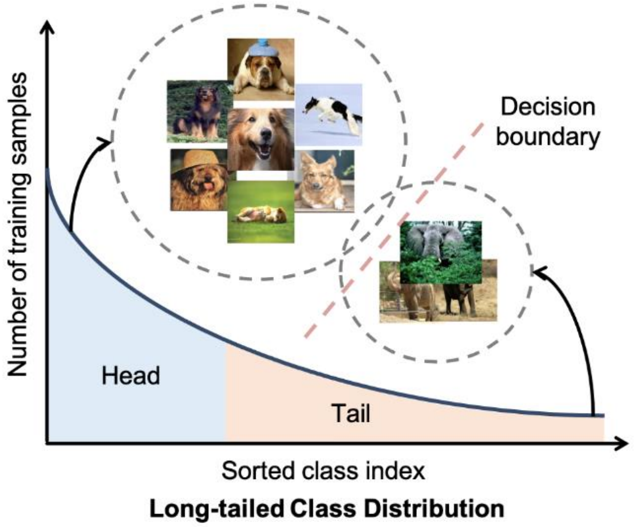
Zhejiang University
College of Information Science and Electronic Engineering

On the Effectiveness of Out-of-Distribution Data in Self-Supervised Long-Tail Learning

Jianhong Bai[1*], Zuozhu Liu[1*], Hualiang Wang[2], Jin Hao[3], Yang Feng[4] , Huanpeng Chu[1], Haoji Hu[1†]

[1] Zhejiang University.
[2] The Hong Kong University of Science and Technology.
[3] Harvard University.
[4] Angelalign Technology.
[*] Equal contribution.
[†] Corresponding Author.

Background.



Long-tailed training set.

- Long-tailed distribution is a common characteristic of real-world data, where a few head classes contribute the majority of data, while most tail classes comprise relatively few instances. Self-supervised learning (SSL) methods suffer from severe performance degradation when the training set is imbalanced.

- Recent work shows that the self-supervised long-tailed learning performance could be boosted by sampling extra in-distribution (ID) data for self-supervised training. However, large-scale ID data which can rebalance the minority classes are expensive to collect.

Main contributions.

1. 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.

2. 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.

3. We conducted extensive experiments on various datasets and SSL frameworks to verify and understand the effectiveness of the proposed method. Our method consistently out-performs baselines by a large margin with the consistent agreement between the superior performance and various feature quality evaluation metrics of contrastive learning.

Overview of the proposed method.

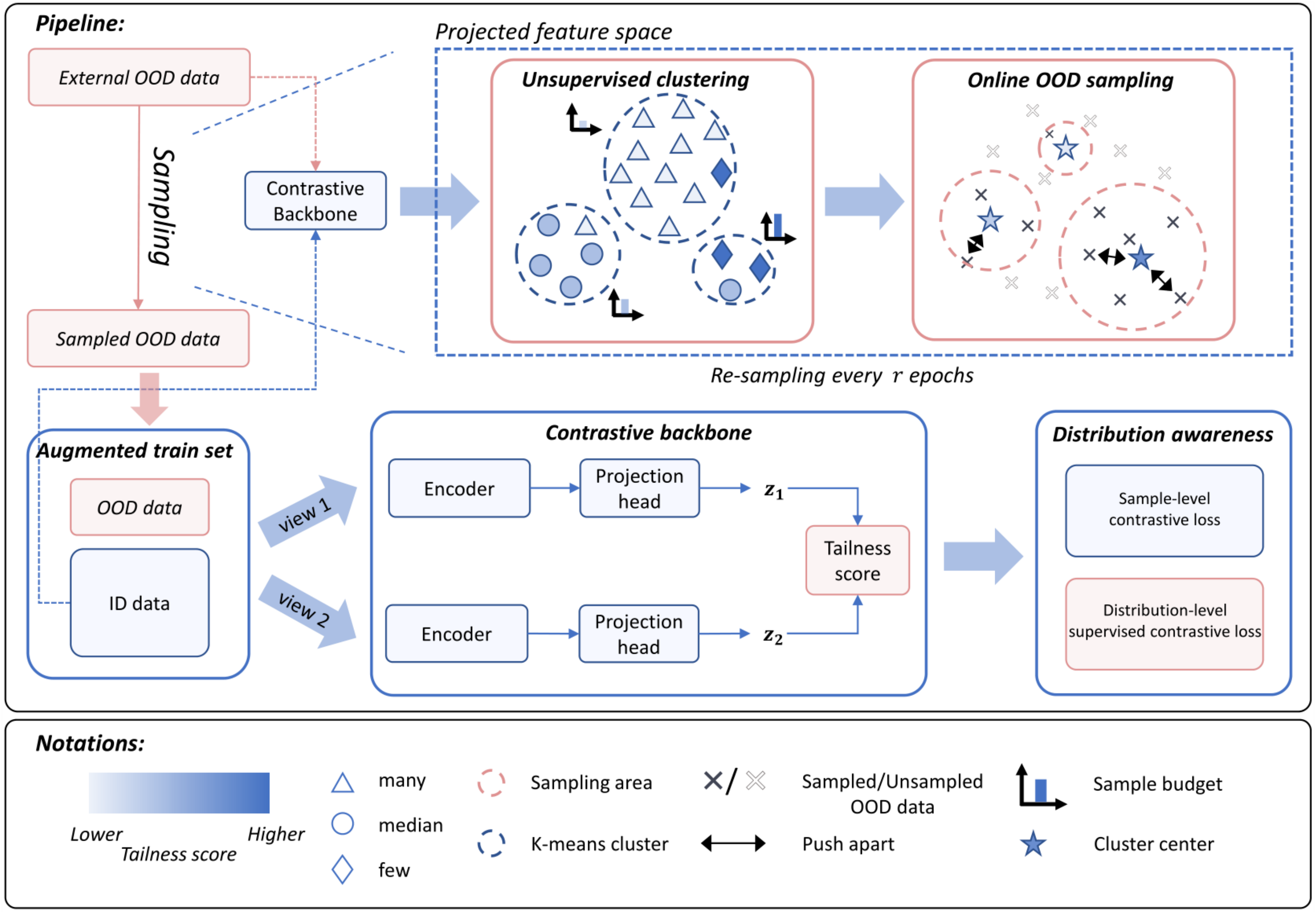


Fig. 1. Overview of **C**ontrastive with **O**ut-of-distribution data for **L**ong-**T**ail learning (COLT). COLT is composed of ‘tailness’ score estimation, dynamic sampling strategies, and an additional contrastive losses for long-tail learning with external OOD samples.

Step1: 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. We observe that the majority classes dominate the feature space and tail instances turn out to be outliers and have lower intra-class consistency. Hence, a sparse neighborhood could be a reliable proxy to identify the tail samples. Specifically, we use top-k% largest negative logits of each sample to depict the feature space neighborhood during training. Given a training sample x_i, its negative logits p_i^- is the following:

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

Then, we define $s_t^i = -\sum_{top-k\%} p_i^-$ as the tailness score for each ID instance x_i. During training, we perform a momentum update to the tailness score:

$$s_t^{i,0} = s_t^i,\ s_t^{i,n} = ms_t^{i,n-1} + (1-m)s_t^{i,n}$$

where $m \in [0,1)$ is the momentum coefficient. A higher value of sitindicates sample x_i has a more sparse neighborhood in the feature space and implies that it belongs to the tail classes with a larger probability.

Step2: 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 z_{c_i} from ID training set S_{id} via K-means clustering. The cluster-wise tailness score $s_t^{c_i}$ is defined as the mean of tailness score in cluster c_i. Then, we obtain each cluster's sampling budget K' as follows:

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

where K refers to the total sampling budget, $K' \in \mathbb{R}^C$ is the sampling budget assigned to each cluster, $\widetilde{s_t^c}$ is the normalized cluster tailness score. To fully exploit the OOD data, we re-sample from the S_{ood} every r epoch. The online sampling process is summarized in Algorithm 2.

Algorithm 2 our online sampling strategy.

Input: ID train set S_{id}, OOD dataset S_{ood}, model θ, sample budget K, cluster number C, similarity metric sim$(\cdot)$, hyper-parameter τ_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 $s_t^{c_i} = \sum_{z_j \in c_i} s_t^j / |c_i|$;
Assign each cluster a sample budget K'_{c_i} with Eq. 3;
Initialize the sample set $S_{sample} = \emptyset$;
for $i = 0, \cdots, C-1$ **do**
  Initialize subset $S^i_{sample} = \emptyset$;
  while $|S_{sample}| < K'_{c_i}$ **do**
    $u = \arg\max_{x_j \in S_{ood}} sim(z_j, z_{c_i})$;
    $S^i_{sample} = S^i_{sample} \cup \{u\}$;
  end while
  $S_{sample} = S_{sample} \cup S^i_{sample}$;
end for
$S_{train} = S_{train} \cup S_{sample}$.

Step3: Awareness of the out-of-distribution data.

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 contrastive loss. However, giving equal treatment to all samples may not be the optimal choice. A natural idea is to let the model be aware of there are two kinds of samples from different domains. Hence, we define an indicator ϕ to provide weakly supervised (distribution only) information:

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

Afterwards, we add a supervised contrastive loss to both ID and OOD samples:

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

where $P(i) \equiv \{p: \phi(x_p) = \phi(x_i)\}$ is the set of indices of the same domain within the mini-batch, $|P(i)|$ is its cardinality and the negative index set $N(i) \equiv \{n: \phi(x_n) \neq \phi(x_i)\}$ contains index from different distribution. Finally, we scale the supervised loss with α and add it to the contrastive loss:

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

Datasets.

CIFAR-10-LT/CIFAR-100-LT are long-tail subsets sampled from the original CIFAR10/CIFAR100. We set the imbalance ratio to 100 in default. Following previous work, we use 300K Random Images as the OOD dataset.

ImageNet-100-LT has 12K images sampled from ImageNet-100 with Pareto distribution. We use ImageNet-R as the OOD dataset.

Places-LT contains about 62.5K images sampled from the large-scale scene-centric Places dataset with Pareto distribution. Places-Extra69 is utilized as the OOD dataset.

Results on improved SSL frameworks for long-tail learning.

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**

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	70.96	65.33	61.89	3.74	67.08	40.02	46.61	49.38	3.93	44.78
+COLT	**75.13**	**71.38**	**66.62**	**3.48**	**72.22**	**41.55**	**48.40**	**50.54**	**3.83**	**46.36**
SDCLR	71.13	66.04	62.31	3.61	67.54	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**

Accuracy, balancedness and versatility.

A step further: OOD vs ID.

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**

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 compare with sampling in an entirely ID dataset.

Take home messages.

1. External unlabeled OOD data can be useful for improving self-supervised learning (SSL) methods performance on long-tailed datasets.
2. We propose a novel SSL pipeline COLT, which extending 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.
3. Future research might should be focusing on how to specify the best OOD dataset that gives the largest improvements for a given long-tail ID dataset.