (Chen et al., 2020) is one of the state-of-the-art unsupervised learning frameworks, and its training objective is defined as:

$$\mathcal{L}_{CL} = \frac{1}{N} \sum_{i=1}^{N} -\log \frac{exp(\boldsymbol{z}_i \cdot \boldsymbol{z}_i^+ / \tau)}{exp(\boldsymbol{z}_i \cdot \boldsymbol{z}_i^+ / \tau) + \sum_{\boldsymbol{z}_i^- \in Z^-} exp(\boldsymbol{z}_i \cdot \boldsymbol{z}_i^- / \tau)}, \tag{1}$$

where $(\boldsymbol{z}_i, \boldsymbol{z}_i^+)$ is the positive pair of instance $i$, $\boldsymbol{z}_i^-$ indicates the negative samples from the negative set $Z^-$, and $\tau$ is the temperature hyper-parameter. In practice, a batch of images is augmented twice in different augmentations, the positive pair is formulated as the two views of the same image, and the negative samples are the views of other images.

### 3.2 LOCALIZE TAIL SAMPLES IN SELF-SUPERVISED TRAINING

Due to the label-agnostic assumption in the pre-training state, the first step of the proposed method is to localize tail samples. As mentioned earlier, the majority classes dominate the feature space, and tail instances turn out to be outliers. Moreover, the minority classes have lower intra-class

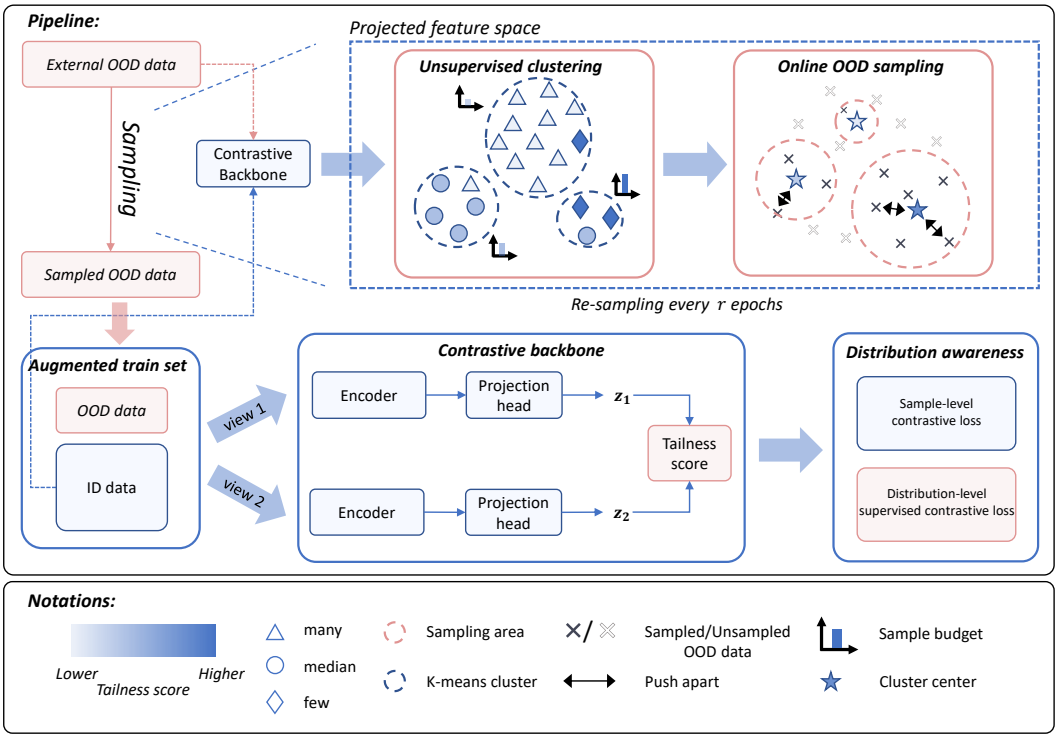



Figure 2: Overview of Contrastive with Out-of-distribution data for Long-Tail learning (COLT). COLT can be easily plugged into most SSL frameworks. Proposed components are denoted as red.

consistency (Li et al., 2022). Hence, a sparse neighborhood could be a reliable proxy to identify the tail samples (More analysis can be found in Section 4.4). Specifically, we use top-$k\%$ ($k = 2$ in practice) largest negative logits of each sample to depict the feature space neighborhood during training. Given a training sample $\boldsymbol{x}_i$, its negative logits $p_i^-$ is the following:

$$p_i^- = \frac{exp(\boldsymbol{z}_i \cdot \boldsymbol{z}_i^- / \tau)}{exp(\boldsymbol{z}_i \cdot \boldsymbol{z}_i^+ / \tau) + \sum_{\boldsymbol{z}_i^- \in Z^-} exp(\boldsymbol{z}_i \cdot \boldsymbol{z}_i^- / \tau)}. \tag{2}$$

Considering implementing SimCLR (Chen et al., 2020) with batch size $B$, each image has $2(B-1)$ negative samples. Then, we define $s_t^i = -\sum_{\text{top}-k\%} p_i^-$ as the tailness score for each ID instance $\boldsymbol{x}_i$. During training, we perform a momentum update to the tailness score, i.e., $s_t^{i,0} = s_t^i, \quad s_t^{i,n} = m s_t^{i,n-1} + (1-m) s_t^{i,n}$ where $m \in [0, 1)$ is the momentum coefficient. The momentum update makes the tailness score more robust and discriminative to the tail samples. A higher value of $s_t^i$ indicates sample $\boldsymbol{x}_i$ has a more sparse neighborhood in the feature space and implies that it belongs to the tail classes with a larger probability. Experiments in Fig 3e empirically demonstrated that tail samples could be effectively discovered by our proposed tailness score.

### 3.3 Dynamically Re-balance the Feature Space with Online Sampling

The core of our approach is to sample OOD images from the sampling pool $S_{ood}$ and further re-balance the original long-tail ID dataset and the feature space. First, we obtain $C$ feature prototypes $\boldsymbol{z}_{c_i}$ from ID training set $S_{id}$ via K-means clustering. Note that we use the features at the last projection layer since the contrastive process is performed on this layer. The cluster-wise tailness score $s_t^{c_i}$ is defined as the mean of tailness score in cluster $c_i$, i.e., $s_t^{c_i} = \sum_{\boldsymbol{z}_j \in c_i} s_t^j / |c_i|$, here $|c_i|$ is the number of instances in cluster $c_i$. Then, we obtain each cluster's sampling budget $K'$ as follows:

$$K' = K \cdot softmax(\widetilde{\boldsymbol{s}}_t^c / \tau_c), \quad \widetilde{\boldsymbol{s}}_t^c = \frac{\boldsymbol{s}_t^c - mean(\boldsymbol{s}_t^c)}{std(\boldsymbol{s}_t^c)}, \tag{3}$$

where $K$ refers to the total sampling budget, $K' \in \mathbb{R}^C$ is the sampling budget assigned to each cluster, $\widetilde{\boldsymbol{s}}_t^c$ is the normalized cluster tailness score. Empirically, we assign more sampling budget to the tailness clusters to be consistent with the idea of re-balancing the feature space. We sample OOD images whose feature is close to (higher cosine similarity) the ID prototypes $\boldsymbol{z}_{c_i}$.

To fully exploit the OOD data, we re-sample from the $S_{ood}$ every $T$ epoch. The motivation behind this is: i) the sampled OOD data can be well-separated from $S_{id}$ after a few epochs, therefore becoming less effective to re-balance the feature space; ii) over-fitting to the OOD data can be toxic to the ID performance (Wei et al., 2022). From another perspective, this online sampling strategy lets the ID training set (especially the tail samples) continuously be exposed to the more newly sampled effective negative samples, forcing the model gives more distinctive embeddings and better fitting the ID distribution. The online sampling process is summarized in Algorithm 2.

**Algorithm 1** The overall pipeline of COLT.

**Input**: ID train set $S_{id}$, OOD dataset $S_{ood}$, sample budget $K$, train epoch $T$, momentum coefficient $m$, warm-up epochs $w$, sample interval $r$, cluster number $C$, hyper-parameter $k, \tau_c$.
**Output**: pre-trained model parameter $\theta_T$.
**Initialize**: model parameter $\theta_0$, the original train set $S_{train} = S_{id}$.

```
if epoch = 0 then
    Train model θ_0 with Eq. 1 and compute s_t^0;
end if
for epoch = 1, ··· , T − 1 do
    if epoch ≥ w then
        if (epoch − w) %r = 0 then
            Sample by performing Algorithm 2.
        end if
        Calculate the supervised contrastive loss
        with Eq. 5
        Use S_train to train θ_epoch with Eq. 6;
    else
        Use S_train to train θ_epoch with Eq. 1;
    end if
    Compute s_t^i, and update s_t^{i,epoch}.
end for
```

**Algorithm 2** our online sampling strategy.

**Input**: ID train set $S_{id}$, OOD dataset $S_{ood}$, model $\theta$, sample budget $K$, cluster number $C$, similarity metric sim$(\cdot)$, hyper-parameter $\tau_c$.
**Output**: new train set $S_{train}$.

```
Calculate both ID features z^{id} and OOD features z^{ood} through model θ;
Obtain C ID prototypes z_{c_i} via K-means clustering in the projected feature space;
Calculate cluster-wise tailness score by s_t^{c_i} = Σ_{z_j ∈ c_i} s_t^j / |c_i|;
Assign each cluster a sample budget K'_{c_i} with Eq. 3;
Initialize the sample set S_sample = ∅;
for i = 0, ··· , C − 1 do
    Initialize subset S^i_sample = ∅;
    while |S_sample| < K'_{c_i} do
        u = arg max_{x_j ∈ S_ood} sim(z_j, z_{c_i});
        S^i_sample = S^i_sample ∪ {u};
    end while
    S_sample = S_sample ∪ S^i_sample;
end for
S_train = S_train ∪ S_sample.
```

### 3.4 Awareness of the Out-Of-Distribution Data

Section 3.2 and Section 3.3 introduce our sampling strategy toward OOD images. To involve the sampled OOD subset $S_{sample}$ in training, a feasible way is directly using the augmented training set (containing both ID and OOD samples) to train the model with Eq. 1. However, we would argue that giving equal treatment to all samples may not be the optimal choice (details in Section 4). One natural idea is to let the model be aware of that there are two kinds of samples from different domains. Hence, we define an indicator $\phi$ to provide weakly supervised (distribution only) information:

$$\phi(x_i) = \begin{cases} +1, & \boldsymbol{x}_i \in S_{id}; \\ -1, & \boldsymbol{x}_i \in S_{ood}. \end{cases} \tag{4}$$

Afterward, we add a supervised contrastive loss (Khosla et al., 2020) to both ID and OOD samples:

$$\mathcal{L}_{SCL} = \frac{1}{N} \sum_{i=1}^{N} \frac{1}{|P(i)|} \sum_{p \in P(i)} -\log \frac{exp(\boldsymbol{z}_i \cdot \boldsymbol{z}_p / \tau)}{exp(\boldsymbol{z}_i \cdot \boldsymbol{z}_p / \tau) + \sum_{n \in N(i)} exp(\boldsymbol{z}_i \cdot \boldsymbol{z}_n / \tau)}, \tag{5}$$

where $P(i) \equiv \{p : \phi(\boldsymbol{x}_p) = \phi(\boldsymbol{x}_i)\}$ is the set of indices of the same domain within the mini-batch, $|P(i)|$ is its cardinality and the negative index set $N(i) \equiv \{n : \phi(\boldsymbol{x}_n) \neq \phi(\boldsymbol{x}_i)\}$ contains index from

Table 1: Test accuracy (%) and balancedness (Std↓) on CIFAR-10-LT and CIFAR-100-LT.

| Method | CIFAR-10-LT | | | | | CIFAR-100-LT | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Metric | Many ↑ | Median ↑ | Few ↑ | Std ↓ | All ↑ | Many ↑ | Median ↑ | Few ↑ | Std ↓ | All ↑ |
| SimCLR | 82.40 | 73.91 | 70.19 | 5.11 | 75.34 | 51.50 | 45.58 | 45.96 | 2.71 | 47.65 |
| +COLT | **87.50** | **81.65** | **80.80** | **2.98** | **83.15** | **57.94** | **56.74** | **57.72** | **0.52** | **57.46** |
| SDCLR | 86.69 | 82.15 | 76.23 | 4.28 | 81.74 | 58.54 | 55.70 | 52.10 | 2.64 | 55.48 |
| +COLT | **90.87** | **84.28** | **81.45** | **3.95** | **85.41** | **63.28** | **60.85** | **59.42** | **1.59** | **61.18** |
| BCL-I | 86.97 | 82.40 | 76.45 | 4.31 | 81.99 | 58.92 | 54.63 | 53.58 | 2.31 | 55.70 |
| +COLT | **89.03** | **85.10** | **80.36** | **3.55** | **84.86** | **61.12** | **57.03** | **55.82** | **2.27** | **57.98** |

different distribution. The proposed loss explicitly separates the ID samples and OOD samples, which can be roughly interpreted as a classifier to separate head and tail samples, since most ID instances belong to majority classes and the sampled OOD images' feature is similar to the minority classes. Fig 3c illustrates that the proposed distribution-awareness loss improves not only the overall performance but also facilitates a more balanced feature space. It's worth noting that the proposed loss only utilizes the distribution information as the supervised term, while the labels for both ID and OOD samples are unavailable during the self-supervised training stage. Finally, we scale the supervised loss with $\alpha$ and add it to the contrastive loss in Eq 1:

$$\mathcal{L}_{COLT} = \mathcal{L}_{CL} + \alpha \mathcal{L}_{SCL}. \tag{6}$$

# 4 Experiments

In this section, we first introduce the datasets and experimental settings (Section 4.1) and evaluate the proposed COLT in three aspects: accuracy and balancedness(Section 4.2), versatility and complexity (Section 4.3). Then, we verify whether our method can **1),** localize tail samples, **2),** re-balance the feature space. Finally, we conduct experiments on the hyper-parameters in COLT (Section 4.4).

## 4.1 Datasets and Settings

We conduct experiments on four popular datasets. **CIFAR-10-LT/CIFAR-100-LT** are long-tail subsets sampled from the original CIFAR10/CIFAR100 (Cui et al., 2019a). We set the imbalance ratio to 100 in default. Following (Wei et al., 2022), we use 300K Random Images (Hendrycks et al., 2018b) as the OOD dataset. **ImageNet-100-LT** is proposed by (Jiang et al., 2021b) with 12K images sampled from ImageNet-100 (Tian et al., 2020) with Pareto distribution. We use ImageNet-R (Hendrycks et al., 2021) as the OOD dataset. **Places-LT** (Liu et al., 2019) contains about 62.5K images sampled from the large-scale scene-centric Places dataset (Zhou et al., 2017) with Pareto distribution. Places-Extra69 (Zhou et al., 2017) is utilized as the OOD dataset.

**Evaluation protocols** To verify the balancedness and separability of the feature space, we report performance under two widely-used evaluation protocols in SSL: ***linear-probing*** and ***few-shot***. For both protocols, we first perform self-supervised training on the encoder model to get the optimized visual representation. Then, we fine-tune a linear classifier on top of the fixed encoder. The only difference between linear-probing and few-shot learning is that we use the full dataset for linear probing and 1% samples of the full dataset for few-shot learning during fine-tuning.

**Measurement metrics** As a common practice in long tail learning, we divide each dataset into three disjoint groups in terms of the instance number of each class: {*Many, Median, Few*}. By calculating the standard deviation of the accuracy of the three groups, we can quantitatively analyze the balancedness of a feature space (Jiang et al., 2021b). The linear separability of the feature space is evaluated by the overall accuracy.

**Training settings** We evaluate our method with SimCLR (Chen et al., 2020) framework in default. We also conduct experiments on several state-of-the-art methods in self-supervised long tail learning (Jiang et al., 2021b; Zhou et al., 2022). We adopt Resnet-18 (He et al., 2016) for small datasets (CIFAR-10-LT/CIFAR-100-LT), and Resnet-50 for large datasets (ImageNet-100-LT/Places-LT), respectively. More details can be found in Appendix.

Table 2: Test accuracy (%) and balancedness (Std↓) on ImageNet-100-LT and Places-LT.

| Method | ImageNet-100-LT | | | | | Places-LT | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Metric | Many ↑ | Median ↑ | Few ↑ | Std ↓ | All ↑ | Many ↑ | Median ↑ | Few ↑ | Std ↓ | All ↑ |
| SimCLR | 69.53 | 63.74 | 59.88 | 3.97 | 65.50 | 40.02 | 46.61 | 49.38 | 3.93 | 44.78 |
| +COLT | **76.49** | **70.27** | **68.04** | **3.58** | **72.41** | **41.61** | **48.08** | **50.89** | **3.88** | **46.30** |
| SDCLR | 70.10 | 65.04 | 60.92 | 3.75 | 66.48 | 40.13 | 46.61 | 48.90 | **3.71** | 44.73 |
| +COLT | **75.13** | **70.25** | **67.69** | **3.08** | **71.82** | **41.72** | **48.42** | **50.78** | 3.84 | **46.47** |

Table 3: Compare the proposed COLT with random sample and MAK under the same sampling pool and sampling budget. The best performance under each setting is marked as **bold**.

| ID dataset | Sampling pool | Budget | Method | Protocol | Many ↑ | Median ↑ | Few ↑ | Std ↓ | All ↑ |
|---|---|---|---|---|---|---|---|---|---|
| Image Net-100 | Image Net-R | 5K | random | few-shot | 51.76 | 41.45 | 38.12 | 5.81 | 45.04 |
| | | | | linear-probing | 72.22 | 66.49 | 63.45 | 3.64 | 68.33 |
| | | | MAK | few-shot | 52.49 | 43.54 | 39.00 | 5.65 | 46.48 |
| | | | | linear-probing | 73.24 | 67.62 | 64.14 | 3.75 | 69.36 |
| | | | COLT | few-shot | **54.10** | **46.01** | **42.32** | **4.92** | **48.69** |
| | | | | linear-probing | **74.52** | **69.67** | **67.53** | **2.92** | **71.28** |
| | | 10K | random | few-shot | 52.82 | 42.88 | 40.27 | 5.41 | 46.42 |
| | | | | linear-probing | 74.33 | 68.52 | 62.65 | 4.77 | 70.02 |
| | | | MAK | few-shot | 54.33 | 45.01 | 40.12 | 5.90 | 48.01 |
| | | | | linear-probing | 75.57 | 68.20 | 66.29 | 4.00 | 70.83 |
| | | | COLT | few-shot | **55.04** | **45.67** | **44.97** | **4.59** | **49.23** |
| | | | | linear-probing | **75.34** | **70.83** | **69.45** | **2.52** | **72.41** |
| Places 365 | Places 69 | 10K | random | few-shot | 30.61 | 33.94 | 37.55 | **2.83** | 33.45 |
| | | | | linear-probing | 40.21 | 47.59 | 50.51 | 4.33 | 45.51 |
| | | | MAK | few-shot | 30.71 | 34.75 | 38.04 | 3.00 | 33.94 |
| | | | | linear-probing | 40.83 | 47.78 | 50.72 | 4.15 | 45.86 |
| | | | COLT | few-shot | **31.44** | **35.74** | **40.13** | 3.55 | **35.05** |
| | | | | linear-probing | **41.72** | **48.42** | **50.78** | **3.84** | **46.47** |

### 4.2 COLT'S ACCURACY, BALANCEDNESS AND VERSATILITY

The main results of the proposed approach in various datasets and settings are presented in Table 1 and Table 2. We sample $K = 10,000$ OOD images on every $r = 25$ epoch for CIFAR-10-LT/CIFAR-100-LT and $r = 50$ for ImageNet-100-LT. We only sample $K = 10,000$ OOD images once for Places-LT for better computational efficiency. COLT significantly outperforms the baseline (vanilla SimCLR) by a large margin (about 10% for long-tail CIFAR, 7% for ImageNet-100-LT, 1.6% for Places-LT). Besides, the performance gain of the minority classes (*Median* & *Few*) is more notable (e.g., about 12% for long-tailed CIFAR-100). Meanwhile, COLT yields a balanced feature space. Following previous works (Jiang et al., 2021b) (Zhou et al., 2022), we measure the balancedness of a feature space through the accuracy's standard deviation from *Many, Median* and *Few*. COLT significantly narrows the performance gap between the three groups (much lower Std), which indicates we learn a more balanced feature space.

To evaluate the versatility of COLT, we carry out experiments on top of several improved SSL frameworks for long-tail learning, i.e., SDCLR (Jiang et al., 2021b) and BCL (Zhou et al., 2022). Table 1 and Table 2 also summarized COLT performance on these two methods. We can observe that incorporating our method into existing state-of-the-art methods can consistently improve their performance, which indicates that our method is robust to the underlying SSL frameworks.

Table 4: Compare the test accuracy (%) on ImageNet-100-LT of the proposed COLT with MAK which use ID data. The best performance is marked as **bold**.

| Method | Extra type | Sample set | Many ↑ | Median ↑ | Few ↑ | Std ↓ | All ↑ |
|---|---|---|---|---|---|---|---|
| MAK | ID | IN-900 | **75.7±0.5** | 70.4±0.6 | 66.9±0.6 | 3.0±0.4 | 72.0±0.5 |
| | ID & OOD | IPM | 74.7±0.2 | 69.2±0.7 | 66.6±0.7 | 3.3±0.3 | 71.1±0.5 |
| | OOD | ImageNet-R | 75.6±0.4 | 68.2±0.8 | 66.3±0.8 | 4.1±0.6 | 70.8±0.5 |
| COLT | OOD | ImageNet-R | 75.3±0.3 | **70.9±0.8** | **69.5±0.3** | **2.4±0.7** | **72.4±0.3** |

### 4.3 COLT vs Baselines with Auxiliary Data

We also compare COLT with methods that make use of external data. MAK (Jiang et al., 2021a) is the state-of-the-art method that proposes a sampling strategy to re-balance the training set by sample in-distribution tail class instances from an open-world sampling pool. We compare the proposed COLT with MAK in Table 3, noting that "random" refers to random sampling from the external dataset according to the budget. We observe both higher accuracy and balancedness under different sample budgets on ImageNet-100 and Places. It indicates COLT leverages OOD data in a more efficient way. Furthermore, we ask the question that *whether OOD samples can replace ID samples to help long-tail learning.* We obtain a positive answer from empirical results in Table 4. We compare the result of COLT and MAK on auxiliary data which involve ID samples. COLT achieves better performance on most of the metrics, even compared with sampling in an entirely ID dataset.

On the other, COLT has less computational overhead. MAK applies a three-stage pipeline: pre-train the model with ID samples, sample from the sampling pool, and re-train a model from the beginning. In contrast, COLT samples during the training process, resulted in a single-stage pipeline. The online sampling strategy not only fully utilized the external datasets but also reduced the computation overhead significantly[1]. Open-Sampling (Wei et al., 2022) also uses OOD data to help long-tail learning. Different from ours, they mainly carry out experiments in a supervised manner. Moreover, they use an extremely large data budget (300K for CIFAR), while COLT improves the baselines with a much smaller budget (10K for CIFAR).

### 4.4 Analysis and Ablation Study

**The choice of OOD dataset** We conduct experiments on CIFAR-100-LT and replace the OOD dataset while maintaining other settings unchanged. As shown in Fig 3a, our method improves the ID accuracy when using 300K Random Images (Hendrycks et al., 2018b), STL (Coates et al., 2011), ImageNet-100, Places with 9.81%, 5.19%, 9.43%, 5.74% respectively. Besides, sampling on Gaussian Noise provides limited help (less than 1%) or degradation to ID accuracy.

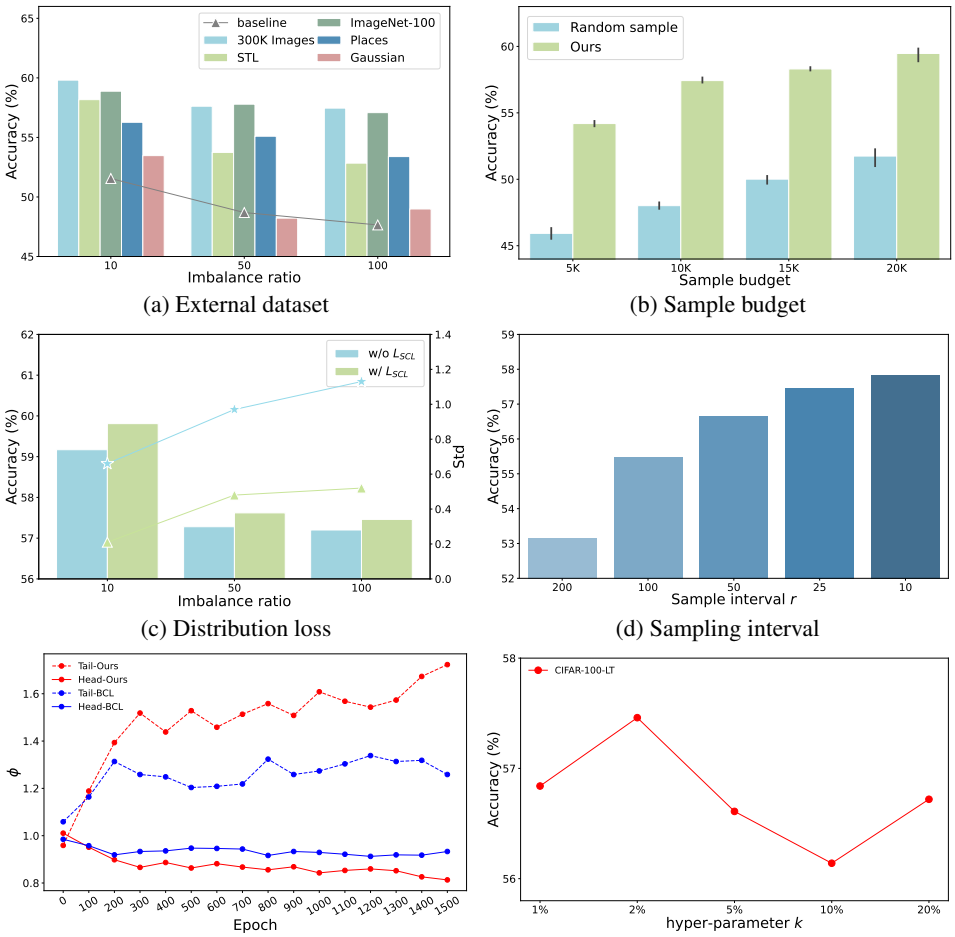
**The effectiveness of distribution-awareness loss** COLT introduces a supervised contrastive loss to explicitly separate samples from different distributions. We conduct an ablation study on the proposed loss in Fig 3c, and the results show that the proposed loss not only improves the overall accuracy but also significantly alleviates the imbalance (i.e., much lower std).

**The effect of sampling budget** In Fig 3b, we compare the performance gains of COLT under different budgets. COLT consistently outperforms the random sampling strategy, i.e., leveraging OOD samples more effectively. Moreover, though a larger budget will give better performance, the performance gain almost plateaus with a budget of 10-15k in COLT, indicating better data efficiency.

**Changing of hyper-parameters** We also show the effect of hyper-parameters involved in COLT. Fig 3d shows the empirical results of changing resample interval $r$. We can observe that reasonably small intervals lead to higher accuracy. Fig 3f shows the classification accuracy when changing $k$. The limited fluctuations in performance prove that COLT is robust to hyper-parameters.

**The ability of tail sample mining** Recall in Section 3.2, we localize tail samples by assigning a predefined "tailness score" to each sample. In order to verify the effectiveness of *tailness score*, we select the top 10% samples with the highest tailness score as a subset and calculate the ratio of the percentage of {*Major* / *Minor*} samples in this subset to the percentage of whole dataset:$\phi = \frac{\mathcal{T} \cap S_{id}^{sub}}{\mathcal{T} \cap S_{id}}$

[1]Although we sample for multiple times (every $r$ epochs) in COLT, the time of performing sampling once is less than training for one epoch; therefore can be ignored compared to 1.7x training epochs brought by MAK.




(a) External dataset (b) Sample budget

(c) Distribution loss (d) Sampling interval

(e) Tail discovery. (f) Ablation study on hyper-parameter $k$.

Figure 3: Analytical experiments of COLT on CIFAR-100-LT. (3a): accuracy when changing the external OOD dataset. (3b): accuracy when sampling different numbers of OOD samples on 300K Random Images. (3c): Top-1 accuracy and standard derivation (Std) of COLT with or without the proposed distribution loss. (3d): accuracy with various sampling intervals $r$. (3e): A higher $\phi_{tail}$ and a lower $\phi_{head}$ implies mining tail samples more precisely. (3f): accuracy with various $k$.

where $\mathcal{T}$ denotes the target group, $S_{id}$ is the whole in-distribution dataset, $S_{id}^{sub}$ is the subset of samples which have top $\gamma\%$ highest tailness score, $\gamma$ is set to 0.1. $\phi$ reflects the ability to identify tail samples: when the target group is *Minor*/*Major*, higher/lower $\phi$ indicates a method localizes tail samples well. As illustrated in Fig 3e, COLT discovers more samples from the tail than BCL, which utilizes the momentum loss to mine tail samples.

## 5 Conclusion and Limitations

In this paper, we propose a novel SSL pipeline COLT, which is, to our best, the first attempt to extend additional training samples from OOD datasets for improved SSL long-tailed learning. COLT includes three steps, unsupervised localizing head/tail samples, re-balancing the feature space by online sampling, and SSL with additional distribution-level supervised contrastive loss. Extensive experiments show that our method significantly and consistently improves the performance of SSL on various long-tailed datasets. There are nevertheless some limitations. First, more theoretical analyses are needed to better understand the effectiveness of OOD samples. Besides, for a given long-tail ID dataset, how to specify the best OOD dataset that gives the largest improvements is also worth exploring. We hope our work can promote the exploration of OOD data in long-tail scenarios.

## REFERENCES


Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katie Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *arXiv preprint arXiv:2204.14198*, 2022.

Devansh Arpit, Stanisław Jastrzębski, Nicolas Ballas, David Krueger, Emmanuel Bengio, Maxinder S Kanwal, Tegan Maharaj, Asja Fischer, Aaron Courville, Yoshua Bengio, et al. A closer look at memorization in deep networks. In *International conference on machine learning*, pp. 233–242. PMLR, 2017.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

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

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pp. 1597–1607. PMLR, 2020.

Adam Coates, Andrew Ng, and Honglak Lee. An analysis of single-layer networks in unsupervised feature learning. In *Proceedings of the fourteenth international conference on artificial intelligence and statistics*, pp. 215–223. JMLR Workshop and Conference Proceedings, 2011.

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, 2019a.

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, 2019b.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

Charles Elkan. The foundations of cost-sensitive learning. In *International joint conference on artificial intelligence*, volume 17, pp. 973–978. Lawrence Erlbaum Associates Ltd, 2001.

Gongfan Fang, Yifan Bao, Jie Song, Xinchao Wang, Donglin Xie, Chengchao Shen, and Mingli Song. Mosaicking to distill: Knowledge distillation from out-of-domain data. *Advances in Neural Information Processing Systems*, 34:11920–11932, 2021.

Yonatan Geifman and Ran El-Yaniv. Deep active learning over the long tail. *arXiv preprint arXiv:1711.00941*, 2017.

Spyros Gidaris, Praveer Singh, and Nikos Komodakis. Unsupervised representation learning by predicting image rotations. In *International Conference on Learning Representations*, 2018.

Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent-a new approach to self-supervised learning. *Advances in neural information processing systems*, 33:21271–21284, 2020.

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.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9729–9738, 2020.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16000–16009, 2022.

Dan Hendrycks, Mantas Mazeika, and Thomas Dietterich. Deep anomaly detection with outlier exposure. In *International Conference on Learning Representations*, 2018a.

Dan Hendrycks, Mantas Mazeika, and Thomas Dietterich. Deep anomaly detection with outlier exposure. In *International Conference on Learning Representations*, 2018b.

Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, et al. The many faces of robustness: A critical analysis of out-of-distribution generalization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 8340–8349, 2021.

Sara Hooker, Aaron Courville, Gregory Clark, Yann Dauphin, and Andrea Frome. What do compressed deep neural networks forget? *arXiv preprint arXiv:1911.05248*, 2019.

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

Ziyu Jiang, Tianlong Chen, Ting Chen, and Zhangyang Wang. Improving contrastive learning on imbalanced data via open-world sampling. *Advances in Neural Information Processing Systems*, 34:5997–6009, 2021a.

Ziyu Jiang, Tianlong Chen, Bobak J Mortazavi, and Zhangyang Wang. Self-damaging contrastive learning. In *International Conference on Machine Learning*, pp. 4927–4939. PMLR, 2021b.

Lie Ju, Xin Wang, Lin Wang, Tongliang Liu, Xin Zhao, Tom Drummond, Dwarikanath Mahapatra, and Zongyuan Ge. Relational subsets knowledge distillation for long-tailed retinal diseases recognition. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 3–12. Springer, 2021.

Bingyi Kang, Saining Xie, Marcus Rohrbach, Zhicheng Yan, Albert Gordo, Jiashi Feng, and Yannis Kalantidis. Decoupling representation and classifier for long-tailed recognition. In *International Conference on Learning Representations*, 2019.

Bingyi Kang, Yu Li, Sa Xie, Zehuan Yuan, and Jiashi Feng. Exploring balanced feature spaces for representation learning. In *International Conference on Learning Representations*, 2020.

Salman H Khan, Munawar Hayat, Mohammed Bennamoun, Ferdous A Sohel, and Roberto Togneri. Cost-sensitive learning of deep feature representations from imbalanced data. *IEEE transactions on neural networks and learning systems*, 29(8):3573–3587, 2017.

Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. *Advances in Neural Information Processing Systems*, 33:18661–18673, 2020.

Ivan Krasin, Tom Duerig, Neil Alldrin, Vittorio Ferrari, Sami Abu-El-Haija, Alina Kuznetsova, Hassan Rom, Jasper Uijlings, Stefan Popov, Andreas Veit, Serge Belongie, Victor Gomes, Abhinav Gupta, Chen Sun, Gal Chechik, David Cai, Zheyun Feng, Dhyanesh Narayanan, and Kevin Murphy. Openimages: A public dataset for large-scale multi-label and multi-class image classification. *Dataset available from https://github.com/openimages*, 2017.

Ananya Kumar, Aditi Raghunathan, Robbie Matthew Jones, Tengyu Ma, and Percy Liang. Fine-tuning can distort pretrained features and underperform out-of-distribution. In *International Conference on Learning Representations*, 2021.

Saehyung Lee, Changhwa Park, Hyungyu Lee, Jihun Yi, Jonghyun Lee, and Sungroh Yoon. Removing undesirable feature contributions using out-of-distribution data. In *International Conference on Learning Representations*, 2020.

Bo Li, Yezhen Wang, Shanghang Zhang, Dongsheng Li, Kurt Keutzer, Trevor Darrell, and Han Zhao. Learning invariant representations and risks for semi-supervised domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1104–1113, 2021.

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

Shiyu Liang, Yixuan Li, and R Srikant. Enhancing the reliability of out-of-distribution image detection in neural networks. In *International Conference on Learning Representations*, 2018.

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

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.

Ziwei Liu, Zhongqi Miao, Xingang Pan, Xiaohang Zhan, Dahua Lin, Stella X Yu, and Boqing Gong. Open compound domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12406–12415, 2020.

Mingsheng Long, Yue Cao, Jianmin Wang, and Michael Jordan. Learning transferable features with deep adaptation networks. In *International conference on machine learning*, pp. 97–105. PMLR, 2015.

Zhongqi Miao, Ziwei Liu, Kaitlyn M Gaynor, Meredith S Palmer, Stella X Yu, and Wayne M Getz. Iterative human and automated identification of wildlife images. *Nature Machine Intelligence*, 3 (10):885–895, 2021.

Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. 2017.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pp. 8748–8763. PMLR, 2021.

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.

Kihyuk Sohn, David Berthelot, Nicholas Carlini, Zizhao Zhang, Han Zhang, Colin A Raffel, Ekin Dogus Cubuk, Alexey Kurakin, and Chun-Liang Li. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. *Advances in neural information processing systems*, 33:596–608, 2020.

Merrielle Spain and Pietro Perona. Measuring and predicting importance of objects in our visual world. 2007.

Yanmin Sun, Mohamed S Kamel, Andrew KC Wong, and Yang Wang. Cost-sensitive boosting for classification of imbalanced data. *Pattern recognition*, 40(12):3358–3378, 2007.

Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive multiview coding. In *European conference on computer vision*, pp. 776–794. Springer, 2020.

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.

Tongzhou Wang and Phillip Isola. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In *International Conference on Machine Learning*, pp. 9929–9939. PMLR, 2020.

Yifei Wang, Qi Zhang, Yisen Wang, Jiansheng Yang, and Zhouchen Lin. Chaos is a ladder: A new theoretical understanding of contrastive learning via augmentation overlap. In *International Conference on Learning Representations*, 2021.

Hongxin Wei, Lue Tao, Renchunzi Xie, and Bo An. Open-set label noise can improve robustness against inherent label noise. *Advances in Neural Information Processing Systems*, 34:7978–7992, 2021.

Hongxin Wei, Lue Tao, Renchunzi Xie, Lei Feng, and Bo An. Open-sampling: Exploring out-of-distribution data for re-balancing long-tailed datasets. In *International Conference on Machine Learning*, pp. 23615–23630. PMLR, 2022.

Yuzhe Yang and Zhi Xu. Rethinking the value of labels for improving class-imbalanced learning. *Advances in neural information processing systems*, 33:19290–19301, 2020.

Bowen Zhang, Yidong Wang, Wenxin Hou, Hao Wu, Jindong Wang, Manabu Okumura, and Takahiro Shinozaki. Flexmatch: Boosting semi-supervised learning with curriculum pseudo labeling. *Advances in Neural Information Processing Systems*, 34:18408–18419, 2021a.

Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning (still) requires rethinking generalization. *Communications of the ACM*, 64(3):107–115, 2021b.

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.

Zhihan Zhou, Jiangchao Yao, Yan-Feng Wang, Bo Han, and Ya Zhang. Contrastive learning with boosted memorization. In *International Conference on Machine Learning*, pp. 27367–27377. PMLR, 2022.

Yang Zou, Zhiding Yu, BVK Kumar, and Jinsong Wang. Unsupervised domain adaptation for semantic segmentation via class-balanced self-training. In *Proceedings of the European conference on computer vision (ECCV)*, pp. 289–305, 2018.

## REPRODUCIBILITY

To ensure the reproducibility of the proposed method, we uploaded the code anonymously, which can be found at: `https://anonymous.4open.science/r/COLT-05D4/`

## APPENDIX A DATASETS AND TRAINING DETAILS

**CIFAR-10-LT/CIFAR-100-LT** are first introduced by (Cui et al., 2019a), which are long-tail subsets sampled from the original CIFAR10/CIFAR100. The imbalance ratio is defined as the instance number of the largest class divided by the smallest class. To better reflect the performance difference, we set the imbalance ratio to 100 in default. Following (Wei et al., 2022), we use 300K Random Images (Hendrycks et al., 2018b) [2] as the OOD dataset. In addition, we also conduct experiments with OOD datasets as STL-10 (Coates et al., 2011), which contains 5,000 labeled images and 100,000 unlabeled images in 10 classes with a resolution of 96x96. We use all the unlabeled images as external OOD data.

**ImageNet-100-LT** is proposed by (Jiang et al., 2021b). it contains about 12K images sampled from ImageNet-100 (Tian et al., 2020) with Pareto distribution. The instance number of each class ranges from 1,280 to 5. We use ImageNet-R (Hendrycks et al., 2021) as the OOD dataset. The dataset contains 30K images with several renditions (e.g., art, cartoons, deviantart) of ImageNet classes.

**Places-LT** The original Places (Zhou et al., 2017) is a large-scale scene-centric dataset. Places-LT (Liu et al., 2019) contains about 62.5K images sampled from Places with Pareto distribution. The instance number of each class ranges from 4,980 to 5. Places-Extra69 (Zhou et al., 2017) is utilized as the OOD dataset. It includes 98,721 images for 69 scene categories besides the 365 scene categories in Places.

**Training details** We implement all our techniques using PyTorch (Paszke et al., 2017) and conduct the experiments using RTX3090 GPUs. We evaluate our method with SimCLR (Chen et al., 2020) framework with batch size 512 for small datasets (CIFAR-10-LT/CIFAR-100-LT) and 256 for large datasets (ImageNet-100-LT/Places-LT) in default. We adopt Resnet-18 (He et al., 2016) for small datasets and Resnet-50 for large datasets, respectively. In our paper, we evaluate COLT's performance under two evaluation protocols in self-supervised learning: ***linear-probing*** and ***few-shot***. For both protocols, we first perform self-supervised training on the encoder model to get the optimized visual representation. Then, we fine-tune a linear classifier on top of the encoder (fixed during training the classifier). The only difference between linear probing and few-shot learning is we use the full dataset for linear probing and 1% samples of the full dataset for few-shot learning during fine-tuning. We keep all settings in the fine-tuning stage (e.g., optimizer, learning rate, batch size) the same as (Jiang et al., 2021b).

Mainly Following (Jiang et al., 2021b; Zhou et al., 2022; Jiang et al., 2021a), we pre-train all the baselines and COLT with 2000 epochs on CIFAR10/100, 1000 epochs on ImageNet-100, 500 epochs on Places. As for the fine-tuning stage, the "linear-probing" and "few-shot" results are produced by fine-tuning the classifier for 30 epochs and 100 epochs, respectively. To make a fair comparison, we implement COLT and all baselines with the same data augmentation strategies. We sample $K = 10,000$ OOD images on every $r = 25$ epoch for CIFAR-10-LT/CIFAR-100-LT and $r = 50$ for ImageNet-100-LT. We only sample $K = 10,000$ OOD images once for Places-LT for better computational efficiency.

## APPENDIX B MORE EMPIRICAL RESULTS

We present the experiment results on ImageNet-100 with ImageNet-R or Places69 as the external OOD dataset in Table 5. We compare COLT with methods that make use of external data. MAK (Jiang et al., 2021a) is the state-of-the-art method that proposes a sampling strategy to re-balance the training set by sample in-distribution tail class instances. Note that "random" refers to randomly sampling from the external dataset according to the sampling budget. We observe both higher accuracy and balancedness under different sample budgets on different OOD datasets. It indicates COLT

[2] A debiased subset of 80 Million Tiny Images (Torralba et al., 2008) with 300K images. 80 Million Tiny Images is constructed by the corresponding images of 53,464 nouns from internet search engines.

Table 5: Accuracy (%) and balancedness (Std) on ImageNet-100 with different OOD datasets.

| Sampling pool | Budget | Method | Protocol | Many ↑ | Median ↑ | Few ↑ | Std ↓ | All ↑ |
|---|---|---|---|---|---|---|---|---|
| None | - | - | few-shot | 48.92 | 39.54 | 34.15 | 6.10 | 42.51 |
| | | | linear-probing | 69.53 | 63.74 | 59.88 | 3.97 | 65.50 |
| Places69 | 10K | random | few-shot | 52.97 | 44.03 | 39.98 | 5.43 | 46.99 |
| | | | linear-probing | 73.49 | 67.57 | 63.05 | 4.28 | 69.29 |
| | | MAK | few-shot | 54.16 | 45.17 | 41.43 | 5.34 | 48.19 |
| | | | linear-probing | 74.69 | 68.60 | 64.63 | 4.14 | 70.46 |
| | | COLT | few-shot | **55.82** | **48.32** | **43.49** | **4.83** | **49.24** |
| | | | linear-probing | **75.23** | **71.42** | **66.46** | **3.59** | **72.26** |
| Image Net-R | 5K | random | few-shot | 51.76 | 41.45 | 38.12 | 5.81 | 45.04 |
| | | | linear-probing | 72.22 | 66.49 | 63.45 | 3.64 | 68.33 |
| | | MAK | few-shot | 52.49 | 43.54 | 39.00 | 5.65 | 46.48 |
| | | | linear-probing | 73.24 | 67.62 | 64.14 | 3.75 | 69.36 |
| | | COLT | few-shot | **54.10** | **46.01** | **42.32** | **4.92** | **48.69** |
| | | | linear-probing | **74.52** | **69.67** | **67.53** | **2.92** | **71.28** |
| | 10K | random | few-shot | 52.82 | 42.88 | 40.27 | 5.41 | 46.42 |
| | | | linear-probing | 74.33 | 68.52 | 62.65 | 4.77 | 70.02 |
| | | MAK | few-shot | 54.33 | 45.01 | 40.12 | 5.90 | 48.01 |
| | | | linear-probing | 75.57 | 68.20 | 66.29 | 4.00 | 70.83 |
| | | COLT | few-shot | **55.04** | **45.67** | **44.97** | **4.59** | **49.23** |
| | | | linear-probing | **75.34** | **70.83** | **69.45** | **2.52** | **72.41** |

leverages OOD data in a more efficient way. Besides, the performance gain of the minority classes (*Median* & *Few*) is more notable. Meanwhile, COLT yields balancedness feature space. Following previous works (Jiang et al., 2021b) (Zhou et al., 2022), we measure the balancedness of a feature space through the accuracy's standard deviation (Std) from *Many, Median* and *Few*. To further demonstrate that our proposed COLT is also effective on the non-curated open-world datasets, we conduct experiments on ImageNet-100-LT with a 50K subset of Open Images (Krasin et al., 2017) (a dataset of about 9 million images belonging to over 6000 categories.) as the OOD dataset. We can also observe a significant improvement in both the accuracy (especially for {Median, Few}) and the balancedness of the feature space.

We also conduct experiments on iNaturalist[3] to evaluate the effectiveness of the proposed COLT, the results are summerized in Tab. 7. We per-train the ResNet-50 for 200 epochs and fine-tuned it for 90 epochs. We select ImageNet as the OOD dataset, and the sampling budget is set to 100K. We can see COLT brings significant performance gains and alleviates the imbalances of feature space.

## Appendix C How do OOD Data Contribute to Long-Tail Learning?

### OOD Data Re-balance the Feature Space

As suggested in previous research, the standard contrastive learning would naturally put more weight on the loss of majority classes and less weight on that of minority classes, resulting in imbalanced feature spaces and poor linear separability on tail samples(Kang et al., 2020; Li et al., 2022), i.e., the majority classes dominate the feature space and tail instances turn out to be outliers.

[3]A large scale long-tailed dataset composed of 8,142 species, with about 437K training images, and 24,426 validation images.

Table 6: Accuracy (%) and balancedness (Std) on ImageNet-100 with external 50K Open-Images (Krasin et al., 2017) as OOD dataset.

| Method | Budget | Protocol | Many ↑ | Median ↑ | Few ↑ | Std ↓ | All ↑ |
|---|---|---|---|---|---|---|---|
| SimCLR | - | few-shot | 49.85 | 41.67 | 37.08 | 5.28 | 44.26 |
| | | linear-probing | 69.53 | 63.74 | 59.88 | 3.97 | 65.50 |
| SimCLR+COLT | 5K | few-shot | 54.46 | 47.12 | 43.23 | 4.66 | 49.48 |
| | | linear-probing | 75.62 | 70.87 | 67.29 | 3.41 | 72.26 |
| | 10K | few-shot | 57.54 | 48.12 | 44.00 | 5.67 | 51.26 |
| | | linear-probing | 76.87 | 71.00 | 69.23 | 3.27 | 73.06 |

Table 7: Accuracy (%) and balancedness (Std) on iNaturalist with ImageNet as OOD dataset.

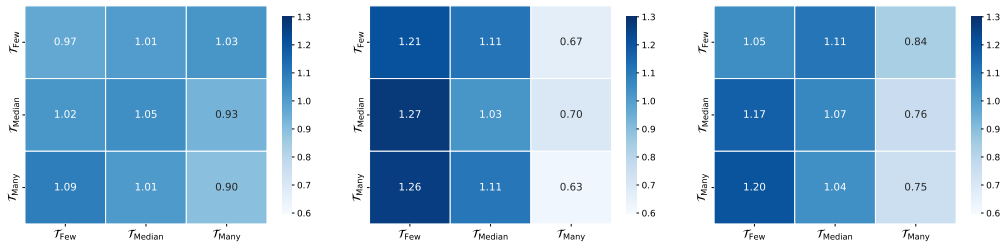
| Method | Many ↑ | Median ↑ | Few ↑ | Std ↓ | All ↑ |
|---|---|---|---|---|---|
| SimCLR | 60.73 | 52.47 | 48.28 | 5.17 | 51.67 |
| SimCLR+COLT | 62.35 | 55.71 | 52.60 | 4.07 | 55.29 |

To quantitatively analyze the imbalance of the feature space, we define a metric called Normalized Misclassification Matrix (NMM):

$$\text{NMM}_{ij} = \frac{m_{ij} / \sum_{k=1}^{n} m_{ik}}{|\mathcal{T}_j| / \sum_{k=1}^{n} |\mathcal{T}_k|}, \tag{7}$$

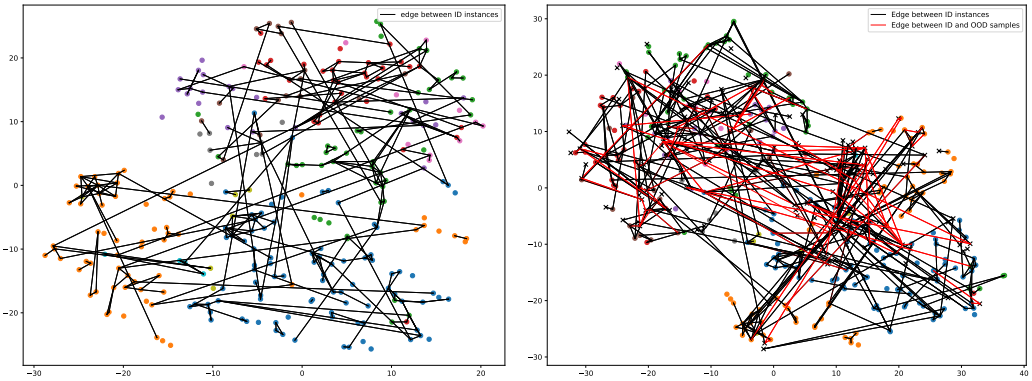
where $\mathcal{T}_k$ is the $k$-th split of the long-tailed train set, satisfying $S_{id} = \cup_{k=1}^{n} \mathcal{T}_k$ and $\mathcal{T}_i \cap \mathcal{T}_j = \emptyset, \forall i \neq j$. In this paper, we follow the common practice in long-tail learning that split the dataset to $S_{id} = \mathcal{T} = \{\mathcal{T}_{\text{Few}}, \mathcal{T}_{\text{Median}}, \mathcal{T}_{\text{Many}}\}$ according to the instance number in each class, and $|\mathcal{T}_j|$ denote the class number in split $\mathcal{T}_j$. $m_{ij}$ represents the number of (misclassified) instances belonging to split $\mathcal{T}_i$ but are classified to split $\mathcal{T}_j$. Note that $m_{ii}$ indicates both the ground truth label and the (wrong) prediction fall into split $\mathcal{T}_i$. Intuitively, if the feature space is perfectly balanced (i.e., equal margin between different classes), each element in NMM is approximately 1.0, i.e., the misclassified samples nearly randomly fall into each split. On the contrary, higher (lower) mean margins between classes in $\mathcal{T}_i$ and $\mathcal{T}_j$ result in lower (higher) $m_{ij}$.

The results are shown in Fig 4. Fig 4a indicates when the train set is balanced, the mean margin of different classes is approximately equal ($\text{NMM}_{ij} \approx 1$), note that the split $\{\mathcal{T}_{\text{Few}}, \mathcal{T}_{\text{Median}}, \mathcal{T}_{\text{Many}}\}$ in Fig 4a is consistent with Fig 4b and Fig 4c. Fig 4b reflects that majority classes have a higher margin to other classes than minority classes. In other words, the model is more likely to confuse




(a) SimCLR(BL) (b) SimCLR(LT) (c) SimCLR+COLT(LT)

Figure 4: Normalized Misclassification Matrix (NMM) on the test set of CIFAR-100 with different frameworks and train sets. (4a): SimCLR trained on balanced CIFAR-100. (4b): SimCLR trained on long-tailed CIFAR-100. (4c): implement COLT on top of SimCLR trained on long-tailed CIFAR-100.




(a) SimCLR (majority classes' mean connectivity: 0.73 minority classes' mean connectivity: 0.46)

(b) SimCLR+COLT (majority classes' mean connectivity: 0.76 minority classes' mean connectivity: 0.67)

Figure 5: The augmentation graph of CIFAR-10. Similar to Wang et al. (2021), We choose a random subset of test images and randomly augment them 20 times. Then, we calculate the instance distance in the representation space and draw edges for image pairs whose smallest view distance is below a small threshold. We visualize the samples with t-SNE and denote edges between ID instances in black and edges between ID and OOD samples, forming new connections in red.

samples from a minority class with other minority classes, implying an imbalanced feature space. Fig 4c exhibits the result of COLT based on SimCLR with 10K OOD data. It's observed that COLT alleviates the imbalance issue since COLT augments minority classes with more instances which could be interpreted as an implicit loss re-weighting strategy.

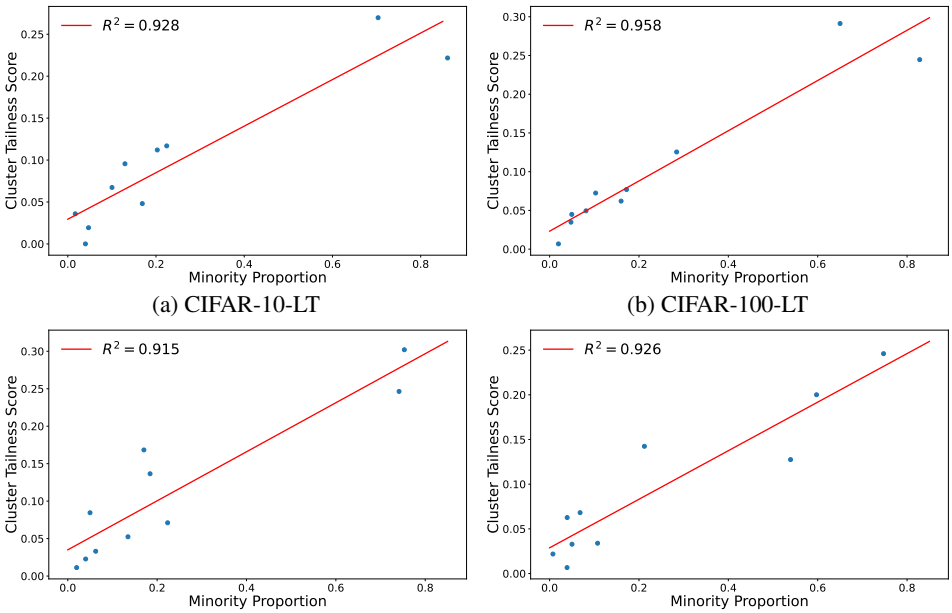
### OOD Data Bridge Instances from Minority Classes

In this section, we demonstrate the effect of OOD data on the perspective of contrastive learning. A recent work (Wang et al., 2021) gives a theoretical understanding of contrastive learning based on *augmentation overlap*. Concretely, they suggest that the samples of the same class could be very alike after aggressive data augmentations. Thus, the pretext task of aligning positive samples can facilitate the model to learn class-separable representations. They define the augmentation graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ as: $N$ samples are the vertices of the graph, and there exists an edge $e_{ij}$ when sample $i$ and sample $j$ has overlapped views. According to their theory, intra-class augmentation overlap is a sufficient condition for gathering features from the same class. In this case, we compute the ratio of connected nodes (degree is not zero) to measure the extent of intra-class augmentation overlap in class $k$, denoted as score $s_c^k$. Ideally, each class should have $s_c^k = 1$, which indicates all samples from the same class will be clustered together during the contrastive training process. A smaller $s_c^k$ indicates lower intra-class consistency and vice versa.

We visualize the augmentation graph of CIFAR-10 in Fig 5 and compute the connectivity score for each class. We observe a smaller $s_c$ for minority classes than for majority classes, which is also consistent with the theoretical analysis in (Wang et al., 2021). Nevertheless, since we dynamically add OOD samples around ID (especially minority classes) samples, more instances from the ID minority class can be reachable on the augmentation graph via OOD samples (majority/minority connectivity improvement: 0.03/0.31). The red edges in Fig 5b illustrate that OOD samples help long-tail learning by bridging samples from minority classes, further improving intra-class consistency.

## Appendix D More Analysis About COLT

**Different tail estimation strategies** In our paper, we defined a tailness score based on top-$k\%$ ($k = 2$ in practice) largest negative logits to localize the head and tail samples in an unsupervised manner. We also make a comparison to other alternative strategies. In Tab. 8, we provide results on locating tail samples by a radius-based definition (with radius as the sum of negative logits and select different radius thresholds) of tailness score or simply using the method in BCL Zhou et al. (2022). The sampling budget is set to 5K, and the ID and OOD dataset is CIFAR-100-LT and 300K




(a) CIFAR-10-LT (b) CIFAR-100-LT (c) ImageNet-100-LT (d) Places-365-LT

Figure 6: Linear regression results between the minority proportion in a cluster and the cluster's tailness score on long-tailed CIFAR, ImageNet-100, and Places. We set cluster number $C = 10$.

Random Images respectively. We can notice that there is no significant difference between COLT with Top-k% and radius-based methods, while both of them surpass BCL.

**Is the unsupervised clustering reliable?** In the paper, we propose to perform a K-means clustering to the samples from $S_{id}$, then calculate the cluster-wise sampling budget via the tailness score. Since our ultimate goal is to sample more (less) OOD data similar to the minority (majority) samples, it's natural to ask *how well does this clustering work*, *whether we assigned more budget to the tail samples*. To this end, we propose to measure the clustering and sampling quality by the ratio of minority samples in a cluster and its corresponding cluster-wise tailness score. Results are visualized in Fig 6. It's observed that the minority proportion of some clusters is close to 1, while others are composed of samples from majority classes. Besides, the cluster-wise tailness score shows a linear correlation of the minority proportion, implying we do allocate more sampling budget to tail classes according to Eq. 3.

**Impact of the cluster number** Recall in Sec 3.3 we first perform a K-means clustering to the ID samples, then select OOD samples close (with a large cosine similarity in the feature space) to the head or tail classes based on the budget allocation function. To verify 1), whether the cluster number will dramatically affect the performance 2), how the performance is when clustering based on the ground-truth label information rather than a self-supervised manner. We conduct experiments with various cluster number $C$ in Tab. 9, the OOD dataset is 300K Random Images Hendrycks et al. (2018b), and the sampling budget is set to 5K. Note that the supervised clustering method is referred to as Oracle. The results show the performance of COLT will not be significantly affected by the number of clusters. Furthermore, cluster according to the label information (Oracle) achieves a similar performance compared to the unsupervised clustering, implying the K-means clustering can be used as a feasible alternative to roughly separates the minority from the majority classes in self-supervised scenarios.

**The scale / balancedness of the OOD dataset** In our work, we evaluate the proposed COLT on both balanced datasets, which are used as OOD datasets in previous works Kumar et al. (2021); Wei et al. (2022) and non-curated open-world datasets such as 300K Random Images Hendrycks et al. (2018b). Although COLT performs well on the aforementioned datasets, we intend to further ask *will COLT perform well when the OOD dataset is also long-tailed?* Tab. 10 shows the performance of COLT on CIFAR-100 when the OOD dataset (ImageNet-100) has various imbalance ratios. The

Table 8: Comparison on different tail estimation strategies.

| Method | COLT | Radius-based | | | | | BCL |
|---|---|---|---|---|---|---|---|
| Threshold | - | 0.08 | 0.09 | 0.10 | 0.11 | 0.12 | - |
| Accuracy | $54.20 \pm 0.35$ | $53.19 \pm 0.47$ | $53.33 \pm 0.28$ | $54.00 \pm 0.21$ | $53.87 \pm 0.22$ | $53.0 \pm 0.34$ | $52.97 \pm 0.30$ |

Table 9: Ablation study on the cluster number $C$ on CIFAR-100-LT.

| Cluster Number C | 10 | 20 | 50 | 100 | 100 (Oracle) |
|---|---|---|---|---|---|
| Accuracy | $54.20 \pm 0.35$ | $53.61 \pm 0.18$ | $54.06 \pm 0.22$ | $53.81 \pm 0.29$ | $54.16 \pm 0.10$ |

similar accuracy suggests COLT is robust to the imbalancedness of the OOD dataset to some extent. This could be attributed to the dynamic sampling procedure filtering out most of the unhelpful OOD samples so that the performance will not be dramatically affected by changes in the external OOD dataset.

Another observation can be found from Tab. 11, where we measured how the OOD dataset's scale influences COLT. We form the external OOD dataset of different scales by gradually increasing the number of samples in Random 300K Images Hendrycks et al. (2018b), and implementing COLT on those subsets of Random 300K Images. Conform to intuition, the scale of the OOD dataset is positively correlated with the performance of COLT since we may select more desired samples in a larger candidate set. However, the problem can be tricky when the dataset is also changed. For example, we observe a similar performance gain on ImageNet-100-LT when utilizing Places-69 (about 98K images) and ImageNet-R (30K images) as the auxiliary OOD dataset, which implies the scale of the OOD dataset is not the only factor affecting the performance.

## APPENDIX E COMPARISON TO SEMI-SUPERVISED METHODS

In our work, we propose a SSL framework COLT to incorporate OOD data into self-supervised long-tail learning to help the minority classes. An alternative and promising choice of utilizing external unlabeled data is to perform semi-supervised learning. We implement several popular semi-supervised methods, i.e., FixMatch Sohn et al. (2020) and FlexMatch Zhang et al. (2021a) on long-tailed CIFAR-100, and compare the results in such semi-supervised learning scenarios (labeled: CIFAR-100-LT, unlabeled: 300K Random Images Hendrycks et al. (2018b)) to supervised, self-supervised and our COLT methods. Note that we only replace the WRN-28-8 with ResNet-18 (also used in COLT) and keep other settings unchanged for a fair comparison. It can be observed that 1), external unlabeled OOD data can also be helpful when performing semi-supervised learning 2), the performance gains of COLT (about 10% on SimCLR) are more significant than incorporating OOD data via semi-supervised training (3.25% improvement on FixMatch and 6.27% on FlexMatch).

Table 10: COLT's accuracy when OOD dataset is also imbalanced.

| Imbalance ratio | 1 | 10 | 50 | 100 |
|---|---|---|---|---|
| Accuracy | $52.98 \pm 0.33$ | $52.63 \pm 0.19$ | $52.74 \pm 0.25$ | $52.66 \pm 0.41$ |

Table 11: COLT's accuracy with different scale OOD dataset.

| OOD Images Number (K) | 100 | 200 | 300 |
|---|---|---|---|
| Accuracy | $53.43 \pm 0.26$ | $53.99 \pm 0.18$ | $54.20 \pm 0.35$ |

Table 12: Comparison of semi-supervised and self-supervised methods when leveraging OOD data.

| Method | ID-Supervised | FixMatch | FlexMatch | SimCLR | SimCLR+COLT(10K) |
|---|---|---|---|---|---|
| Accuracy | 44.13 | 47.38 | 50.40 | 47.65 | 57.46 |