SuperDisco: Super-Class Discovery Improves Visual Recognition for the Long-Tail

Yingjun Du[1], Jiayi Shen[1], Xiantong Zhen[1,2*], Cees G. M. Snoek[1]

[1]AIM Lab, University of Amsterdam [2]Inception Institute of Artificial Intelligence

Abstract

Modern image classifiers perform well on populated classes, while degrading considerably on tail classes with only a few instances. Humans, by contrast, effortlessly handle the long-tailed recognition challenge, since they can learn the tail representation based on different levels of semantic abstraction, making the learned tail features more discriminative. This phenomenon motivated us to propose SuperDisco, an algorithm that discovers super-class representations for long-tailed recognition using a graph model. We learn to construct the super-class graph to guide the representation learning to deal with long-tailed distributions. Through message passing on the super-class graph, image representations are rectified and refined by attending to the most relevant entities based on the semantic similarity among their super-classes. Moreover, we propose to meta-learn the super-class graph under the supervision of a prototype graph constructed from a small amount of imbalanced data. By doing so, we obtain a more robust super-class graph that further improves the long-tailed recognition performance. The consistent state-of-the-art experiments on the long-tailed CIFAR-100, ImageNet, Places and iNaturalist demonstrate the benefit of the discovered super-class graph for dealing with long-tailed distributions.

1. Introduction

This paper strives for long-tailed visual recognition. A computer vision challenge that has received renewed attention in the context of representation learning, as real-world deployment demands moving from balanced to imbalanced scenarios. Three active strands of work involve class re-balancing [15, 22, 32, 43, 65], information augmentation [34, 51, 54] and module improvement [29, 31, 76]. Each of these strands is intuitive and has proven empirically successful. However, all these approaches seek to improve the classification performance of the original feature space. In this paper, we instead explore a graph learning algorithm to discover the imbalanced super-class space hidden in the original feature representation.
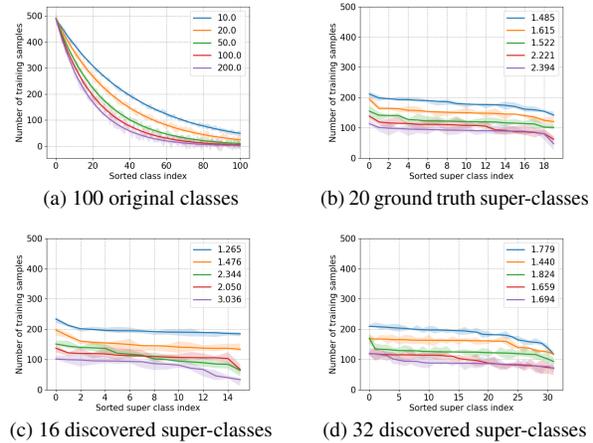
Figure 1. **SuperDisco learns to project the original class space (a) into a relatively balanced super-class space.** Different color curves indicate the different imbalance factors on the long-tailed CIFAR-100 dataset. Like the 20 super-class ground truth (b) our discovered super-classes for 16 super-classes (c) or 32 super-classes (d) provide a much better balance than the original classes.

The fundamental problem in long-tailed recognition [18, 32, 44, 77] is that the head features and the tail features are indistinguishable. Since the head data dominate the feature distribution, they cause the tail features to fall within the head feature space. Nonetheless, humans effortlessly handle long-tailed recognition [2, 16] by leveraging semantic abstractions existing in language to gain better representations of tail objects. This intuition hints that we may discover the semantic hierarchy from the original feature space and use it for better representations of tail objects. Moreover, intermediate concepts have been shown advantageous for classification [5, 36] by allowing the transfer of shared features across classes. Nevertheless, it remains unexplored to exploit intermediate super-classes in long-tailed visual recognition that rectify and refine the original features.

In the real world, each category has a corresponding super-class, *e.g.*, *bus*, *taxi*, and *train* all belong to the *vehicle* super-class. This observation raises the question: *are super-classes of categories also distributed along a long-tail?* We find empirical evidence that within the super-class space of popular datasets, the long-tailed distribution almost disappears, and each super-class has essentially the same number of samples.

---

*Currently with United Imaging Healthcare, Co., Ltd., China.

In Figure 1, we show the number of training samples for each of the original classes and their corresponding super-classes in the long-tailed CIFAR-100 dataset. We observe the data imbalance of super-classes is considerably lower than those of the original classes. This reflects the fact that the original imbalanced data hardly affects the degree of imbalance of the super-classes, which means the distribution of the super-classes and original data is relatively independent. These balanced super-class features could be used to guide the original tail data away from the dominant role of the head data, thus making the tail data more discriminative. Therefore, if the super-classes on different levels of semantic abstraction over the original classes can be accurately discovered, it will help the model generalize over the tail classes. As not all datasets provide labels for super-classes, we propose to learn to discover the super-classes in this paper.

Inspired by the above observation, we make in this paper two algorithmic contributions. First, we propose in Section 3 an algorithm that learns to discover the super-class graph for long-tailed visual recognition, which we call SuperDisco. We construct a learnable graph that discovers the super-class in a hierarchy of semantic abstraction to guide feature representation learning. By message passing on the super-class graph, the original features are rectified and refined, which attend to the most relevant entities according to the similarity between the original image features and super-classes. Thus, the model is endowed with the ability to free the original tail features from the dominance of the head features using the discovered and relatively balanced super-class representations. Even when faced with the severe class imbalance challenges, e.g., iNaturalist, our SuperDisco can still refine the original features by finding a more balanced super-class space using a more complex hierarchy. As a second contribution, we propose in Section 4 a meta-learning variant of our SuperDisco algorithm to discover the super-class graph, enabling the model to achieve even more balanced image representations. To do so, we use a small amount of balanced data to construct a prototype-based relational graph, which captures the underlying relationship behind samples and alleviates the potential effects of abnormal samples. Last, in Section 5 we report experiments on four long-tailed benchmarks: CIFAR-100-LT, ImageNet-LT, Places-LT, and iNaturalist, and verify that our discovered super-class graph performs better for tail data in each dataset. Before detailing our contributions, we first embed our proposal in related work.

2. Related work

Long-tailed recognition. Several strategies have been proposed to address class imbalance in recognition. We categorize them into three groups. Those in the first group are based on class re-balancing [8, 30, 44, 75], which balance the training sample numbers of different classes during model training. Class re-balancing methods also could be catego-

rized into three different groups, i.e., re-sampling [22, 32, 43, 67], cost-sensitive learning [14, 37, 56, 78, 79, 86] and logit adjustment [27, 45, 57, 59]. Class re-balancing methods improve the performance of the tail classes at the expense of the performance of the head classes. The second group is based on information augmentation, introducing additional information into model training to improve long-tailed learning performance. We identify four kinds of methods in the information augmentation scope, i.e., transfer learning, which includes head-to-tail knowledge transfer [6, 42, 64, 73], knowledge distillation [28, 40, 71], model pre-training [9, 33, 72] and self-training [24, 68, 74]. The third group focuses on improving network modules in long-tailed learning. This group includes representation learning [13, 46, 76], classifier learning [32, 41, 42, 69, 73], decoupled training [31, 32, 82], and ensemble learning [20, 83]. These methods introduce additional computation costs for increased performance. Our method belongs to the third group as it aims to learn a better representation of unbalanced training samples by the super-class graph, which is unexplored for long-tail recognition.

Super-class learning. Super-class learning adds super-class labels as intermediate supervision into traditional deep learning. A super-class guided network [38] integrated the high-level semantic information into the network for image classification and object detection, which took two-level class annotations that contain both super-class and finer class labels. In [11], a two-phase multi-expert architecture was proposed for still image action recognition, which includes fine-grained and coarse-grained phases. However, they leveraged the ground truth of the super-class as supervision during the coarse-grained phase. Wu et al. [70] propose a taxonomic classifier to address the long-tail recognition problem, which classified each sample to the level that the classifier is competent. Zhou et al. [85] clustered the original categories into super-classes to produce a relatively balanced distribution in the super-class space, which also leveraged the ground truth of the super-class in the training phase. In contrast with the previous super-class learning, we do not use ground truth to group the original categories into the super-class space. To the best of our knowledge, no work exists that relies on graph learning to discover the super-class for long-tailed visual recognition, thus motivating this work.

Graph neural networks. Recently, several graph neural network models (GNN) have been proposed to exploit the structures underlying graphs to benefit a variety of applications. There are two main research lines of GNN methods: non-spectral methods and spectral methods. The spectral methods [4, 10, 26, 35] focus on learning graph representations in a spectral domain, in which the learned filters are based on Laplacian matrices. The non-spectral methods [21, 63] develop an aggregator to aggregate a local set of features. Note that, message passing [19] is a key mechanism that allows GCNs and other graph neural networks to capture
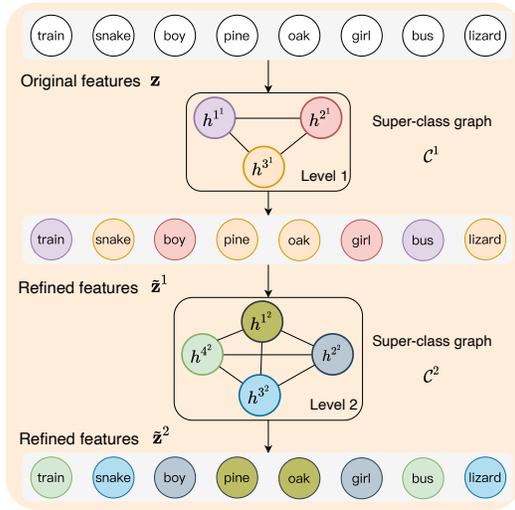
Figure 2. **Illustration of proposed SuperDisco** (a) and visualization of the similarity between the classes and discovered super-class at different levels (b), (c). In (a), we show two levels of super-class graphs $\mathcal{C}^1$ and $\mathcal{C}^2$. The colour in each graph represents the discovered super-class. SuperDisco discovers the potential super-class at different levels hidden in each category from (b) and (c). $\mathcal{C}^1$ roughly categorizes the original classes into three relatively balanced super-classes, and $\mathcal{C}^2$ then finely categorize them into four more balanced super-classes.

complex relationships and dependencies between nodes in a graph, and is a major reason why they have been successful in a variety of tasks involving graph-structured data. Our method belongs to the non-spectral methods, which leverage a GNN as the base architecture to discover the super-class representation. Our proposed super-class graph would refine and rectify the original imbalanced feature to a relatively balanced feature space, which has not been explored for long-tail recognition either.

Meta-learning for the long-tail. Meta-learning or learning to learn [3, 53, 58, 80, 81], is a learning paradigm where a model is trained on the distribution of tasks so as to enable rapid learning on new tasks. Ren *et al.* [51] first proposed meta-learning for the long-tailed problem by reweighting training examples. Shu *et al.* [54] proposed Meta-weight-Net to adaptively extract sample weights to guarantee robust recognition in the presence of training data bias. Li *et al.* [39] introduced meta-semantic augmentation for long-tailed recognition, which produces diversified augmented samples by translating features along many semantically meaningful directions by meta-learning. Our uniqueness is that our model aims to discover an improved super-class representation by meta-learning, which enables the original feature representation to adjust its corresponding higher-level super-class space.

3. Learning to discover the super-class graph

In this section, we discuss how to learn to discover the super-class graph from the training samples and then expand on how to leverage such a graph to benefit the unbalanced data by refining the feature representations of samples. The overall illustration of SuperDisco and a visualization of the discovered super-class hierarchy are shown in Figure 2.

Preliminary. For long-tailed visual recognition, the goal is to learn an image classification model from an imbalanced training set and to evaluate the model on a balanced test set. We first define the notation for long-tailed recognition used throughout our paper. We define a training input $x_k \in \mathbb{R}, i \in \{1, \cdots, n\}$, where n is the number of training samples in the dataset. The corresponding labels are $y_k \in 1, 2, \cdots, C$, where C is the number of classes. Let n_j denote the number of training samples for the class j. Here, we assume that $n_i \geq n_j$ when $i < j$ shows the long-tailed problem simply. In this work, we typically consider a deep network model with three main components: a feature extractor $f(\cdot)$, a proposed graph model $g(\cdot)$ and a classifier $h(\cdot)$. The feature extractor $f(\cdot)$ first extracts an image representation as $\mathbf{z}=f(x;\theta)$, which is then fed into the proposed graph model to refine a new representation as $g(\mathbf{z};\phi)=\tilde{\mathbf{z}}$. The final class prediction $\tilde{y}$ is given by a classifier function $h(\cdot)$, *i.e.*, $\tilde{y}= \arg\max h(\tilde{\mathbf{z}};\psi)$. Before detailing our approach, we add Table 1 to detail the meaning of each symbol for easy lookup.

SuperDisco. We construct the super-class graph to organize and distill knowledge from the training process. The vertices represent different types of super-classes (*e.g.*, the common contour between *birds* and *airplanes*) and the edges are automatically constructed to reflect the relationship between different super-classes. Our super-class graph contains multiple levels, which is closer to the relationship between various objects in the real world. Before detailing the structure, we first explicate why the multi-level super-class graphs are preferred over a flat super-class graph: a single level of

super-class groups is likely insufficient to model complex task relationships in real-world applications; for example, the similarities among different bird species are high, but there are also similarities between birds and mammals, *e.g.*, they are both animals.

We assume the vertex representation g as $\mathbf{h}^g \in \mathbb{R}^d$, and define the super-class graph as $\mathcal{C}^l=(\mathbf{H}_\mathcal{C}^l, \mathbf{A}_\mathcal{C}^l)$, where $\mathbf{H}_\mathcal{C}^l=\{\mathbf{h}^{i^l}|\forall i^l \in [1, C^l]\} \in \mathbb{R}^{C^l \times d}$ is the vertex feature matrix of the l-th super-class level and $\mathbf{A}_\mathcal{C}^l=\{A_\mathcal{C}^l(\mathbf{h}^{i^l}, \mathbf{h}^{j^l})|\forall i^l, j^l \in [1, C^l]\} \in \mathbb{R}^{C^l \times C^l}$ is the vertex adjacency matrix in the l-th super-class level, C^l denotes the number of vertices in the l-th super-class level. Our vertex representation $\mathbf{H}_\mathcal{C}^l$ of the super-class graph is defined to get parameterized and learned during training. The initial vertex representations of each super-class level are randomly initialized, which encourages diversity of the discovered super-classes.

Next, we introduce how to compute the edge weight $A_\mathcal{C}^l$ in the super-class graph. The edge weight $A_\mathcal{C}^l(\mathbf{h}^{i^l}, \mathbf{h}^{j^l})$ between a pair of vertices i and j is gauged by the similarity between them. Formally:

$$A_\mathcal{C}^l(\mathbf{h}^{i^l}, \mathbf{h}^{j^l})=\sigma(\mathbf{W}_c^l(|\mathbf{h}^{i^l} - \mathbf{h}^{j^l}|/\gamma_c^l) + \mathbf{b}_c^l), \qquad (1)$$

where $\mathbf{W}_c^l$ and $\mathbf{b}_c^l$ indicate learnable parameters of the l-th super-class level, γ_c^l of l-th super-class level is a scalar and σ indicates the Sigmoid function, which normalizes the weight between 0 and 1. To adjust the representation of training samples by the involvement of super-classes, we first query the training samples in the super-class graph to obtain the relevant super-class. In light of this, we construct a new graph $\mathcal{R}$, which adds the original sample feature $\mathbf{z}$ to the super-class graph. We define $\mathbf{z}^l$ as the refined feature after the l-th super-class graph. Here we define graph $\mathcal{R}^l=(\mathbf{H}_\mathcal{R}^l, \mathbf{A}_\mathcal{R}^l)$, where $\mathbf{H}_\mathcal{R}^l=\{[\mathbf{z}^l, \mathbf{h}^{i^l}]|\forall i^l \in [1, C^l]\} \in \mathbb{R}^{(C^l+1) \times d}$ denotes the vertex feature matrix of the l-th super-class level, and $\mathbf{A}_\mathcal{R}^l=\{[A_\mathcal{R}^l(\mathbf{h}^{i^l}, \mathbf{z}), A_\mathcal{R}^l(\mathbf{h}^{i^l}, \mathbf{h}^{j^l})]|\forall i^l, j^l \in [1, C^l]\} \in \mathbb{R}^{C^{l+1} \times C^{l+1}}$ denotes the vertex adjacency matrix in the l-th super-class level. The link between $\mathbf{z}^l$ and vertex $\mathbf{h}^i$ in the hierarchical graph is constructed by their similarity. In particular, analogous to the definition of weight in the super-class graph in Eq. (1), the weight $A_\mathcal{R}^l(\mathbf{h}^{i^l}, \mathbf{z}^l)$ is constructed as:

$$A_\mathcal{R}^l(\mathbf{h}^{i^l}, \mathbf{z}^l)=\sigma(\mathbf{W}_r^l(|\mathbf{h}^{i^l} - \mathbf{z}^l|/\gamma_r^l) + \mathbf{b}_r^l), \qquad (2)$$

where $\mathbf{W}_r^l$ and $\mathbf{b}_c^r$ indicate learnable parameters of the l-th super-class level, γ_r^l of the l-th super-class level is a scalar.

After constructing the new graph $\mathcal{R}$, we propagate the most relevant super-class by message passing [19] from the discovered super-classes $\mathcal{C}$ to the features $\mathbf{z}^l$ by introducing a Graph Neural Network (GNN). The message passing operation over the graph is formulated as:

$$\mathbf{H}_\mathcal{R}^{(m+1)}=\text{MP}(\mathbf{A}_\mathcal{R}^l, \mathbf{H}_\mathcal{R}^{(m)}; \mathbf{W}^{(m)}), \qquad (3)$$

where $\text{MP}(\cdot)$ is the message passing function, $\mathbf{H}^{(m)}$ is the vertex embedding after m layers of GNN and $\mathbf{W}^{(m)}$ is a

Notation	Description
$\mathbf{h}$	Vertex representation
$\mathbf{z}$	Original sample feature
$\mathcal{C}$	Super-class graph
$\mathbf{H}_\mathcal{C}$	Vertex feature matrix of $\mathcal{C}^l$
$\mathbf{A}_\mathcal{C}$	Vertex adjacency matrix of $\mathcal{C}^l$
$\mathcal{R}$	Graph which adds $\mathbf{z}$ to graph $\mathcal{C}^l$
$\mathcal{P}$	Prototype graph
$\mathbf{C}_\mathcal{P}$	Vertex feature matrix of $\mathcal{P}$
$\mathbf{A}_\mathcal{P}$	Vertex adjacency matrix of $\mathcal{P}$
$\mathcal{S}$	Super graph which connecting $\mathcal{P}$ and $\mathcal{C}$
$\mathbf{A}$	Vertex feature matrix of $\mathcal{S}$
$\mathbf{M}$	Vertex adjacency matrix of $\mathcal{S}$

Table 1. Summary of the core notation used for SuperDisco.

learnable weight matrix of layer m. After stacking M GNN layers, we get the information-propagated feature representation $\tilde{\mathbf{z}}^L$ for each level of the super-class graph $\mathcal{C}$. Once we obtain the refined representation $\tilde{\mathbf{z}}^L$ for a training sample by the super-class graph, we feed them into the classifier to make the predictions and compute the corresponding loss, *i.e.*, Cross-entropy loss for optimization. Using gradient descent, we then update the super-class graph $\mathcal{C}$. To be able to discover a more accurate super-class graph in the face of severe imbalance problems, we propose meta-learning super-class graph discovery in the next section.

4. Meta-learning super-class graph discovery

To explore and exploit a more accurate and richer super-class graph, we propose the Meta-SuperDisco to discover the super-class graph using meta-learning, making the model more robust. In the traditional meta-learning setting [17, 50], it includes meta-training tasks and meta-test task. Each task contains a support set $\mathcal{S}$ and a query set $\mathcal{Q}$. Each task is first trained by $\mathcal{S}$ to get the task-specific learner and $\mathcal{Q}$ optimizes this learner to update the meta-learner. For long-tailed recognition with meta-learning, previous works [51, 54] randomly sample a small amount of balanced data denoted as $\mathcal{M}$. The imbalanced data and the small balanced data can be seen as $\mathcal{S}$ and $\mathcal{Q}$ in the training phase. The goal of meta-learning for long-tailed recognition is to use a small set of balanced data to optimize the model obtained from unbalanced data. We follow [51, 54] by randomly selecting the same number of samples (*e.g.*, ten) as $\mathcal{M}$ per class from the training set.

Meta-SuperDisco. To meta-learn the super-class graph, we construct a prototype graph $\mathcal{P}$ from $\mathcal{M}$, since $\mathcal{M}$ is a balanced dataset. The prototype graph extracts the sample-level relation information, which captures the underlying relationship behind samples and alleviates the potential effects of abnormal samples. For the prototype graph, we need to compute the prototype of each category [55], which is defined as: $\mathbf{c}^k=\frac{1}{N^k} \sum_{i=1}^{N^k} \mathbf{z}_j$, where N^k denotes the number of samples in class k, $\mathbf{z}_j$ is the feature from $\mathcal{M}$.

After calculating all prototype representations $\{\mathbf{c}^k|\forall k \in [1, K]\}$, which serve as the vertices in the prototype graph $\mathcal{P}_i$, we further need to define the edges and the corresponding

edge weights. The edge weight $A_\mathcal{P}(\mathbf{c}^i, \mathbf{c}^j)$ between two prototypes $\mathbf{c}^i$ and $\mathbf{c}^j$ is gauged by the similarity between them. The edge weight is calculated as follows:

$$A_\mathcal{P}(\mathbf{c}^i, \mathbf{c}^j) = \sigma(\mathbf{W}_p(|\mathbf{c}^i - \mathbf{c}^j|/\gamma_p) + \mathbf{b}_p), \quad (4)$$

where $\mathbf{W}_p$ and $\mathbf{b}_p$ are the learnable parameters, γ_p is a scalar. Thus, we denote the prototype graph as $\mathcal{P}=(\mathbf{C}_\mathcal{P}, \mathbf{A}_\mathcal{P})$, where $\mathbf{C}_{\mathcal{P}_i}=\{\mathbf{c}^i|\forall i \in [1, K]\} \in \mathbb{R}^{K \times d}$ represent a set of vertices, with each one corresponding to the prototype from a class, while $\mathbf{A}_\mathcal{P}=\{|A_\mathcal{P}(\mathbf{c}^i, \mathbf{c}^j)|\forall i, j \in [1, K]\} \in \mathbb{R}^{K \times K}$ gives the adjacency matrix, which indicates the proximity between prototypes. We then use the prototype graph to guide the learning of the meta super-class graph. We construct a super graph $\mathcal{S}$ by connecting prototype graph $\mathcal{P}$ to super-class graph $\mathcal{C}$. In the super graph $\mathcal{S}$, the vertices are $\mathbf{M}=(\mathbf{C}_{\mathcal{P}^l}^l; \mathbf{H}_{\mathcal{C}^l}^l)$, where $\mathcal{P}^l$ denotes the refined prototype graph vertex after the l-th level super-class graph. Then, we calculate the link weight $A_\mathcal{S}^l(c^i, \{\mathbf{h}^j\})$ of the super graph as:

$$A_\mathcal{S}^l(\mathbf{c}^i, \mathbf{h}^{j^l}) = \frac{\exp(-\|(\mathbf{c}^i - \mathbf{h}^{j^l})/\gamma_s^l\|_2^2/2)}{\sum_{j^{l'}=1}^J \exp(-\|(\mathbf{c}^i - \mathbf{h}^{j^{l'}})/\gamma_s^l\|_2^2/2)}, \quad (5)$$

where γ_s^l is a scaling factor. Note that, here we use softmax to ensure that the total weight of edges between the prototype graph $\mathcal{P}$ and the super-class graph $\mathcal{C}$ is equal to 1, giving the prototype graph a unique influence on the expression of each super-class. Thus, the adjacent matrix and feature matrix of the super graph $\mathcal{S}^l=(\mathbf{A}^l, \mathbf{M}^l)$ is defined as $\mathbf{A}^l=(\mathbf{A}_\mathcal{P}, \mathbf{A}_\mathcal{S}^l; \mathbf{A}_\mathcal{S}^{l^\mathsf{T}}, \mathbf{A}_\mathcal{C}^l)$ and $\mathbf{M}^l=(\mathbf{C}_{\mathcal{P}^l}^l; \mathbf{H}_{\mathcal{C}^l}^l)$. Once we constructed the super graph $\mathcal{S}$, we use message-passing again to propagate the most relevant knowledge from the prototype graph $\mathcal{P}$ to the super-class graph $\mathcal{C}$. Similar to eq. (3):

$$\mathbf{M}^{(m+1)} = \text{MP}(\mathbf{A}^l, \mathbf{M}^{(m)}; \mathbf{W}^{(m)}). \quad (6)$$

We leverage the graph $\mathcal{S}$ to refine the super-class graph. Finally, we feed the original feature $\mathbf{z}$ into the super-class graph to get the information-propagated feature representation $\tilde{\mathbf{z}}^L$, which refines the original feature by its corresponding discovered super-classes. We provide the complete SuperDisco and Meta-SuperDisco algorithm specifications in the supplemental material.

5. Experiments

Datasets. We apply our method to four commonly used long-tail recognition benchmarks. Sample images and the number of categories for all datasets are provided in the supplement material. ***CIFAR-100-LT*** reduces the number of training samples per class according to an exponential function $n=n_i\mu^i$, where i is the class index, n_i is the original number of training samples, and $\mu \in (0, 1)$. The imbalance factor of a dataset is defined as the number of training samples in the most populated class divided by the minority class. We consider imbalance factors $\{10, 50, 100\}$. ***ImageNet-LT*** [44] is a subset of ImageNet [12] consisting of

	Imbalance ratio				
	10	20	50	100	200
Baseline	60.3	57.3	47.5	44.9	39.3
SuperDisco	65.9	60.7	57.2	50.9	45.2
Meta-SuperDisco	68.5	63.1	58.3	53.8	47.5

Table 2. **Benefit of SuperDisco and Meta-SuperDisco.** SuperDisco achieves better performance compared to a baseline fine-tuning on all imbalance factors, while Meta-SuperDisco is even better for long-tailed recognition.

115.8K images from 1000 categories, with maximally 1,280 images per class and minimally 5 images per class, and a balanced test set. ***Places-LT*** [44] has an imbalanced training set with 62,500 images for 365 classes from Places [84]. It contains images from 365 classes and the number of images per class ranges from 4980 to 5. The test sets are balanced and contain 100 images per class. ***iNaturalist*** [61] is a real-world long-tailed dataset with 675,170 training images for 5,089 classes, where the top 1% most populated classes contain more than 16% of the training images. Additionally, there is also a severe imbalance among the super-classes of iNaturalist. The 13 ground truth super-classes images range from 158,407 to 308.

Implementation details. We follow [32] by first training a feature extractor with instance-balanced sampling, and then training our graph model and classifier based on the trained features. For CIFAR-100-LT, we follow [54] and use a ResNet-32 backbone. For ImageNet-LT, we use ResNeXt-50 [23] as our backbone, following [32]. For Places-LT, we report results with ResNet-152 following [44]. For iNaturalist, we use a ResNet-50 backbone. We train each dataset for 200 epochs with batch size 512. We use random left-right flipping and cropping as our training augmentation. For all experiments, we use an SGD optimizer with a momentum of 0.9 and a batch size of 512. We randomly selected 10 images per class from the training set for all datasets as $\mathcal{M}$. Code available at: https://github.com/Yingjun-Du/SuperDisco.

Benefit of SuperDisco and Meta-SuperDisco. To show the benefit of SuperDisco, we compare it with a fine-tuning baseline, which retrains the classifier only. Table 2 shows SuperDisco improves over fine-tuning on CIFAR-100-LT, and the results for the other long-tailed datasets are provided in the supplemental materials Table 1. In the most challenging setting with the largest imbalance factor of 200, our SuperDisco delivers 45.2%, surpassing the baseline by 5.9%. We attribute improvement to our model's ability to refine original features, allowing the discovered super-class graph to guide the tail features away from the dominant role of head features, thus leading to improvements over the original features. We also investigate the benefit of meta-learning with Meta-SuperDisco. The Meta-SuperDisco consistently surpasses the SuperDisco for all imbalance factors. The con-

	Imbalance ratio				
	10	20	50	100	200
Baseline	60.3	57.3	47.5	44.9	39.3
(20)	61.2	60.1	49.9	47.3	41.9
(2, 4, 8)	65.3	62.7	53.1	49.8	43.2
(4, 8, 16)	**69.1**	**64.2**	55.2	52.3	45.9
(4, 8, 16, 32)	68.5	63.1	58.3	**53.8**	**47.5**
(4, 8, 16, 32, 64)	66.9	62.7	**58.9**	52.9	46.3
Oracle super-classes	66.9	63.2	54.7	51.4	43.2

Table 3. **Effect of number of super-class levels on CIFAR-100-LT**. Compared to a baseline [32] and an oracle setting, Meta-SuperDisco provides higher performance gains with more complex hierarchies.

	Many	Medium	Few	All
Baseline	65.0	66.3	65.5	65.9
(13)	71.8	70.2	66.1	70.8
(2, 4, 8)	70.5	69.3	65.9	69.4
(4, 8, 16)	72.2	70.9	66.4	70.3
(4, 8, 16, 32)	**73.6**	70.2	67.3	70.9
(4, 8, 16, 32, 64)	73.4	**72.9**	**68.3**	**72.3**
(4, 8, 16, 32, 64, 128)	72.1	71.3	66.2	70.9
Oracle super-classes	70.7	70.5	65.9	70.2

Table 4. **Effect of number of super-class levels on iNaturalist**. Meta-SuperDisco achieves consistent performance gains with more complex hierarchies.

sistent improvements confirm that Meta-SuperDisco learns even more robust super-class graphs, leading to a discriminative representation of the tail data.

Effect of the number of super-class levels. A significant challenge with any structure-aware learning algorithm is determining the appropriate complexity for the knowledge structure. So, we further analyze the effect of the super-class hierarchies, including the level (number of depths L) or the number of super-classes in each level. The results are shown in Table 3 and Table 4. The super-class number from the bottom layer to the top layer is saved in a tuple. For example, (2, 4, 8) represents three depth, with two super-classes in the top layer. The baseline is Decouple-LWS [32], which only inputs the original feature to learn a new classifier. The oracle super-classes are first trained on two long-tailed datasets using the ground truth super-class labels for super-class classifications. Once the training is completed, each oracle super-class is obtained by averaging the samples of each super-class. We constructed a one-layer super-class graph using these super-classes, where the vertices of the graph are for each super-class, and the edges of the graph are computed according to Eq. (1). Then, we use the message passing by Eq. (3) to refine the original features and input them into the classifier to get the final predictions. From Table 3, we observe that using oracle super-classes achieves better performance compared to the learned super-class (20) since it uses the ground truth super-classes as supervision. We also conclude that too few levels may not be enough

(a) SuperDisco   (b) Meta-SuperDisco
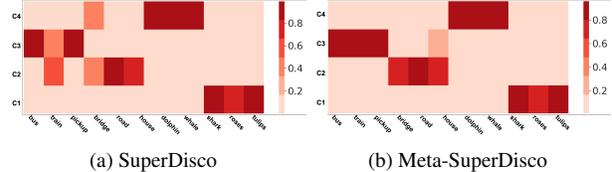
Figure 3. **Similarity between discovered super-classes and classes.** SuperDisco discovers super-classes hidden in each class, while Meta-SuperDisco discovers more accurate super-classes.

to learn the precise super-classes (*e.g.*, tuple (20) or (2, 4, 8)). In this dataset, increasing levels (*e.g.*, tuple (4, 8, 16, 32)) achieves better performance on the smaller imbalance factor (*e.g.*, 10), and similar performance compared with (4, 8, 16). For the real-world long-tailed dataset iNaturalist [62] in Table 4, we find no significant improvement for the few-shot classes in the performance of the oracle super-class compared to the baseline, and the same is true for the performance of the discovered super-class structure (13). This is because the super-class of iNaturalist also have serious long-tailed problems, resulting in the refined features of tail classes remaining indistinguishable from the refined features of head classes. However, with a more complex graph structure (4, 8, 16, 32, 64), the few-shot performance improves by a good margin compared with the baseline, and even the oracle super-classes. We attribute this to our model's ability to explore relatively balanced super-class spaces, thus making the refined tail category features discriminative. By comparing Table 3 and Table 4, we conclude that deeper as well as wider graphs are needed to discover the super-classes in the case of severe class imbalance.

Visualization of SuperDisco. To understand the meaning of the discovered super-classes more clearly, we present a visualization in Figure. 3. We selected 12 different categories from the CIFAR-100 test dataset. We calculate the similarity of each of these 12 categories to the different vertices in the graph we explore. Here we show the similarity with the second layer of graph vertices (C1, C2, C3, C4). We can see different categories mainly activate different vertices, e.g., bus → C3 and road → C2. As shown in this heatmap, we find that C1 reflects the super-class of *flowers*, C2 reflects the super-class of *buildings*, C3 reflects the super-class of *vehicles*, C4 reflects the super-class of *fish*. Another observation is that the second-largest activated super-class is also meaningful, promoting knowledge transfer between super-classes. For example, *road* and *bridge* are related to the C3 super-class, since some vehicles may be on the road and bridge. This visualization reflects that we can use graph models to discover the super-classes and the relationships between each super-class. We also visualize the discovered meta-learning super-classes in Figure 3 (b). The discovered super-classes are even more accurate, e.g., *roses* have high similarity to C2, which mainly reflects the *buildings* super-class, while it has high similarity to C1, which is the *flowers* super-class. This once again validates the benefit of Meta-SuperDisco.

(a) Original features

(c) Refined features by SuperDisco

(b) Training distribution

(d) Refined features by Meta-SuperDisco
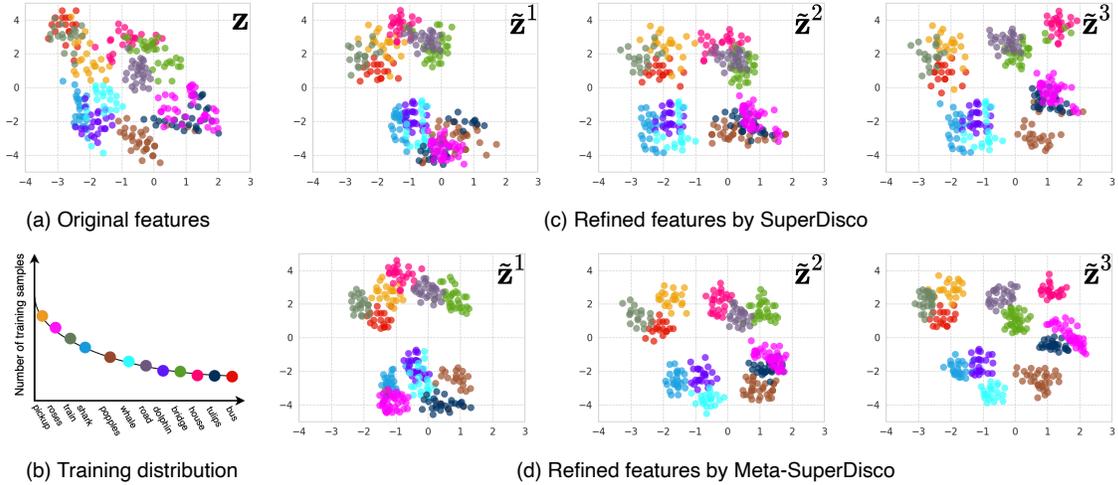
Figure 4. **Visualization of refined features** on CIFAR-100-LT, with the original features (a) and their corresponding training distribution (b). Colours indicate categories. SuperDisco (c) guides the original features on being clustered into the corresponding super-class space at different levels, while Meta-SuperDisco (d) obtains even more discriminative intra-class features.
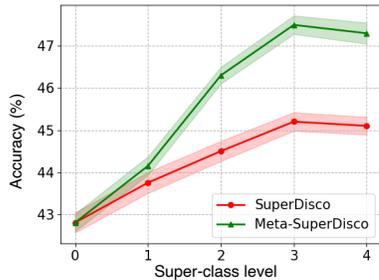
Figure 5. **Effect of refined features.** Accuracy increases along with the increased super-class levels, revealing that more accurate and richer super-classes facilitate better long-tailed recognition.

	Many	Medium	Few	All
Baseline	58.4	49.3	34.8	52.7
Multi-layer perceptron	63.5	51.8	35.9	55.0
Graph convolution network	66.1	53.3	37.1	57.1

Table 5. **Analysis of super-class mechanism** on ImageNet-LT. The super-class mechanism contributes most, the graph convolution network improves results further.

Furthermore, in the (c) learned the hierarchical concept of each class, we can see that bus and bridge have the same concept C2 in the last concept level, which may be due to the possible presence of cars on the bridge.

Visualization of refined features. To understand the empirical benefit of SuperDisco, we visualize in Figure 4 the original features and refined features with super-class graphs of the different levels using t-SNE [60]. We choose the vertices numbers as (2, 4, 6), meaning the super-class graph has three different levels, each with a different number of vertices. The original features of the category with a small sample size will overlap with the (original) features of the category with a large sample size. Super-class graphs discovered by our model project the original features into a high-level super-class space, where the imbalance is relatively small. Hence, its corresponding subcategory can be predicted more accurately. It is worth noting that when comparing the two different super-class graphs on top and below, the features obtained by Meta-SuperDisco are even more distinctive and distant from each other. To better measure the goodness of the refined features obtained at different levels,

we show in Figure 5 the prediction accuracy using different refined features. We find that the accuracy increases along with the increased super-class levels, which shows that using more accurate and richer super-classes facilitates better performance. This again demonstrates that Meta-SuperDisco is most suitable for long-tailed visual recognition.

Analysis of super-class mechanism. To demonstrate that the improved performance of our SuperDisco cannot solely be attributed to the graph convolutional network module, we conducted an experiment where we replaced it with a multi-layer perceptron to obtain the representation per sample. In Table 5, the performance gains of our method are primarily due to the super-classes rather than the graph convolution network. The results suggest that incorporating the super-classes mechanism plays a crucial role in improving the performance of long-tailed problems. Furthermore, the results improve further when we replace the multi-layer perceptron with our graph convolution network module.

Comparison with the state-of-the-art. We evaluate our method on the four long-tailed datasets under different imbalance factors in Table 6 and 7. Our model achieves state-of-the-art performance on the tail data of all datasets. For ImageNet-LT, our model achieves state-of-the-art performance on both few-shot and all data. In the most challenging Places-LT, our model delivers 40.3% on all classes, surpassing the second-best Parisot *et al.* [47] by 1.1%. On the real-

	Venue	ImageNet-LT				Places-LT				iNaturalist			
		Many	Medium	Few	All	Many	Medium	Few	All	Many	Medium	Few	All
Kang *et al.* [32]	ICLR 19	60.2	47.2	30.3	49.9	40.6	39.1	28.6	37.6	65.0	66.3	65.5	65.9
Kang *et al.* [31]	ICLR 21	61.8	49.4	30.9	51.5	-	-	-	-	-	-	-	68.6
He *et al.* [25]	ICCV 21	64.1	50.4	31.5	53.1	-	-	-	-	70.6	70.1	67.6	69.1
Li *et al.* [40]	CVPR 21	**66.8**	51.1	35.4	56.0	-	-	-	-	-	-	-	69.3
Samuel *et al.* [52]	ICCV 21	64.0	49.8	33.1	53.5	-	-	-	-	-	-	-	69.7
Alshammari *et al.* [1]	CVPR 22	62.5	50.4	41.5	53.9	-	-	-	-	71.2	70.4	69.7	70.2
Zhang *et al.* [75]	CVPR 21	61.3	52.2	31.4	52.9	40.4	*42.4*	30.1	39.3	69.0	71.1	*70.2*	70.6
Parisot *et al.* [47]	CVPR 22	63.2	52.1	*36.9*	54.1	39.7	41.0	*34.9*	*39.2*	-	-	-	-
Park *et al.* [48]	CVPR 22	*66.4*	**53.9**	35.6	*56.2*	-	-	-	-	**73.1**	*72.6*	68.7	*72.8*
This paper		66.1	*53.3*	**37.1**	**57.1**	**45.3**	**42.8**	**35.3**	**40.3**	*72.3*	**72.9**	**71.3**	**73.6**

Table 6. **Comparison with the state-of-the-art on ImageNet-LT, Places-LT and iNaturalist.** Best and second best results are highlighted in **bold** and ***italic bold***. Our Meta-SuperDisco achieves either better or comparable performance than state-of-the-art methods under the tail and all data for long-tailed visual recognition.

		Imbalance ratio		
	Venue	10	50	100
Park *et al.* [49]	ICCV 21	59.5	47.4	42.0
Li *et al.* [40]	CVPR 21	62.3	50.5	46.0
Zhong *et al.* [82]	CVPR 21	62.5	51.5	46.8
Samuel *et al.* [52]	ICCV 21	63.4	57.6	47.3
Wang *et al.* [66]	ICLR 21	61.8	51.7	48.0
Zhu *et al.* [87]	CVPR 22	64.9	56.6	51.9
Cui *et al.* [7]	ICCV 21	64.2	56.0	52.0
Alshammari *et al.* [1]	CVPR 22	*68.8*	*57.7*	*53.3*
This paper		**69.3**	**58.3**	**53.8**

Table 7. **Comparison with the state-of-the-art on CIFAR-100-LT.** Our model achieves best performance.

Figure 6. **Limitation.** Accuracy (%) vs. speed (ms) comparison with different methods on balanced CIFAR-100. SuperDisco has little impact on the performance of balanced datasets at the expense of increased inference time.

world long-tailed dataset iNaturalist, our model achieves the three best performances under four different shots. On the long-tailed synthetic dataset CIFAR-100-LT, our model achieves the best performance under each imbalance factor. The consistent improvements on all benchmarks under various configurations confirm that our Meta-SuperDisco is effective for long-tailed visual recognition.

Limitations. We show that SuperDisco and Meta-SuperDisco achieve good performance on tail data while being less successful on the head data. Based on this result, we also perform an experiment on *balanced* CIFAR-100 in Figure 6. With SuperDisco and Meta-SuperDisco, there is only a slight change in performance at the expense of an increased inference time. This reveals that our SuperDisco does not change the original features much through message passing on a balanced dataset. This may be because the obtained super-classes are still the original class itself. In addition, as the computation of graphs involves many matrix operations, our model also requires a relatively long computational speed. Due to introducing a prototype graph and more data, Meta-SuperDisco takes longer to compute. In addition, the training time of SuperDisco and its meta variants is also 1.5 times higher than the baseline. Future work could investigate how to use the discovered super-class graph in balanced datasets and how to reduce the computation time.

6. Conclusions

This paper proposes learning to discover a super-class graph for long-tailed visual recognition. The proposed super-class graph could rectify and refine the original features by message passing, which results in attending to the most relevant entities based on their semantic similarity between concepts for more accurate predictions. To obtain a more informative super-class graph and more balanced image representations, we further propose to meta-learn the super-class graph based on the prototype graph from a small amount of imbalanced data. We conduct thorough ablation studies to demonstrate the effectiveness of the proposed SuperDisco and Meta-SuperDisco algorithms. The state-of-the-art performance on the long-tailed version of four datasets further substantiates the benefit of our proposal.

References

[1] Shaden Alshammari, Yu-Xiong Wang, Deva Ramanan, and Shu Kong. Long-tailed recognition via weight balancing. In *CVPR*, pages 6897–6907, 2022. 8

[2] Chris Anderson. *The long tail: How endless choice is creating unlimited demand*. Random House, 2007. 1

[3] Yoshua Bengio, Samy Bengio, and Jocelyn Cloutier. Learning a synaptic learning rule. In *IJCNN*, 1991. 3

[4] Joan Bruna, Wojciech Zaremba, Arthur Szlam, and Yann LeCun. Spectral networks and locally connected networks on graphs. *arXiv preprint arXiv:1312.6203*, 2013. 2

[5] Zhi Chen, Yijie Bei, and Cynthia Rudin. Concept whitening for interpretable image recognition. In *Nature Machine Intelligence*, 2020. 1

[6] Peng Chu, Xiao Bian, Shaopeng Liu, and Haibin Ling. Feature space augmentation for long-tailed data. In *ECCV*, 2020. 2

[7] Jiequan Cui, Zhisheng Zhong, Shu Liu, Bei Yu, and Jiaya Jia. Parametric contrastive learning. In *ICCV*, 2021. 8

[8] Yin Cui, Menglin Jia, Tsung-Yi Lin, Yang Song, and Serge Belongie. Class-balanced loss based on effective number of samples. In *CVPR*, 2019. 2

[9] Yin Cui, Yang Song, Chen Sun, Andrew Howard, and Serge Belongie. Large scale fine-grained categorization and domain-specific transfer learning. In *CVPR*, 2018. 2

[10] Michaël Defferrard, Xavier Bresson, and Pierre Vandergheynst. Convolutional neural networks on graphs with fast localized spectral filtering. In *NeurIPS*, 2016. 2

[11] Hojat Asgarian Dehkordi, Ali Soltani Nezhad, Hossein Kashiani, Shahriar Baradaran Shokouhi, and Ahmad Ayatollahi. Multi-expert human action recognition with hierarchical super-class learning. *Knowledge-Based Systems*, page 109091, 2022. 2

[12] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, pages 248–255, 2009. 5

[13] Qi Dong, Shaogang Gong, and Xiatian Zhu. Class rectification hard mining for imbalanced deep learning. In *ICCV*, 2017. 2

[14] Charles Elkan. The foundations of cost-sensitive learning. In *IJCAI*, 2001. 2

[15] Andrew Estabrooks, Taeho Jo, and Nathalie Japkowicz. A multiple resampling method for learning from imbalanced data sets. *Computational Intelligence*, 20(1):18–36, 2004. 1

[16] Adam R Ferguson, Jessica L Nielson, Melissa H Cragin, Anita E Bandrowski, and Maryann E Martone. Big data from small data: data-sharing in the'long tail'of neuroscience. *Nature Neuroscience*, 17(11):1442–1447, 2014. 1

[17] Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *ICLR*, pages 1126–1135, 2017. 4

[18] Camille Garcin, Alexis Joly, Pierre Bonnet, Jean-Christophe Lombardo, Antoine Affouard, Mathias Chouet, Maximilien Servajean, Joseph Salmon, and Titouan Lorieul. Pl@ntnet-300k: a plant image dataset with high label ambiguity and a long-tailed distribution. In *NeurIPS 2021-35th Conference on Neural Information Processing Systems*, 2021. 1

[19] Justin Gilmer, Samuel S Schoenholz, Patrick F Riley, Oriol Vinyals, and George E Dahl. Neural message passing for quantum chemistry. In *ICLR*, pages 1263–1272, 2017. 2, 4

[20] Hao Guo and Song Wang. Long-tailed multi-label visual recognition by collaborative training on uniform and re-balanced samplings. In *CVPR*, pages 15089–15098, 2021. 2

[21] Will Hamilton, Zhitao Ying, and Jure Leskovec. Inductive representation learning on large graphs. In *NeurIPS*, 2017. 2

[22] 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, 2005. 1, 2

[23] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 5

[24] Ruifei He, Jihan Yang, and Xiaojun Qi. Re-distributing biased pseudo labels for semi-supervised semantic segmentation: A baseline investigation. In *ICCV*, 2021. 2

[25] Yin-Yin He, Jianxin Wu, and Xiu-Shen Wei. Distilling virtual examples for long-tailed recognition. In *ICCV*, 2021. 8

[26] Mikael Henaff, Joan Bruna, and Yann LeCun. Deep convolutional networks on graph-structured data. *arXiv preprint arXiv:1506.05163*, 2015. 2

[27] Youngkyu Hong, Seungju Han, Kwanghee Choi, Seokjun Seo, Beomsu Kim, and Buru Chang. Disentangling label distribution for long-tailed visual recognition. In *CVPR*, 2021. 2

[28] Xinting Hu, Yi Jiang, Kaihua Tang, Jingyuan Chen, Chunyan Miao, and Hanwang Zhang. Learning to segment the tail. In *CVPR*, 2020. 2

[29] Chen Huang, Yining Li, Chen Change Loy, and Xiaoou Tang. Learning deep representation for imbalanced classification. In *CVPR*, 2016. 1

[30] 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 *CVPR*, pages 7610–7619, 2020. 2

[31] Bingyi Kang, Yu Li, Sa Xie, Zehuan Yuan, and Jiashi Feng. Exploring balanced feature spaces for representation learning. In *ICLR*, 2021. 1, 2, 8

[32] Bingyi Kang, Saining Xie, Marcus Rohrbach, Zhicheng Yan, Albert Gordo, Jiashi Feng, and Yannis Kalantidis. Decoupling representation and classifier for long-tailed recognition. *arXiv preprint arXiv:1910.09217*, 2019. 1, 2, 5, 6, 8

[33] Shyamgopal Karthik, Jérome Revaud, and Chidlovskii Boris. Learning from long-tailed data with noisy labels. *arXiv:2108.11096*, 2021. 2

[34] Jaehyung Kim, Jongheon Jeong, and Jinwoo Shin. M2m: Imbalanced classification via major-to-minor translation. In *CVPR*, 2020. 1

[35] Thomas N Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*, 2016. 2

[36] Pang Wei Koh, Thao Nguyen, Yew Siang Tang, Stephen Mussmann, Emma Pierson, Been Kim, and Percy Liang. Concept bottleneck models. In *ICLR*, 2020. 1

[37] Bo Li, Yongqiang Yao, Jingru Tan, Gang Zhang, Fengwei Yu, Jianwei Lu, and Ye Luo. Equalized focal loss for dense long-tailed object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6990–6999, 2022. 2

[38] Kaidong Li, Nina Y Wang, Yiju Yang, and Guanghui Wang. Sgnet: A super-class guided network for image classification and object detection. In *Conference on Robots and Vision*, pages 127–134, 2021. 2

[39] Shuang Li, Kaixiong Gong, Chi Harold Liu, Yulin Wang, Feng Qiao, and Xinjing Cheng. Metasaug: Meta semantic augmentation for long-tailed visual recognition. In *CVPR*, pages 5212–5221, 2021. 3

[40] Tianhao Li, Limin Wang, and Gangshan Wu. Self supervision to distillation for long-tailed visual recognition. In *CVPR*, pages 630–639, 2021. 2, 8

[41] Bo Liu, Haoxiang Li, Hao Kang, Gang Hua, and Nuno Vasconcelos. Gistnet: a geometric structure transfer network for long-tailed recognition. In *ICCV*, 2021. 2

[42] Jialun Liu, Yifan Sun, Chuchu Han, Zhaopeng Dou, and Wenhui Li. Deep representation learning on long-tailed data: A learnable embedding augmentation perspective. In *CVPR*, 2020. 2

[43] Xu-Ying Liu, Jianxin Wu, and Zhi-Hua Zhou. Exploratory undersampling for class-imbalance learning. *IEEE Transactions on Systems, Man, and Cybernetics*, 39(2):539–550, 2008. 1, 2

[44] Ziwei Liu, Zhongqi Miao, Xiaohang Zhan, Jiayun Wang, Boqing Gong, and Stella X Yu. Large-scale long-tailed recognition in an open world. In *CVPR*, pages 2537–2546, 2019. 1, 2, 5

[45] Aditya Krishna Menon, Sadeep Jayasumana, Ankit Singh Rawat, Himanshu Jain, Andreas Veit, and Sanjiv Kumar. Long-tail learning via logit adjustment. In *ICLR*, 2021. 2

[46] Wanli Ouyang, Xiaogang Wang, Cong Zhang, and Xiaokang Yang. Factors in finetuning deep model for object detection with long-tail distribution. In *CVPR*, pages 864–873, 2016. 2

[47] Sarah Parisot, Pedro M Esperança, Steven McDonagh, Tamas J Madarasz, Yongxin Yang, and Zhenguo Li. Long-tail recognition via compositional knowledge transfer. In *CVPR*, pages 6939–6948, 2022. 7, 8

[48] 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. 8

[49] Seulki Park, Jongin Lim, Younghan Jeon, and Jin Young Choi. Influence-balanced loss for imbalanced visual classification. In *ICCV*, pages 735–744, 2021. 8

[50] Sachin Ravi and Hugo Larochelle. Optimization as a model for few-shot learning. In *ICLR*, 2017. 4

[51] Mengye Ren, Wenyuan Zeng, Bin Yang, and Raquel Urtasun. Learning to reweight examples for robust deep learning. In *ICML*, 2018. 1, 3, 4

[52] Dvir Samuel and Gal Chechik. Distributional robustness loss for long-tail learning. In *ICCV*, 2021. 8

[53] Jürgen Schmidhuber. *Evolutionary principles in self-referential learning, or on learning how to learn: the meta-meta-... hook*. PhD thesis, Technische Universität München, 1987. 3

[54] Jun Shu, Qi Xie, Lixuan Yi, Qian Zhao, Sanping Zhou, Zongben Xu, and Deyu Meng. Meta-weight-net: Learning an explicit mapping for sample weighting. In *NeurIPS*, 2019. 1, 3, 4, 5

[55] Jake Snell, Kevin Swersky, and Richard Zemel. Prototypical networks for few-shot learning. In *NeurIPS*, 2017. 4

[56] 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. 2

[57] Kaihua Tang, Jianqiang Huang, and Hanwang Zhang. Long-tailed classification by keeping the good and removing the bad momentum causal effect. In *NeurIPS*, 2020. 2

[58] Sebastian Thrun and Lorien Pratt, editors. *Learning to Learn*. Kluwer Academic Publishers, USA, 1998. 3

[59] Junjiao Tian, Yen-Cheng Liu, Nathan Glaser, Yen-Chang Hsu, and Zsolt Kira. Posterior re-calibration for imbalanced datasets. In *NeurIPS*, 2020. 2

[60] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(11), 2008. 7

[61] Grant Van Horn, Oisin Mac Aodha, Yang Song, Yin Cui, Chen Sun, Alex Shepard, Hartwig Adam, Pietro Perona, and Serge Belongie. The inaturalist species classification and detection dataset. In *CVPR*, pages 8769–8778, 2018. 5

[62] Grant Van Horn and Pietro Perona. The devil is in the tails: Fine-grained classification in the wild. *arXiv preprint arXiv:1709.01450*, 2017. 6

[63] Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Lio, and Yoshua Bengio. Graph attention networks. *arXiv preprint arXiv:1710.10903*, 2017. 2

[64] Jianfeng Wang, Thomas Lukasiewicz, Xiaolin Hu, Jianfei Cai, and Zhenghua Xu. Rsg: A simple but effective module for learning imbalanced datasets. In *CVPR*, pages 3784–3793, 2021. 2

[65] Tao Wang, Yu Li, Bingyi Kang, Junnan Li, Junhao Liew, Sheng Tang, Steven Hoi, and Jiashi Feng. The devil is in classification: A simple framework for long-tail instance segmentation. In *ECCV*, 2020. 1

[66] Xudong Wang, Long Lian, Zhongqi Miao, Ziwei Liu, and Stella X Yu. Long-tailed recognition by routing diverse distribution-aware experts. In *ICLR*, 2021. 8

[67] Zhihao Wang, Jian Chen, and Steven CH Hoi. Deep learning for image super-resolution: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020. 2

[68] Chen Wei, Kihyuk Sohn, Clayton Mellina, Alan Yuille, and Fan Yang. Crest: A class-rebalancing self-training framework for imbalanced semi-supervised learning. In *CVPR*, 2021. 2

[69] Tong Wu, Ziwei Liu, Qingqiu Huang, Yu Wang, and Dahua Lin. Adversarial robustness under long-tailed distribution. In *CVPR*, pages 8659–8668, 2021. 2

[70] Tz-Ying Wu, Pedro Morgado, Pei Wang, Chih-Hui Ho, and Nuno Vasconcelos. Solving long-tailed recognition with deep

realistic taxonomic classifier. In *ECCV*, pages 171–189, 2020. 2

[71] Liuyu Xiang, Guiguang Ding, and Jungong Han. Learning from multiple experts: Self-paced knowledge distillation for long-tailed classification. In *ECCV*, pages 247–263, 2020. 2

[72] Yuzhe Yang and Zhi Xu. Rethinking the value of labels for improving class-imbalanced learning. In *NeurIPS*, 2020. 2

[73] Xi Yin, Xiang Yu, Kihyuk Sohn, Xiaoming Liu, and Manmohan Chandraker. Feature transfer learning for face recognition with under-represented data. In *CVPR*, pages 5704–5713, 2019. 2

[74] Cheng Zhang, Tai-Yu Pan, Yandong Li, Hexiang Hu, Dong Xuan, Soravit Changpinyo, Boqing Gong, and Wei-Lun Chao. Mosaicos: A simple and effective use of object-centric images for long-tailed object detection. In *ICCV*, 2021. 2

[75] Songyang Zhang, Zeming Li, Shipeng Yan, Xuming He, and Jian Sun. Distribution alignment: A unified framework for long-tail visual recognition. In *CVPR*, pages 2361–2370, 2021. 2, 8

[76] Xiao Zhang, Zhiyuan Fang, Yandong Wen, Zhifeng Li, and Yu Qiao. Range loss for deep face recognition with long-tailed training data. In *ICCV*, 2017. 1, 2

[77] Yifan Zhang, Bingyi Kang, Bryan Hooi, Shuicheng Yan, and Jiashi Feng. Deep long-tailed learning: A survey. *arXiv preprint arXiv:2110.04596*, 2021. 1

[78] Yifan Zhang, Peilin Zhao, Jiezhang Cao, Wenye Ma, Junzhou Huang, Qingyao Wu, and Mingkui Tan. Online adaptive asymmetric active learning for budgeted imbalanced data. In *KDD*, pages 2768–2777, 2018. 2

[79] Peilin Zhao, Yifan Zhang, Min Wu, Steven CH Hoi, Mingkui Tan, and Junzhou Huang. Adaptive cost-sensitive online classification. *IEEE Transactions on Knowledge and Data Engineering*, 31(2):214–228, 2018. 2

[80] Xiantong Zhen, Yingjun Du, Huan Xiong, Qiang Qiu, Cees Snoek, and Ling Shao. Learning to learn variational semantic memory. In *NeurIPS*, 2020. 3

[81] Xiantong Zhen, Haoliang Sun, Yingjun Du, Jun Xu, Yilong Yin, Ling Shao, and Cees Snoek. Learning to learn kernels with variational random features. In *ICML*, 2020. 3

[82] Zhisheng Zhong, Jiequan Cui, Shu Liu, and Jiaya Jia. Improving calibration for long-tailed recognition. In *CVPR*, 2021. 2, 8

[83] Boyan Zhou, Quan Cui, Xiu-Shen Wei, and Zhao-Min Chen. Bbn: Bilateral-branch network with cumulative learning for long-tailed visual recognition. In *CVPR*, pages 9719–9728, 2020. 2

[84] 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*, 2017. 5

[85] Yucan Zhou, Qinghua Hu, and Yu Wang. Deep super-class learning for long-tail distributed image classification. *Pattern Recognition*, 80:118–128, 2018. 2

[86] ZhiHua Zhou and XuYing Liu. Training cost-sensitive neural networks with methods addressing the class imbalance problem. *IEEE Transactions on Knowledge and Data Engineering*, 18(1):63–77, 2005. 2

[87] Jianggang Zhu, Zheng Wang, Jingjing Chen, Yi-Ping Phoebe Chen, and Yu-Gang Jiang. Balanced contrastive learning for long-tailed visual recognition. In *CVPR*, pages 6908–6917, 2022. 8