Frequency-Aware Self-Supervised Long-Tailed Learning

Ci-Siang Lin[1,2]   Min-Hung Chen[2]   Yu-Chiang Frank Wang[1,2]

[1]Graduate Institute of Communication Engineering, National Taiwan University, Taiwan

[2]NVIDIA, Taiwan

{d08942011, ycwang}@ntu.edu.tw, minghungc@nvidia.com

Abstract

*Data collected from the real world typically exhibit long-tailed distributions, where frequent classes contain abundant data while rare ones have only a limited number of samples. While existing supervised learning approaches have been proposed to tackle such data imbalance, the requirement of label supervision would limit their applicability to real-world scenarios in which label annotation might not be available. Without the access to class labels nor the associated class frequencies, we propose **F**requency-**A**ware **S**elf-**S**upervised **L**earning (FASSL) in this paper. Targeting at learning from unlabeled data with inherent long-tailed distributions, the goal of FASSL is to produce discriminative feature representations for downstream classification tasks. In FASSL, we first learn frequency-aware prototypes, reflecting the associated long-tailed distribution. Particularly focusing on rare-class samples, the relationships between image data and the derived prototypes are further exploited with the introduced self-supervised learning scheme. Experiments on long-tailed image datasets quantitatively and qualitatively verify the effectiveness of our learning scheme.*

1. Introduction

Deep neural networks (DNNs) have made remarkable progress in recent years and have been applied in various applications, such as image recognition [13, 34], semantic segmentation [1, 22], image synthesis [4, 24], 3D analysis [8, 28] and video understanding [30, 37]. When abundant labeled data are available for training (i.e., standard supervised scenario), DNNs would achieve satisfactory performance due to its powerful learning ability. However, in real-world applications, collecting labeled data is often costly and time-consuming, which thus limits the scalability and practicality of supervised DNN models. To alleviate this labeling cost, researchers start to explore varying learning strategies using *unlabeled data*, resulting in the recent emergence of self-supervised learning.

Self-supervised learning (SSL) [5, 29, 11, 3] aims to pre-

train DNN models with unlabeled data and to derive their representations, so that downstream tasks can be performed effectively (e.g., image recognition, object detection, etc.). The emergence of self-supervised learning begins from unsupervised representation learning works [10, 25, 33], in which pretext tasks such as rotation prediction [10], jigsaw puzzle solving [25] or image colorization [33] are designed to obtain supervision through data manipulation. As pointed out in [5], these pretext tasks are handcrafted and may not generalize well to downstream tasks. On the other hand, contrastive learning [5, 29] has received increasing research attention for SSL. Its core idea is to create positive pairs from different augmented views of the same image and negative pairs from different images. For instance, SimCLR [5] combines multiple data augmentations and performs contrastive learning by attracting positive features while repelling negative ones. Despite the promising results, these works typically rely on a large number of negative pairs for training purposes. To learn without negative pairs, BYOL [11] attracts positive pairs between the student and teacher network with exponential moving average (EMA) to avoid model collapse. A recent work of SwAV [3] utilizes prototypes to perform clustering, and enforces consistent clustering assignments across positive pairs. While promising results have been presented, these SSL works are not designed to handle data with possible *class imbalance* and may fail to generalize to long-tailed datasets [15].

Long-tailed data learning has been significant in machine learning, where frequent classes contain abundant data while rare ones have only scarce samples. Models learned from such highly imbalanced data would often be biased towards frequent classes and perform poorly on categories with relatively fewer samples. To alleviate this problem, a number of works [12, 7, 6, 27, 36, 26] have been presented and could be divided into two categories: re-sampling and re-weighting. Re-sampling [12, 7, 26] aims to sample instances for each class to eliminate the imbalanced issue. Over-sampling [12, 26] replicates rare-class samples, which could lead to over-fitting, while under-sampling [7] randomly removes frequent-class ones, which may discard

valuable data. Re-weighting [6, 27, 36], on the other hand, assigns larger weights to rare categories during training. An intuitive way is to have weights inverse proportional to class frequencies and therefore amplify the importance of rare-class samples. Specifically, [6] proposes a theoretical framework to estimate the effective number of samples and perform re-weighting. However, such methods require full label supervision to long-tailed data.

Without observing data labels during pre-training, works like [14, 23, 31] investigate the performance of SSL methods on class-imbalanced or long-tailed data. A recent work of [23] also studies the improved robustness of self-supervised learning models compared to supervised ones. Assuming that rare-class samples might be "forgotten" by DNNs during training, SDCLR [15] chooses to perform pruning when training the neural network, simulating the "forgetting" mechanism for producing DNNs which are robust to rare categories. However, since their learning mechanism treats each sample equally, the resulting model might still favor frequent classes.

To address the long-tailed data learning problem without label supervision, we propose a *Frequency-Aware Self-Supervised Learning* (FASSL) scheme in this paper. We present a *Frequency-Aware Prototype Learning* strategy in FASSL, which learns image prototypes from class-imbalanced yet unlabeled data, aiming to reflect the inherent long-tailed distribution. With such derived prototypes, we utilize a teacher-student learning scheme and present *Prototypical Re-balanced Self-Supervised Learning* to train deep neural networks for producing discriminative feature representations. As noted above, this entire learning scheme does *not* observe any label supervision. As confirmed later by our experiments, the frequency-aware prototypes learned by FASSL would properly describe image data with long-tailed distributions. More importantly, the CNN backbones pre-trained by FASSL can be effectively applied for downstream classification tasks, and performs favorably against state-of-the-art SSL or supervised models.

We now highlight our contributions as below:

- We propose *Frequency-Aware Self-Supervised Learning* (FASSL) to pre-train CNNs using unlabeled data with inherent long-tailed distributions.

- In FASSL, we present a *Frequency-Aware Prototype Learning* stage, which identifies frequency-aware prototypes from unlabeled data, reflecting the implicit long-tailed data distribution.

- With the observed image prototypes, our *Prototypical Re-balanced Self-Supervised Learning* trains CNN models from long-tailed yet unlabeled image data, benefiting downstream visual classification tasks.

2. Related Works

2.1. Self-Supervised Learning

In real-world applications, collecting dense annotations could be costly or sometime infeasible. To alleviate such annotation costs, self-supervised learning (SSL) [5, 29, 11, 3] aims to learn data representations from unlabeled data. To introduce discriminative capability for the learned representation, SimCLR [5] pulls positive samples from another augmentation view while pushes negative ones farther away from each other. In order to learn with only positive pairs while preventing models from collapse, BYOL [11] introduces student and teacher networks to attract positive pairs with exponential moving average. On the other hand, SwAV [3] assigns soft clustering codes with prototypes, and it performs swap predictions to enforce consistent clustering assignments. While promising results are presented, existing SSL methods generally assume that the training data are balanced. As verified later in our experiments, such techniques cannot generalize to long-tailed data.

2.2. Learning from Long-Tailed Data

Supervised Long-Tailed Learning. Real-world data typically exhibit long-tailed distributions, where head (frequent) classes contain abundant data while tail (rare) classes contain a limited amount of samples. Learning deep models which generalize well to rare classes is therefore of broad research interest. To address the imbalanced data learning problem, a number of works have been proposed [12, 7, 6, 27, 20, 36, 26, 21, 19, 9]. Specifically, re-sampling approaches [12, 7, 26] design sampling strategies either to remove data from frequent classes (under-sampling) or replicate rare-class samples (over-sampling) to generate a balanced dataset. However, these methods may discard valuable data or lead to over-fitting on sampled rare-class data. On the other hand, re-weighting techniques [6, 27, 36] weight each sample by inverse class frequencies to emphasize rare classes. Despite the promising results, these methods share a common constraint of label supervision (i.e., known label distributions). When such supervision is not available, neither strategy can be easily applied.

Self-Supervised Long-Tailed Learning. Without observing class labels or knowing data frequencies in advance, self-supervised long-tailed learning [2, 15, 17, 35] aims to pre-train deep learning models which generalize to rare categories. Specifically, COLT [2] alleviates class imbalance by adding additional training data sampled from an auxiliary dataset. When such external data are not available, SDCLR [15] emphasizes the learning on rare samples by performing model pruning and enforcing a consistency loss between the pruned model and the original one. On the other hand, BCL [35] designs data augmentation with different
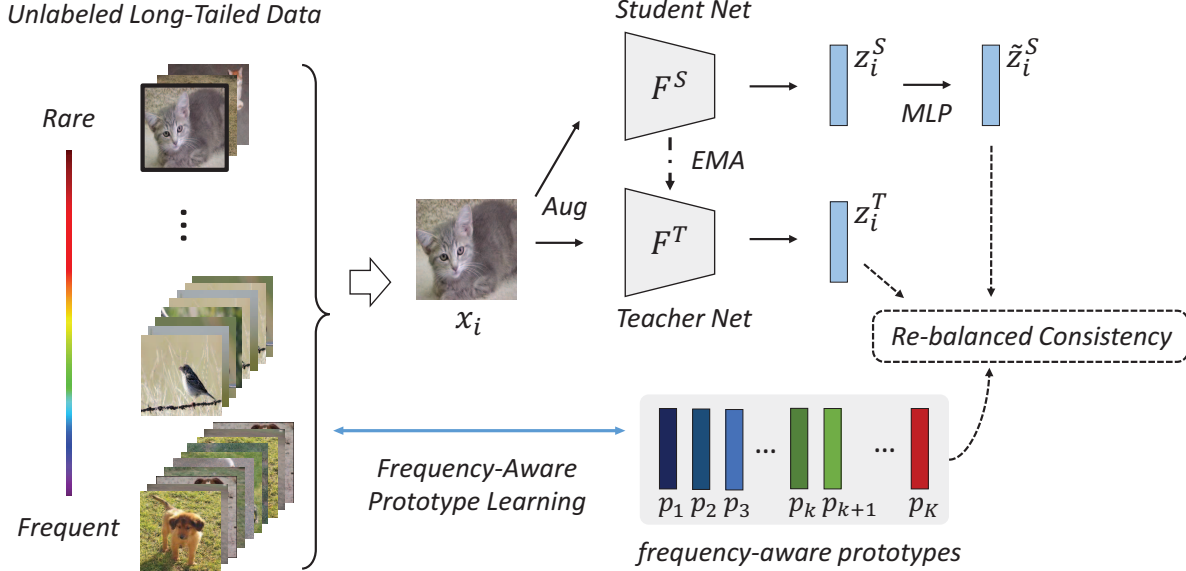
Figure 1. Overview of the proposed *Frequency-Aware Self-Supervised Learning* (FASSL). With image prototypes $\{p_1, p_2, \ldots, p_K\}$ derived to reflect the inherent long-tailed data distribution, each image instance is uniquely exploited into *Prototypical Re-balanced Self-Supervised Learning* via a teacher-student network (F^T and F^S) to perform representation learning.
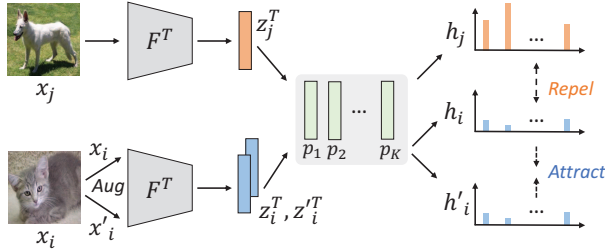

Figure 2. *Frequency-Aware Prototype Learning*. Extended from contrastive learning, we learn image prototypes $\{p_1, p_2, \ldots, p_K\}$ from unlabeled data inputs with objectives allowing the prototypes aligning with the implicit long-tailed distribution.

intensities and applies stronger augmentation to data with higher training losses. However, take BCL [35] as an example, it requires defining a large set of specific augmentation types by heuristic to achieve satisfactory performance.

3. Proposed Method

3.1. Notations and Overview

We first provide the problem definition and the notations used in this paper. Assume there is a set of N images $X = \{x_i\}_{i=1}^N$ with the associated imbalanced label set $Y = \{y_i\}_{i=1}^N$, where $x_i \in \mathbb{R}^{H \times W \times 3}$ and $y_i \in \mathbb{R}$ represent the ith image and its corresponding class label, respectively. Following the standard SSL setting, we do *not* observe the label set Y during pre-training, while X is expected to exhibit a long-tailed distribution. It is worth repeating that, without label supervision, existing approaches

for imbalanced data such as re-sampling [12, 7, 26] and re-weighting [6, 27, 36] cannot be directly applied.

In this paper, we propose *Frequency-Aware Self-Supervised Learning* (FASSL) for pre-training deep learning models using unlabeled long-tailed data. As illustrated in Figure 1, we exploit the inherent imbalanced data distribution *without* any label supervision during the training stage, which allows us to derive discriminative representation for downstream classification tasks. We propose to learn frequency-aware prototypes $\{p_1, p_2, \ldots, p_K\}$ from unlabeled data, which reflect the long-tailed data distribution. With such prototypes obtained, we design self-supervised objectives for learning the CNN backbone model. Note that we have F^T and F^S indicate the teacher and student networks in our framework (parameterized by θ^T and θ^S, respectively). We now detail our proposed method in the following subsections.

3.2. Frequency-Aware Self-Supervised Learning

3.2.1 Frequency-Aware Prototype Learning

To start our proposed learning strategy, we first utilize unlabeled long-tailed data to derive image representatives, aiming to reflect the inherent *imbalanced data distributions*. We view this as the learning stage of *Frequency-Aware Prototype Learning*, which advances contrastive learning technique to produce the desirable image prototypes $\{p_1, p_2, \ldots, p_K\}$. We now detail this learning stage.

In order to derive frequency-aware prototypes describing long-tailed data distribution in an unsupervised setting, we first follow SimCLR [5] and perform data augmentation on

sampled input images x_i to create positive pairs and extract the image features z_i^T and $z_i'^T$ with the network F^T. We note that, rather than directly imposing the contrastive loss on image features z_i^T as SimCLR did, we choose to perform contrastive learning on the *similarity score distribution* level h_i, which is derived from the inner products between x_i and $\{p_1, p_2, \ldots, p_K\}$. Since such inner product operation suggests the similarity between each image and the prototypes, we expect that such derived prototypes would learn as visual exemplars for long-tailed data.

More specifically, we form the prototype matrix P by taking the prototypes in each row, and the resulting matrix size is $K \times D$, indicating K D-dimensional prototypes. We perform matrix-vector product from the matrix P and the feature z_i^T to produce the similarity scores h_i (which could be implemented with a linear layer):

$$h_i = P \cdot z_i^T, \quad \text{where} \quad z_i^T = F^T(x_i). \quad (1)$$

And, a standard contrastive loss L_{contra} is calculated as:

$$L_{contra} = \mathbb{E}_{x_i \sim X} \left[-\log \frac{\exp(sim(h_i, h_i')/\beta)}{\sum_j \exp(sim(h_i, h_j)/\beta)} \right], \quad (2)$$

where $sim(\cdot, \cdot)$ denotes the cosine similarity and β is a hyperparameter of temperature. From the above design, the image prototypes are encouraged to align h_i with its positive sample h_i' (derived from $z_i'^T$) while repelling negative ones h_j. As rare-class data are less frequently sampled and learned, the prototypes are more likely to be updated and to describe frequent categories. As a result, the resulting prototypes would be expected to exhibit the inherent long-tailed distribution.

We note that, both the *network* F^T and *prototypes* $\{p_1, p_2, \ldots, p_K\}$ will be updated by L_{contra} via back propagation. Once complete, the resulting prototypes would implicitly describe the long-tailed data distribution (i.e., a large portion of $\{p_1, p_2, \ldots, p_K\}$ would correspond to frequent classes, while only few are associated with the rare classes). Later in the experiments, we will verify and visualize the derivation of image prototypes at this stage.

3.2.2 Prototypical Re-balanced SSL

As pointed out by [15], existing SSL approaches like [5, 29, 11, 3] are vulnerable to class imbalance, which might fail to generalize to long-tailed data problems. To address this particular challenge, our FASSL utilizes the frequency-aware prototypes derived above as a guidance for learning discriminative representations from unlabeled long-tailed data, as discussed below.

Our proposed strategy is to perform self-supervised learning from unlabeled data with inherent data distribution utilized, while jointly achieving the goal of training

the network F^S for producing discriminative feature representations. Given the aforementioned prototypes implicitly reflecting long-tailed data distribution, we would expect different degrees of similarity when relating each sampled image to these prototypes, depending on its corresponding class frequency. To be more precise, given an input image x_i, we first extract its feature z_i^T from the teacher network F^T (initialized from Section 3.2.1). If x_i is from a rare class, it would be expected to be dissimilar to a majority of prototypes and therefore be viewed as more important during the SSL process. As a result, as a reweighting technique, we calculate the weight $\phi(z_i^T, P)$ for x_i, which is inversely proportional to the similarity sum to all prototypes.

With the above reweighting and regularization strategy, we adopt the asymmetric teacher-student framework to perform *Prototypical Re-balanced Self-Supervised Learning*, as depicted in Figure 1. To avoid possible model collapse when attract positive pairs [11], we further deploy a MLP for the student network F^S. It can be seen that we encourage the semantic similarity between $\tilde{z}_i^S$ and z_i^T derived from F^S and F^T with the re-balanced self-supervised consistency loss L_{reb}:

$$L_{reb} = \mathbb{E}_{x_i \sim X} \left[\phi(z_i^T, P) \cdot L_{consis}(\tilde{z}_i^S, z_i^T) \right],$$

$$\text{where} \quad \phi(z_i^T, P) = \frac{1}{exp(\sum_k \langle z_i^T, p_k \rangle)} \quad (3)$$

$$\text{and} \quad L_{consis} = \left\| \frac{\tilde{z}_i^S}{\|\tilde{z}_i^S\|_2} - \frac{z_i^T}{\|z_i^T\|_2} \right\|_2^2.$$

Thus, the student network F^S would be updated by the above re-balanced loss L_{reb}, while the teacher network F^T is updated by exponential moving average (EMA) from F^S:

$$\theta_S \leftarrow \theta_S - \gamma \frac{\partial L_{reb}}{\partial \theta_S} \quad \text{and}$$

$$\theta_T \leftarrow \tau \cdot \theta_T + (1 - \tau) \cdot \theta_S, \quad (4)$$

where τ controls the decay rate of the teacher network F^T, and θ_S and θ_T are parameters of F^S and F^T. It can be seen that, with the prototypes derived in Section 3.2.1 and the introduced re-balanced loss L_{reb}, we are able to identify rare-class samples and focus on the associated representation learning. It is also worth repeating that, the above learning scheme is implemented with no label supervision.

Finally, we note that we do not alternate between the above two stages during training. This is because, alternative optimization between these two stages tends to hinder the prototypes from describing the associated long-tailed distribution. With F^T initialized by SDCLR [15], we simply jointly train F^T and produce the prototypes $\{p_1, p_2, \ldots, p_K\}$ in *Frequency-Aware Prototype Learn-*

Algorithm 1: Training of FASSL
---
1 **Input**: Images $X = \{x_i\}_{i=1}^N$ with long-tailed class distribution

2 *Frequency-Aware Prototype Learning*

3 $\{p_1, p_2, \ldots, p_K\} \leftarrow$ randomly initialize

4 $\theta_T \leftarrow$ initialize from [15]

5 **for** *num. of iterations* **do**

6      $x_i, x_i', x_j \leftarrow$ random sample from X

7      $h_i, h_i', h_j \leftarrow$ derived by (1) with F^T and P

8      $L_{contra} \leftarrow$ calculated by (2)

9      $\theta_T, p_1, p_2, \ldots, p_K \leftarrow$ update by L_{contra}

10 **end**

11 *Prototypical Re-balanced Self-Supervised Learning*

12 $\theta_S \leftarrow$ initialize from θ_T

13 **for** *num. of iterations* **do**

14      $x_i, x_i' \leftarrow$ random sample from X

15      $\tilde{z}_i^S, z_i^T \leftarrow$ produced by F^T and F^S

16      $L_{reb} \leftarrow$ calculated by (3)

17      $\theta_S, \theta_T \leftarrow$ update by L_{reb} and EMA (4)

18 **end**

19 **Output**: the student network F^S
---

ing. As for the stage of *Prototypical Re-balanced Self-Supervised Learning*, we perform the proposed re-weighted SSL to learn the network F^S for data with long-tailed distribution. The pseudo code for our FASSL is summarized in Algorithm 1.

4. Experiments

4.1. Experimental Settings

Datasets. Following [15] and [6], we consider the benchmarks of Long-Tailed CIFAR-10/CIFAR-100 and Long-Tail ImageNet-100 for experiments. The original CIFAR-10/CIFAR-100 [16] datasets contain 10/100 classes in a total of $60,000$ 32×32 images. To create the long-tailed setting, [6] samples imbalanced subsets from the originals to create Long-Tailed CIFAR-10/CIFAR-100. The imbalance factor is defined by the number of samples in the most frequent class divided by the least one. Moreover, we follow SDCLR [15] and set the imbalance factor as 100, which makes the imbalanced problem challenging. For both datasets, we randomly choose one of the five splits for the experiments. As for Long-Tail ImageNet-100, the sample number per class ranges from 1280 to 5. In addition, we also evaluate the proposed method on Tiny-ImageNet [18]. Tiny-ImageNet contains 200 classes and each class contains 500 training images. We sample an imbalanced pre-training subset from the training split for the long-tailed setting and

take the validation split as the testing data.

Settings and Evaluation. We consider the standard linear evaluation in self-supervised learning to measure the quality of learned representations. That is, we freeze the parameters of the backbone model and fine-tune a linear classifier on top of it. Since our student network F^S contains a ResNet-18 [13] as the CNN backbone model and an additional MLP as the projection head, we follow SimCLR [5] and remove the projection head during fine-tuning. In addition to standard pre-train/fine-tune setting, we follow SDCLR [15] and consider the *few-shot* setting, where only 1% data in the standard setting are used to fine-tune the classifier. Depending on the sample frequencies, we divide all classes into three groups, *Frequent*, *Medium* and *Rare*. *Frequent* stands for the top thrid of most frequent classes, while *Rare* stands for the lowest third. We report the accuracy of each group and also the average and standard deviation of three groups.

4.2. Implementation Details

We perform data augmentation of random cropping, horizontal flipping, color jittering and random gray scaling for pre-training, with ResNet-18/50 [13] as the backbone. We set the temperature β as 0.2 and the number of prototypes K as 128 by default. We clip outliers and normalize the weights in each mini-batch for stability. For fine-tuning, we use the standard cross-entropy loss and train 30/100 epochs for the standard and few-shot setting, respectively.

4.3. Quantitative Comparisons

In Table 1, we perform linear evaluation on Long-Tailed CIFAR-10/CIFAR-100, Long-Tail ImageNet-100 and Tiny-ImageNet. We note that, while no label supervision during pre-training, neither SimCLR [5] nor SwAV [3] consider the long-tailed or imbalanced setting. From the results shown in Table 1, we see that our approach achieved the best performance compared to existing self-supervised methods. Specifically, we achieved the accuracy of 45.85% in average and 43.15% on rare classes on Tiny-ImageNet. Since our approach explicitly weighted rare samples higher to tackle class imbalance, we performed favorably against SDCLR [15] by 1% on rare classes. It is worth noting that, since BYOL [11] adopts teacher-student learning with a uniform-weighting loss, it could be viewed as an ablation study of our re-balanced loss L_{reb}. We see that BYOL only reported 34.91% on rare classes, which is over 8% lower compared to our method, verifying the effectiveness of our loss L_{reb} on unlabeled long-tailed data. We also note that, BCL-I [35] requires defining a large set of specific augmentation types by heuristic. When only common augmentation types in SSL are applied (as we do), BCL-I reported the rare-class accuracy of 48.27% on CIFAR100-LT, which is over 5% lower compared to our FASSL. In ad-

Table 1. Evaluation on Tiny-ImageNet, ImageNet-100-LT, CIFAR100-LT and CIFAR10-LT with the standard setting (i.e., use of all labeled data for fine-tuning). ↑ denotes the higher the better, and ↓ denotes the lower the better. **Bold** denotes the best averaged results except for the supervised method. †: Note that BYOL [11] can be viewed as an ablation study of ours (*i.e.*, teacher-student learning with an uniform-weighting loss).

Dataset	Method	All ↑	Rare ↑	Medium ↑	Frequent ↑	Std ↓
Tiny-ImageNet	SimCLR [5]	32.11	32.12	31.73	32.47	0.30
	SwAV [3]	33.21	32.15	30.61	36.88	2.67
	BYOL† [11]	37.58	34.91	35.45	42.38	3.40
	SDCLR [15]	45.23	42.15	43.39	50.15	3.51
	FASSL (Ours)	**45.85**	43.15	43.88	50.53	3.32
	Supervised	43.57	40.00	40.45	50.26	4.74
ImageNet-100-LT	SimCLR [5]	65.46	59.69	63.71	69.54	4.04
	SDCLR [15]	66.48	60.92	65.04	70.10	3.75
	FASSL (Ours)	**68.92**	66.00	68.06	72.71	2.81
CIFAR100-LT	LDAM-DRW+SSP [31]	43.43	-	-	-	-
	SwAV [3]	47.00	44.00	47.03	49.97	2.44
	BYOL† [11]	48.86	45.55	47.42	53.62	3.45
	SimCLR [5]	49.76	47.58	50.36	51.35	1.60
	BCL-I [35]	52.22	48.27	53.03	55.35	2.95
	SDCLR [15]	54.94	51.00	55.03	58.79	3.18
	FASSL (Ours)	**55.27**	53.55	54.52	57.74	1.79
	Supervised	54.06	51.39	54.18	56.62	2.13
CIFAR10-LT	SimCLR [5]	75.37	69.33	73.33	83.45	5.94
	BYOL† [11]	75.66	75.43	69.83	81.70	4.85
	SwAV [3]	76.60	73.30	71.10	85.40	6.29
	LDAM-DRW+SSP [31]	77.83	-	-	-	-
	SDCLR [15]	80.49	75.10	78.07	88.30	5.66
	FASSL (Ours)	**80.69**	78.80	76.73	86.55	4.23
	Supervised	80.76	75.43	80.93	85.93	4.95

Table 2. Evaluation on CIFAR100-LT/CIFAR10-LT with the few-shot setting (i.e., use of 1% of labeled data for fine-tuning). †: Note that BYOL [11] could be viewed as an ablation study of ours (*i.e.*, teacher-student learning with an uniform-weighting loss).

Dataset	Method	All ↑	Rare ↑	Medium ↑	Frequent ↑	Std ↓
CIFAR100-LT	SwAV [3]	20.14	13.91	20.06	26.44	5.12
	BYOL† [11]	20.62	14.03	19.61	28.24	5.84
	SimCLR [5]	22.51	15.45	22.48	29.59	5.77
	SDCLR [15]	25.39	19.91	25.52	30.74	4.42
	FASSL (Ours)	**27.11**	21.18	27.88	32.26	4.56
	Supervised	28.97	13.88	32.03	41.00	11.28
CIFAR10-LT	BYOL† [11]	62.33	50.40	60.53	76.05	10.55
	SimCLR [5]	62.98	56.67	62.97	69.30	5.16
	SwAV [3]	65.16	52.90	61.97	80.60	11.53
	SDCLR [15]	67.23	56.63	61.87	83.20	11.49
	FASSL (Ours)	**68.54**	62.43	61.93	81.25	8.99
	Supervised	69.82	65.07	66.47	77.93	5.76

Table 3. Hyper-parameter analysis on the number of prototypes K (**left**) and the decay rate τ of EMA (**right**).

| Dataset	K	All ↑	Rare ↑	Medium ↑	Frequent ↑	τ	All ↑	Rare ↑	Medium ↑	Frequent ↑
	64	54.94	52.67	54.30	57.85	0.0	33.35	30.15	34.88	35.03
	96	54.99	52.15	54.24	58.59	0.9	53.63	50.91	52.97	57.00
CIFAR100-LT	128	**55.27**	53.55	54.52	57.74	0.99	**55.27**	53.55	54.52	57.74
	256	54.99	52.15	54.88	57.94	0.999	55.02	51.70	55.00	58.35
	512	55.22	52.88	54.61	58.18	1.0	53.00	50.73	53.24	55.03
	64	80.69	78.43	75.33	88.30	0.0	49.74	44.27	47.83	57.13
	96	**80.81**	77.07	77.20	88.15	0.9	79.74	75.40	76.57	87.25
CIFAR10-LT	128	80.69	78.80	76.73	86.55	0.99	**80.69**	78.80	76.73	86.55
	256	80.79	78.97	75.53	87.88	0.999	80.16	77.57	74.30	88.63
	512	80.69	77.13	76.70	88.05	1.0	79.63	75.33	75.83	87.73
	64	45.80	43.06	43.76	50.59	0.0	30.70	25.61	27.61	38.88
	96	45.50	43.03	43.42	50.06	0.9	**45.86**	42.76	44.12	50.71
Tiny-ImageNet	128	45.85	43.15	43.88	50.53	0.99	45.85	43.15	43.88	50.53
	256	**46.11**	43.18	44.09	51.06	0.999	45.75	42.91	43.94	50.41
	512	45.72	42.97	43.70	50.50	1.0	45.83	42.48	44.48	50.53
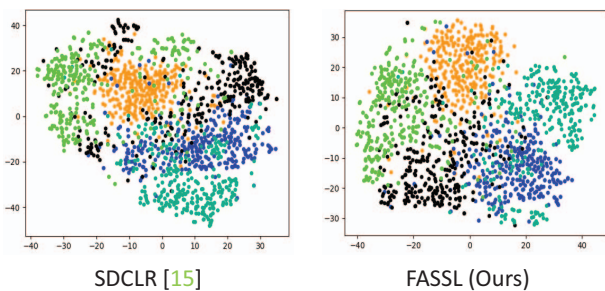
Figure 3. T-SNE of latent representations on CIFAR10-LT. Different colors indicate different categories. Compared with SDCLR [15], our FASSL results in improved discriminative representations.

dition, our method is even *higher* than supervised learning on CIFAR100-LT, which further verifies the effectiveness of our *Frequency-Aware Self-Supervised Learning* framework.

We further consider the challenging few-shot setting, in which only 1% labeled data are used to train the linear classifier. This would assess the model ability in deriving robust representations and avoiding over-fitting on few labeled data. In Table 2, we observe that our model again performed favorably against existing methods under the few-shot setting. In particular, we achieved 62.43% on rare-classes on CIFAR10-LT, which is over 5% higher compared to SDCLR [15]. To further verify the effectiveness of our approach, we provide qualitative comparisons as shown in Figure 3. We select five categories from CIFAR10-LT and apply t-SNE visualization for the learned features of the student network F^S. We see that the representations derived by our scheme are better separated according to the associated categories.

4.4. Analysis of Hyper-parameters

Impact of the number of prototypes K. In Table 3 (left), we conduct sensitivity analysis with respect to the number of prototypes K. When varying K from 64 to 512 on CIFAR100-LT, we see that our model produced consistent results within a 0.5% drop compared to the best average accuracy. It is worth noting that, different from other SSL works [32] which utilize prototypes to perform clustering, we do *not* require knowing the number of classes in advance to set the number of prototypes K. Thus, we simply set K to 128 throughout our experiments.

Impact of the decay rate τ of EMA. In our proposed FASSL, we adopt exponential moving average (EMA) to update the teacher network F^T with the decay rate τ. We now analyze the impact of τ in Table 3 (right). When setting $\tau = 0$, no moving average is adopted, and the teacher network F^T shares the same parameters with the student network F^S at all times. As noted in [11], attracting positive pairs between identical feature encoders could lead to model collapse and therefore yielded the poor performance of 33.35% on CIFAR100-LT. When the decay rate $\tau = 1$, the teacher network F^T is frozen and never updated, resulting in suboptimal results. In practice, we set $\tau = 0.99$ to progressively distill the knowledge from the student network F^S while avoiding collapse.

4.5. Analysis of Long-Tailed Prototypes

In *Frequency-Aware Prototype Learning*, we learn prototypes to reflect the long-tailed distribution from unlabeled data. With the imbalance factor $\rho = 100$, the amount of frequent samples is about ten times of that of the medium

Frequent (116 prototypes)
- frog (52)
- ship (32)
- deer (20)
- automobile (12)

Medium (10 prototypes)
- horse (5)
- airplane (3)
- cat (2)

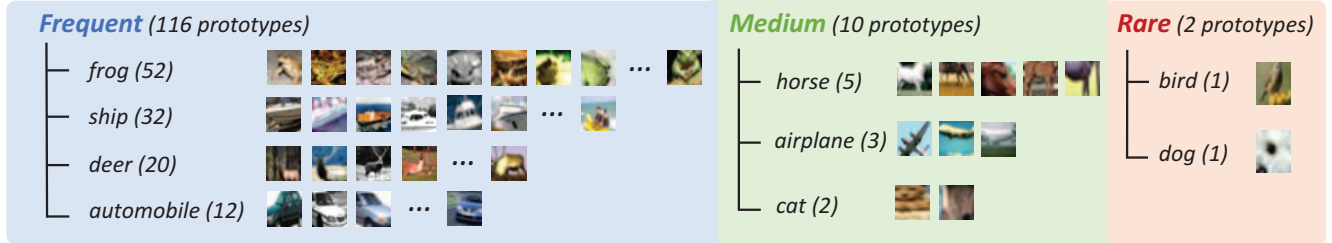**Rare (2 prototypes)**
- bird (1)
- dog (1)

Figure 4. Visualization of 128 frequency-aware prototypes and the associate classes on CIFAR10-LT. The prototypes are visualized by retrieving their closest images in the latent space. Note that the number in each parenthesis denotes the number of prototypes in each group/category.

Table 4. Different numbers of prototypes K and the corresponding **class distributions** on Tiny-ImageNet, CIFAR100-LT and CIFAR10-LT with the imbalanced factor ρ set as 100.

Dataset	K	Frequent	Medium	Rare
Tiny-ImageNet	64	78.13%	21.87%	0.00%
	96	86.46%	13.54%	0.00%
	128	84.38%	14.06%	1.56%
	256	84.77%	13.67%	1.56%
	512	89.26%	9.77%	0.97%
CIFAR100-LT	64	89.06%	7.81%	3.13%
	96	82.29%	15.63%	2.08%
	128	88.28%	10.16%	1.56%
	256	85.94%	12.89%	1.17%
	512	91.21%	6.84%	1.95%
CIFAR10-LT	64	90.63%	9.38%	0.00%
	96	90.63%	8.33%	1.04%
	128	90.63%	7.81%	1.56%
	256	88.28%	9.38%	2.34%
	512	85.35%	12.50%	2.15%

Table 5. **Class distributions** of prototypes when varying imbalanced factor $\rho = 1 - 500$.

Dataset	ρ	Frequent	Medium	Rare
Tiny-ImageNet	1	34.38%	33.59%	32.03%
	20	72.66%	21.09%	6.25%
	100	84.38%	14.06%	1.56%
	200	87.50%	10.94%	1.56%
	500	91.41%	7.81%	0.78%

increased. This validates that our prototypes effectively describe the data distribution, and thus are applicable for our *Prototypical Re-balanced Self-Supervised Learning*.

5. Conclusion

Learning representations from long-tailed data has been among active research topics for machine learning communities, and it becomes particularly challenging when no label supervision is available during training (or pre-training) stage. In this paper, we proposed *Frequency-Aware Self-Supervised Learning* (FASSL), which exploits and identifies the inherent imbalanced data distribution to derive frequency-aware prototypes. While self-supervised learning can be easily deployed on the unlabeled input data, our FASSL further utilizes the aforementioned image prototypes for guiding the learning process, which aligns with the data distributions while producing desirable image representations. Once the network backbone is pre-trained by FASSL, downstream classification tasks can be tackled with satisfactory performances. Experiments on long-tailed benchmarks confirmed the effectiveness of our FASSL against state-of-the-art methods, while ablation studies and analyses were conducted to verify and visualize our derived prototypes and representations.

ones (same for medium vs. rare classes). Therefore, most of the prototypes are expected to represent frequent classes. To confirm this, we take 128 prototypes to retrieve closest images in the latent space and visualize them in Figure 4. We see that a total of 116 prototypes were from frequent-class images (e.g., frog, ship, etc.), and only 2 were related to rare classes of bird and dog on CIFAR10-LT.

In addition to visualization, we also present the class distributions of the retrieved images in Table 4. With $K = 64$ on CIFAR10-LT, we observe that the number of prototypes may not be sufficient to cover rare classes. With $K = 128$ and above, the distributions were aligned with the long-tailed data. Similar results are observed on CIFAR100-LT and Tiny-ImageNet. In Table 5, we further vary the imbalanced factor ρ from 1 to 500 to control the degree of class imbalance and report the corresponding class distributions of prototypes on Tiny-ImageNet. We see that when $\rho = 1$ indicating no class imbalance presents, the class distribution was nearly uniform as desired. With larger ρ, the prototypes number ratio of frequency classes over rare class was

References

[1] Vijay Badrinarayanan, Alex Kendall, and Roberto Cipolla. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE transactions on pattern analysis and machine intelligence*, 39(12):2481–2495, 2017. 1

[2] Jianhong Bai, Zuozhu Liu, Hualiang Wang, Jin Hao, YANG FENG, Huanpeng Chu, and Haoji Hu. On the effectiveness of out-of-distribution data in self-supervised long-tail learning. In *The Eleventh International Conference on Learning Representations*, 2022. 2

[3] Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. *Advances in Neural Information Processing Systems*, 33:9912–9924, 2020. 1, 2, 4, 5, 6

[4] Qifeng Chen and Vladlen Koltun. Photographic image synthesis with cascaded refinement networks. In *Proceedings of the IEEE international conference on computer vision*, pages 1511–1520, 2017. 1

[5] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020. 1, 2, 3, 4, 5, 6

[6] Yin Cui, Menglin Jia, Tsung-Yi Lin, Yang Song, and Serge Belongie. Class-balanced loss based on effective number of samples. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9268–9277, 2019. 1, 2, 3, 5

[7] Chris Drummond, Robert C Holte, et al. C4. 5, class imbalance, and cost sensitivity: why under-sampling beats over-sampling. In *Workshop on learning from imbalanced datasets II*, volume 11, pages 1–8. Citeseer, 2003. 1, 2, 3

[8] Kai Fischer, Martin Simon, Florian Olsner, Stefan Milz, Horst-Michael Gross, and Patrick Mader. Stickypillars: Robust and efficient feature matching on point clouds using graph neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 313–323, June 2021. 1

[9] Siming Fu, Huanpeng Chu, Xiaoxuan He, Hualiang Wang, Zhenyu Yang, and Haoji Hu. Meta-prototype decoupled training for long-tailed learning. In *Proceedings of the Asian Conference on Computer Vision*, pages 569–585, 2022. 2

[10] Spyros Gidaris, Praveer Singh, and Nikos Komodakis. Unsupervised representation learning by predicting image rotations. In *International Conference on Learning Representations*, 2018. 1

[11] 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. 1, 2, 4, 5, 6, 7

[12] Hui Han, Wen-Yuan Wang, and Bing-Huan Mao. Borderline-smote: a new over-sampling method in imbalanced data sets learning. In *International conference on intelligent computing*, pages 878–887. Springer, 2005. 1, 2, 3

[13] 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*, pages 770–778, 2016. 1, 5

[14] 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, 2021. 2

[15] Ziyu Jiang, Tianlong Chen, Bobak J Mortazavi, and Zhangyang Wang. Self-damaging contrastive learning. In *International Conference on Machine Learning*, pages 4927–4939. PMLR, 2021. 1, 2, 4, 5, 6, 7

[16] Alex Krizhevsky and Geoffrey Hinton. Learning multiple layers of features from tiny images. Technical Report 0, University of Toronto, Toronto, Ontario, 2009. 5

[17] Anna Kukleva, Moritz Böhle, Bernt Schiele, Hilde Kuehne, and Christian Rupprecht. Temperature schedules for self-supervised contrastive methods on long-tail data. In *The Eleventh International Conference on Learning Representations*, 2023. 2

[18] Ya Le and Xuan Yang. Tiny imagenet visual recognition challenge. *CS 231N*, 7(7):3, 2015. 5

[19] Jun Li, Zichang Tan, Jun Wan, Zhen Lei, and Guodong Guo. Nested collaborative learning for long-tailed visual recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6949–6958, 2022. 2

[20] Tianhong Li, Peng Cao, Yuan Yuan, Lijie Fan, Yuzhe Yang, Rogerio S Feris, Piotr Indyk, and Dina Katabi. Targeted supervised contrastive learning for long-tailed recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6918–6928, 2022. 2

[21] Tianhao Li, Limin Wang, and Gangshan Wu. Self supervision to distillation for long-tailed visual recognition. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 630–639, 2021. 2

[22] Chenxi Liu, Liang-Chieh Chen, Florian Schroff, Hartwig Adam, Wei Hua, Alan L Yuille, and Li Fei-Fei. Auto-deeplab: Hierarchical neural architecture search for semantic image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 82–92, 2019. 1

[23] Hong Liu, Jeff Z HaoChen, Adrien Gaidon, and Tengyu Ma. Self-supervised learning is more robust to dataset imbalance. In *NeurIPS 2021 Workshop on Distribution Shifts: Connecting Methods and Applications*, 2021. 2

[24] Moustafa Meshry, Yixuan Ren, Larry S. Davis, and Abhinav Shrivastava. Step: Style-based encoder pre-training for multi-modal image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3712–3721, June 2021. 1

[25] Mehdi Noroozi and Paolo Favaro. Unsupervised learning of visual representations by solving jigsaw puzzles. In *European conference on computer vision*, pages 69–84. Springer, 2016. 1

[26] Seulki Park, Youngkyu Hong, Byeongho Heo, Sangdoo Yun, and Jin Young Choi. The majority can help the minority: Context-rich minority oversampling for long-tailed classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6887–6896, 2022. 1, 2, 3

[27] Seulki Park, Jongin Lim, Younghan Jeon, and Jin Young Choi. Influence-balanced loss for imbalanced visual classification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 735–744, 2021. 1, 2, 3

[28] Charles R Qi, Hao Su, Kaichun Mo, and Leonidas J Guibas. Pointnet: Deep learning on point sets for 3d classification and segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 652–660, 2017. 1

[29] Aaron Van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv e-prints*, pages arXiv–1807, 2018. 1, 2, 4

[30] Chao-Yuan Wu and Philipp Krahenbuhl. Towards long-form video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1884–1894, June 2021. 1

[31] 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. 2, 6

[32] Xiangyu Yue, Zangwei Zheng, Shanghang Zhang, Yang Gao, Trevor Darrell, Kurt Keutzer, and Alberto Sangiovanni Vincentelli. Prototypical cross-domain self-supervised learning for few-shot unsupervised domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13834–13844, 2021. 7

[33] Richard Zhang, Phillip Isola, and Alexei A Efros. Colorful image colorization. In *European conference on computer vision*, pages 649–666. Springer, 2016. 1

[34] Heliang Zheng, Jianlong Fu, Zheng-Jun Zha, and Jiebo Luo. Looking for the devil in the details: Learning trilinear attention sampling network for fine-grained image recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5012–5021, 2019. 1

[35] Zhihan Zhou, Jiangchao Yao, Yan-Feng Wang, Bo Han, and Ya Zhang. Contrastive learning with boosted memorization. In *International Conference on Machine Learning*, pages 27367–27377. PMLR, 2022. 2, 3, 5, 6

[36] Jianggang Zhu, Zheng Wang, Jingjing Chen, Yi-Ping Phoebe Chen, and Yu-Gang Jiang. Balanced contrastive learning for long-tailed visual recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6908–6917, 2022. 1, 2, 3

[37] Mohammadreza Zolfaghari, Kamaljeet Singh, and Thomas Brox. Eco: Efficient convolutional network for online video understanding. In *Proceedings of the European conference on computer vision (ECCV)*, pages 695–712, 2018. 1