# Self Supervision to Distillation for Long-Tailed Visual Recognition

Tianhao Li Limin Wang✉ Gangshan Wu
State Key Laboratory for Novel Software Technology, Nanjing University, China

## Abstract

*Deep learning has achieved remarkable progress for visual recognition on large-scale balanced datasets but still performs poorly on real-world long-tailed data. Previous methods often adopt class re-balanced training strategies to effectively alleviate the imbalance issue, but might be a risk of over-fitting tail classes. The recent decoupling method overcomes over-fitting issues by using a multi-stage training scheme, yet, it is still incapable of capturing tail class information in the feature learning stage. In this paper, we show that soft label can serve as a powerful solution to incorporate label correlation into a multi-stage training scheme for long-tailed recognition. The intrinsic relation between classes embodied by soft labels turns out to be helpful for long-tailed recognition by transferring knowledge from head to tail classes.*

*Specifically, we propose a conceptually simple yet particularly effective multi-stage training scheme, termed as Self Supervised to Distillation (SSD). This scheme is composed of two parts. First, we introduce a self-distillation framework for long-tailed recognition, which can mine the label relation automatically. Second, we present a new distillation label generation module guided by self-supervision. The distilled labels integrate information from both label and data domains that can model long-tailed distribution effectively. We conduct extensive experiments and our method achieves the state-of-the-art results on three long-tailed recognition benchmarks: ImageNet-LT, CIFAR100-LT and iNaturalist 2018. Our SSD outperforms the strong LWS baseline by from* $2.7\%$ *to* $4.5\%$ *on various datasets.*

## 1. Introduction

Deep learning has achieved remarkable progress for visual recognition in both image and video domains by training powerful neural networks on large-scale balanced and curated datasets (e.g., ImageNet [8] and Kinetics [21]). Distinct from these artificially balanced datasets, real-world data always follows long-tailed distribution [30, 29], which makes collecting balanced datasets more challenging, especially for classes naturally with rare samples. However, learning directly from long-tailed data induces significant performance degeneration due to the highly imbalanced data distribution.

A common series of approaches to alleviate the deterioration caused by long-tailed training data is based on class re-balanced strategies [13, 22, 2, 10, 33], including re-sampling training data [13, 10, 6, 33] and designing cost-sensitive re-weighting loss functions [22, 39]. These methods can effectively diminish the domination of head classes during the training procedure, and thus can yield more precise classification decision boundaries. However, they are often confronted with the risk of over-fitting tail classes since the original data distribution is distorted and over-parameterized deep networks easily fit this synthetic distribution. To overcome these issues, the recent work [20, 48] decouples the tasks of representation learning and classifier training. This two-stage training scheme first learns visual representation under the original data distribution, and then trains a linear classifier on frozen features under class-balanced sampling. This simple two-stage training scheme turns out to be able to handle the over-fitting issue and sets new state-of-the-art performance on the standard long-tailed benchmarks. Nevertheless, this two-stage training scheme fails to deal with imbalance label distribution issues well, particularly for representation learning stage.

In this paper, our objective is to design a new learning paradigm for long-tailed visual recognition, with the hope of sharing the merits of both types of long-tailed recognition methods, i.e., robust to the over-fitting issue, and effectively handling imbalance label issue. To meet this objective, our idea is to study *how to incorporate the label correlation into a multi-stage training scheme*? Inspired by the work of knowledge distillation [16] in model compression, we observe that soft labels produced by a teacher network are able to capture the inherent relation between classes, which might be helpful for long-tailed recognition by transferring knowledge from head classes to tail classes, as shown in Figure 1. Thus, soft labels provide a practical solution for the multi-stage training strategy with label modeling.

Based on the above analysis, we present a conceptu-

✉: Corresponding author (lmwang@nju.edu.cn).

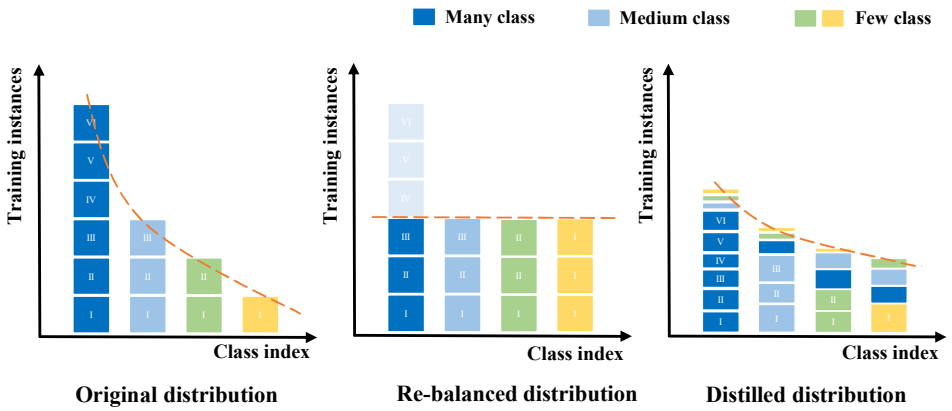


Figure 1. Real-world data always follows long-tailed data distribution, which is dominated by several head classes with abundant samples (i.e, blue cubes) but also contains many tail classes with scarce data (i.e. green and yellow cubes), termed as *original distribution*. Learning directly from long-tailed data can cause a significant performance drop. A common way to deal with the imbalance problem is re-sampling by randomly dropping images from head classes and repeatedly sampling images from tail classes (identical image is marked by unique Roman numeral), resulting in a *re-balanced distribution*. This strategy might lead to over-fitting tailed classes and under-fitting head classes. Inspired by the work of knowledge distillation in model compression, we propose to use the soft labels to deal with imbalance distribution that reflect the inherent relation between classes. The *distilled distribution* acts as a naturally balanced distribution by transferring knowledge from data-rich classes to data-poor classes. Best viewed in color.

ally simple yet particularly effective multi-stage training scheme for long-tailed visual recognition, termed as *Self Supervision to Distillation* (SSD). The key contribution of our SSD is two folds: (1) a self-distillation framework for learning effective long-tailed recognition network; (2) a self-supervision guided distillation label generation module to provide less biased but more informative soft labels for self-distillation. Specifically, we first streamline the multi-stage long-tailed training pipeline within a simple self-distillation framework, in which we are able to naturally mine the label relation automatically and incorporate this intrinsic label structure to improve the generalization performance of multi-stage training. Then, to further improve the robustness of the self-distillation framework, we present an enhanced distillation label generation module by self-supervision from the long-tailed training set itself. Self-supervised learning learns effective visual representation without labels, and can treat each image equally, thus relieving the effect of imbalanced label distribution on soft label generation.

Specifically, we first train an initial teacher network under label supervision and self-supervision simultaneously using instance-balanced sampling. Then, we train a separate linear classifier on top of the visual representation by refining the class decision boundaries with class-balanced sampling. This new classifier yields soft labels of training samples for self-distillation. Finally, we train a self-distillation network under the hybrid supervision of soft labels from previous stages and hard labels from the original training set. As a semantic gap exists between hard labels and soft labels on whether it is biased to head classes, we adopt two classification heads for these two supervisions respectively. We evaluate our SSD training framework for long-tailed visual recognition on the datasets of ImageNet-LT [28], CIFAR100-LT [2], and iNaturalist 2018 [37]. Our approach outperforms other methods on these datasets by a large margin, which verifies the effectiveness of our proposed multi-stage training scheme.

To sum up, the main contribution of this paper is as follows:

- We introduce a simple yet effective multi-stage training framework (SSD). In this framework, we share the merits of re-balanced sampling and decoupled training strategy, by leveraging soft labels modeling into the feature learning stage.
- We propose a self-supervision guided soft label generation module which produces robust soft labels from both data and label domains. These soft labels provide effective information by transferring knowledge from head to tail classes.
- Our SSD achieves the state-of-the-art performance on three challenging long-tailed recognition benchmarks including ImageNet-LT, CIFAR100-LT and iNaturalist 2018 datasets.

## 2. Related Work

### 2.1. Long-tailed Classification

Class re-balanced training [13, 22, 2, 10, 39, 19] has been comprehensively studied for imbalance classification

and long-tailed recognition. Data re-sampling achieved class-balanced training by over-sampling tail classes [3, 13, 6, 33], or under-sampling head classes [10, 1], yet, they might incur generalization problems due to over-fit data-scarce classes or under-fit data-abundant classes. Recent methods overcame the over-fitting problem by augmenting tail class samples with head classes [6, 23]. Another way for re-balanced training was to design class-balanced loss, which gave tail classes larger weights [7, 22, 18, 39] or margins [2], grouped classes with similar number of training samples [26] or ignored negative gradients for tail classes [34]. In addition, researchers tackled long-tailed recognition by transferring knowledge from head classes to tail classes [28, 49, 46, 20, 48]. Features from head classes were used to augment tail classes by maintaining memory banks [28, 49] or modeling intra-class variance [46]. Recent proposed decoupling methods [20, 48] also can be regarded as transferring head classes frozen feature to tail classes in the classifier training stage.

### 2.2. Learning with Distilled Labels

Distilled labels were first adopted by knowledge distillation [16] to transfer knowledge from large teacher models to small student models. BAN [11] proposed to transfer to student models that have the identical architecture to teacher models in a sequential way and made an ensemble of multiple student generations. Our method is different from BAN in that we focus on long-tailed data distribution and produce distilled label by an additional stage of classifiers adjustment. Self-training scheme [42] generated distilled labels for unlabeled data and trains student model with a combination of labeled and unlabeled data.

LFME [41] and RIDE [38] also got help from knowledge distillation for long-tailed visual recognition. LFME [41] grouped categories by the number of training samples and trained multiple experts on these groups. They distilled experts to a unified student model for re-balanced training. RIDE [38] trained multiple experts jointly to reduce the model variance and applied distillation from a powerful model with more experts into a model with fewer experts for model compression. Differing from these methods, our SSD applies distilled labels in a multi-stage training schema for transferring knowledge from head classes to tail classes.

### 2.3. Self-supervised Learning

Self-supervised learning [24, 9, 47, 40, 14, 4, 36] has achieved remarkable progress in recent years, especially in the image representation field, by training models on carefully designed proxy tasks without manual annotations. These tasks could be predicting the image context [9, 32] or rotation [24], image colorization [47], solving jigsaw puzzles [31], maximizing mutual information of global and local features [17] and instance discrimination [40, 14, 4]. Self-supervised methods also inspired studies in many other supervised fields, such as few-shot learning [12] and knowledge distillation [44]. The recent work [45] observed self-supervised pre-trained initialization would benefit long-tail visual recognition, while our goal is to improve the quality of distilled labels with an auxiliary self-supervised task, which is different to [45].

## 3. Methodology

In this section, we provide a detailed description of our long-tailed recognition approach. First, we present an overview of our framework. Then, we introduce how to generate self-supervision guided distillation labels. Finally, we propose to learn generalizable features via self-distillation.

### 3.1. Overall framework

The overall framework of *Self Supervision to Distillation* (SSD) is illustrated in Figure 2. We present a multi-stage long-tailed training pipeline within a self-distillation framework. Our SSD is composed of three steps: (1) self-supervision guided feature learning; (2) intermediate soft labels generation; (3) joint training with self-distillation.

**I-Self-supervision guided feature learning.** We train networks on classification tasks under the original long-tailed data distribution during this phase. The classification tasks consist of two parts: the conventional $C$-way classification task that aims to classify images into $C$ semantic categories, and the balanced self-supervised classification task purely from data itself. Although the $C$-way classification task provides rich semantic information, it is also biased by the long-tailed labels. Samples of tailed classes might be overwhelmed by data-rich classes, resulting in an under-representation issue [34, 6]. Therefore, we construct balanced self-supervised classification tasks, e.g., predicting the image rotation [24] and instance discrimination [40, 14], which consider each image equally without the influence of labels. Rotation prediction [24] recognizes the rotation angle among $\{0°, 90°, 180°, 270°\}$. Instance discrimination [40, 14] regards each image as a single category which is equal to $N$-way classification, where $N$ is the number of images in the training set. We will describe the details of self-supervised learning methods in Section 3.2.

**II-Intermediate soft labels generation.** During this phase, the classifier needs to be tuned under the class-balanced setting on top of the frozen features to generate distilled labels. We choose the Learnable weight scaling (LWS) approach following [20] for its consistently good performance in various settings. It learns to re-scale the weights of the classifier to avoid the tendency to head classes. Given an image $\mathbf{x}$, the fine-tuned classifier provides relatively balanced and soft labels $\widetilde{\mathbf{y}}$ that integrates both label-driven and data-driven in-

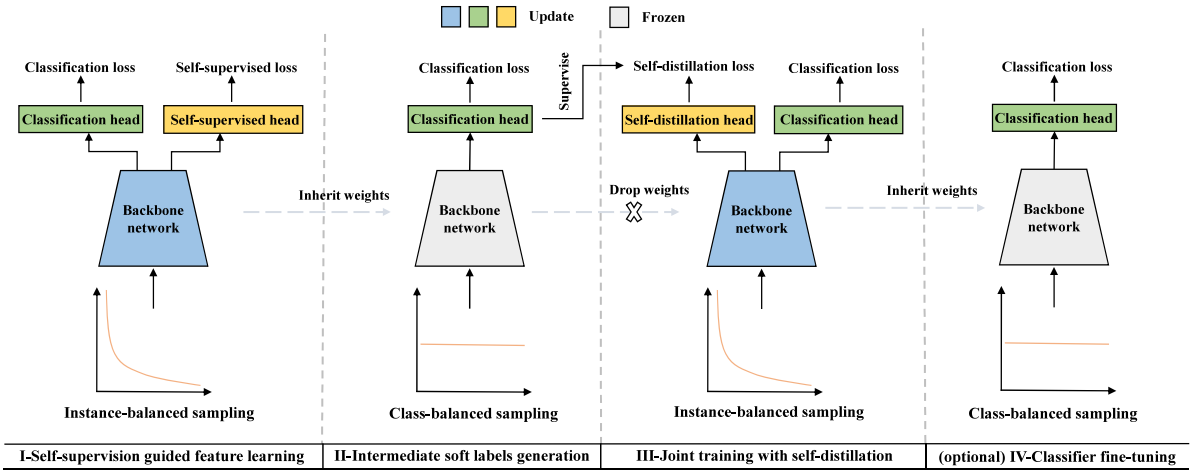


Figure 2. The pipeline of our Self Supervision to Distillation (SSD) framework. First, we train an initial feature network under label supervision and self-supervision jointly using instance-balanced sampling. Then, we refine the class decision boundaries with class-balanced sampling to generate soft labels by fixing the feature backbone. Finally, we train a self-distillation network with two classification heads under the supervision of both soft labels from previous stages and hard labels from the original training set.

formation, acting as teacher supervision for the next step of self-distillation.

**III-Joint training with self-distillation.** As the representation and classifier are trained separately under different sampling strategies, the entire network might not be optimal. However, direct fine-tuning of the backbone network during classifier learning stage will hurt the generalization power [20]. Instead, we propose to jointly train another backbone network and classifier under the original long-tailed data distribution with hybrid supervisions of original labels and balanced distilled labels. We **re-initialize** the network at this stage as previous representations are still relatively biased, and it is hard to escape from the local minimal via fine-tuning. In addition, other self-training paper [42] finds similar conclusions that training the student from scratch is better than initializing the student with the teacher. Training details of self-distillation can be found in Section 3.3. After learning from hybrid supervision, the final model can achieve higher performance than the teacher model. **Also, an extra classifier fine-tuning step is optional but recommended for further performance improvement (IV-Classifier fine-tuning)**.

### 3.2. Feature learning enhanced by self-supervision

In phase-I of the feature learning stage, we choose to train the backbone network using a standard supervised task and a self-supervised task in a multi-task learning way. The supervised task might ignore images of data-scarce classes due to highly biased labels, while the self-supervised task treats every sample equally without the influence of long-tailed labels. Formally, let $\theta$ be parameters of the shared backbone network, and $\omega_{sup}$ and $\omega_{self}$ are parameters for the supervised task and self-supervised task respectively. Then the loss function of self-supervised task for an input image $\mathbf{x}$ with label $\mathbf{y}$ can be written as $\mathcal{L}_{self}(\mathbf{x};\theta,\omega_{self})$, and $\mathcal{L}_{sup}(\mathbf{x},\mathbf{y};\theta,\omega_{sup})$ is for the supervised cross-entropy loss. The total loss of this stage is illustrated as:

$$\mathcal{L} = \alpha_1 \mathcal{L}_{sup}(\mathbf{x};\theta,\omega_{sup}) + \alpha_2 \mathcal{L}_{self}(\mathbf{x},\mathbf{y};\theta,\omega_{self}), \quad (1)$$

where $\alpha_1$ and $\alpha_2$ are hyper-parameters and equal to 1 in our experiments. We choose rotation prediction and instance discrimination as self-supervised proxy tasks. The network can learn to represent images properly by solving these proxy tasks.

**Rotation prediction.** Predicting image rotation is a classical self-supervised task that is simple yet effective. Given an image $\mathbf{x}$, we randomly rotate it by an angle among $\{0°, 90°, 180°, 270°\}$ to obtain a rotated image $\mathbf{x}'$. These two images are sent to the network simultaneously. The original image $\mathbf{x}$ is used for the original cross-entropy loss. The rotated image $\mathbf{x}'$ is chosen for predicting the rotation degree which can be formulated to a 4-way balanced classification problem. In this case, the specific parameters $\omega_{self}$ are implemented as a conventional 4-way linear classifier.

**Instance discrimination.** In the instance discrimination task, each image is treated as a distinct class, and it would learn a non-parametric classifier to categorize each image. Formally, let $\mathbf{v}_i$ denote the $\ell_2$-normalized embedding of image $i$ and $\mathbf{v'}_i$ is the $\ell_2$-normalized embedding extracted from a copy of image $i$ with different transformations. The loss of instance discrimination can be:

$$\mathcal{L}_{self} = -\log\left(\frac{\exp(\mathbf{v}_i \mathbf{v'}_i/\tau)}{\exp(\mathbf{v}_i \mathbf{v'}_i/\tau) + \sum_K \exp(\mathbf{v}_i \mathbf{v'}_k/\tau)}\right), \quad (2)$$

where $\tau$ is the temperature, and $K$ is the number of other images as negative samples, which can be retrieved from

memory bank [40, 14] and current mini-batch [4]. Following [5], we maintain a momentum network with a feature queue to produce a large number of negative samples and utilize MLP projection head $\omega_{self}$ to transform the backbone output to a low-dimensional feature space.

### 3.3. Long-tailed recognition via self-distillation

Knowledge distillation [16] is first introduced for transferring knowledge from high-capability networks (teacher models) to small networks (student models) via soft labels. Our SSD method is inspired by knowledge distillation, yet exhibits essential difference with it. In our SSD method, student models are identical to teacher models, but are learned under different sampling strategies. Also, in particular for long-tailed recognition, the dark knowledge in soft labels can be helpful by transferring knowledge from head classes to tail classes. Due to the complementary properties of soft and hard labels, we propose a customized design by applying two separate classifiers supervised by hard and soft labels, respectively.

More formally, we denote $\mathbf{x}$ a training image with its hard label $\mathbf{y}$ and soft label $\widetilde{\mathbf{y}}$. We aim to learning an embedding function $\mathcal{F}$ that encodes $\mathbf{x}$ into feature vector $\mathbf{f} = \mathcal{F}(\mathbf{x};\theta)$, as well as two classifiers $\mathcal{G}_{hard}$ and $\mathcal{G}_{soft}$. The feature vector $\mathbf{f}$ will be sent to two linear classifiers $\mathcal{G}_{hard}$ and $\mathcal{G}_{soft}$ to get output logits $\mathbf{z}^{hard} = \mathcal{G}_{hard}(\mathbf{f})$ and $\mathbf{z}^{soft} = \mathcal{G}_{soft}(\mathbf{f})$. Let $\widetilde{\mathbf{z}}$ denote the output logits of teacher model, then the soft label is given by:

$$\widetilde{y}_i = \frac{\exp(\widetilde{z}_i/T)}{\sum_{k=1}^{C}\exp(\widetilde{z}_k/T)}, \tag{3}$$

where $i$ is the category index and $T$ is the temperature which is set to 2 by default. Then, the knowledge distillation loss is written as:

$$\mathcal{L}_{kd}(\widetilde{\mathbf{y}}, \mathbf{z}^{soft}) = -T^2\sum_{i=1}^{C}\widetilde{y}_i\log(\frac{\exp(z_i^{soft}/T)}{\sum_{k=1}^{C}\exp(z_k^{soft}/T)}). \tag{4}$$

For hard label supervision, we utilize the standard cross entropy loss $\mathcal{L}_{ce}$. Thus, the final loss is the combination of these two losses:

$$\mathcal{L} = \lambda_1\mathcal{L}_{ce}(\mathbf{y}, \mathbf{z}^{hard}) + \lambda_2\mathcal{L}_{kd}(\widetilde{\mathbf{y}}, \mathbf{z}^{soft}), \tag{5}$$

where both of $\lambda_1$ and $\lambda_2$ are the weight of each loss and set to 1 in our experiments.

## 4. Experiments

### 4.1. Experimental settings

**Datasets.** We perform extensive experiments on three long-tailed visual recognition benchmarks: ImageNet-LT [28], CIFAR100-LT [2] and iNaturalist 2018 [37].

**ImageNet-LT** is constructed from ImageNet-2012 [8] by sampling a subset following the Pareto distribution with the power value $\alpha = 6$, which contains 1000 classes. The training set has 115.8K images and the number of images per class range from 1280 to 5 images. Both the validation set and the test set are balanced which contain 20K and 50K images respectively. We select the hyper-parameters on the validation set and report numerical results on the test set.

**CIFAR100-LT** is a set of long-tailed datasets with different imbalance factors sampled from CIFAR-100 [25] with 100 categories. The imbalance factor is defined as the ratio between the number of images for the most frequent class and the least frequent class, which is set to 10, 50 and 100 in our experiments. There are 100 images per class in the validation set.

**iNaturalist 2018** is a real-world, naturally long-tailed dataset which is composed of 8,142 fine-grained species. The training set contains 437.5K images and its imbalance factor is equal to 500. We use the official validation set to test our approach which has 3 images per class.

**Evaluation protocol.** We evaluate our proposed SSD methods on the corresponding balanced validation/test datasets and report the top-1 accuracy, denoted by overall accuracy. Following the previous studies [28], we also report the accuracy of three splits according to the number of training samples per class: Many-shot ($\geq$100), Medium-shot (20$\sim$100) and Few-shot ($\leq$20) for diagnosing the source of improvement in the ImageNet-LT dataset.

### 4.2. Comparisons with the state-of-the-art methods

In this section, we demonstrate the effectiveness of our SSD method by comparing its performance to other state-of-the-art methods on ImageNet-LT, CIFAR100-LT and iNaturalist 2018 datasets. Numerical results can be found in Table 1, Table 2, and Table 3.

#### 4.2.1 Experimental results on ImageNet-LT

We conduct extensive experiments on the ImageNet-LT dataset and report the results of each split in Table 1. We compare to methods that use ResNeXt-50 [43] as backbone network. We follow the same training strategy in [20] except for changing batch size to 256 due to GPU memory limitation and linearly decreasing the learning rate from 0.2 to 0.1.

For phase-I of feature learning, we utilize instance discrimination as the self-supervised task for ImageNet-LT. Following MoCov2 [5], we set $\tau = 0.2$, $K = 65536$, and update the momentum encoder with the momentum equal to 0.999. Stronger data augmentation from MoCov2 [5] is adopted only for the input of the momentum encoder. Since

| Methods | Many | Medium | Few | Overall |
|---|---|---|---|---|
| Cross Entropy | 65.9 | 37.5 | 7.7 | 44.4 |
| OLTR [28] | - | - | - | 46.3 |
| NCM [20] | 56.6 | 45.3 | 28.1 | 47.3 |
| cRT [20] | 61.8 | 46.2 | 27.4 | 49.6 |
| LWS [20] | 60.2 | 47.2 | 30.3 | 49.9 |
| De-confound [35] | 62.7 | 48.8 | 31.6 | 51.8 |
| cRT* | 62.6 | 46.9 | 27.9 | 50.3 |
| LWS* | 61.1 | 48.0 | 31.5 | 50.7 |
| **SSD (ours)** | 64.2 (+3.1) | 50.8 (+2.8) | 34.5 (+3.0) | 53.8 (+3.1) |
| cRT*‡ | 64.2 | 47.7 | 27.8 | 51.3 |
| LWS*‡ | 63.4 | 48.6 | 32.3 | 52.1 |
| **SSD (ours)**‡ | **66.8** (+3.4) | **53.1** (+4.5) | **35.4** (+3.1) | **56.0** (+3.9) |

Table 1. Top-1 accuracy on ImageNet-LT dataset. Comparison to the state-of-the-art methods with ResNeXt-50 as backbone. We report absolute improvements against LWS with the same hyper-parameters. * indicates our reproduced results with the released code. Results marked with ‡are trained with $1.5\times$ scheduler.

the contrastive loss needs longer training iterations to converge [14, 4], we also test a longer training scheduler of 135 epochs ($1.5\times$ of 90 epochs in original setting) to sufficiently unleash the performance gain of the self-supervised task, term as $1.5\times$ scheduler. After joint training with soft labels and hard labels, another stage for class boundary adjustment is adopted by default for all datasets.

For a fair comparison, we re-implement baseline models of cRT and LWS [20] with our hyper-parameters. As shown in Table 1, smaller batch size and learning rate benefits long-tailed recognition ($50.7\%$ v.s. $49.9\%$), and we can get further improvements with $1.5\times$ scheduler especially for LWS ($+1.4\%$). Compared with these strong baselines, our method brings consistent performance improvement on every split by a large margin, with $+3.1\%$ and $+3.9\%$ improvements for $1\times$ and $1.5\times$ training scheduler. We arrive at $56.0\%$ for overall performance, which sets a new state-of-the-art performance on the ImageNet-LT dataset.

#### 4.2.2 Experimental results on CIFAR100-LT

We evaluate our method on the CIFAR100-LT dataset with the imbalance factor of 100, 50 and 10. We use ResNet-32 [15] as our backbone network for fair comparison.

For stage I and III, we completely follow the setting of BBN [48] including data augmentation strategies, warming-up training scheduler and batch size. Due to the smaller size of training set, we only adopt rotation prediction as the self-supervised task during phase-I. The rotated images are used for predicting their degree of rotation. The images with the original orientation are applied for the supervised cross entropy loss. As for distilled labels generation, we train the LWS classifier for five epochs with the learning rate of 0.2 and the batch size of 512, following [20]. We re-implement some baseline models, including cross-entropy loss under uniform sampling and decoupled training with LWS classifier to fairly compare with our method.

| Methods | Imbalance factor | | |
|---|---|---|---|
| | 100 | 50 | 10 |
| Cross Entropy (CE)* | 39.1 | 44.0 | 55.8 |
| Focal [27] | 38.4 | 44.3 | 55.8 |
| LDAM-DRW [2] | 42.0 | 46.6 | 58.7 |
| LWS* [20] | 42.3 | 46.0 | 58.1 |
| CE-DRW [48] | 41.5 | 45.3 | 58.2 |
| CE-DRS [48] | 41.6 | 45.5 | 58.1 |
| BBN [48] | 42.6 | 47.0 | 59.1 |
| M2m [23] | 43.5 | - | 57.6 |
| LFME [41] | 43.8 | - | - |
| Domain Adaption [19] | 44.1 | 49.1 | 58.0 |
| De-confound [35] | 44.1 | 50.3 | 59.6 |
| **SSD (ours)** | **46.0** | **50.5** | **62.3** |

Table 2. Top-1 accuracy on CIFAR100-LT dataset with the imbalance factor of 100, 50 and 10. We compare with state-of-the-art methods with ResNet-32 as backbone network. * indicates our reproduced results with the released code.

| Methods | Top-1 Acc. | |
|---|---|---|
| | $1\times$ | $2\times$ |
| CB-Focal [2] | 61.1 | - |
| LDAM [2] | 64.6 | - |
| LDAM+DRW [2] | 68.0 | - |
| LDAM+DRW† [2] | 64.6 | 66.1 |
| $\tau$-norm‡ [20] | 65.6 | 69.3 |
| cRT‡ [20] | 65.2 | 68.5 |
| LWS‡ [20] | 65.9 | 69.5 |
| CE-DRW [48] | 63.7 | - |
| CE-DRS [48] | 63.6 | - |
| BBN [48] | 66.3 | 69.6 |
| FSA [6] | 65.9 | - |
| LWS‡* [20] | 66.6 | 69.5 |
| **SSD (ours)**‡ | **69.3** | **71.5** |

Table 3. Top-1 accuracy on iNaturalist 2018 dataset with $1\times$ and $2\times$ schedulers and comparison to state-of-the-art methods with ResNet-50 as backbone. * indicates our reproduced results. Results marked by † are cited from [48]. $2\times$ means using 200 epochs training scheduler for methods marked by ‡ and 180 epochs for other methods.

Experimental results and the comparison with other state-of-the-art methods are reported in Table 2. Our SSD outperforms the strong baseline LWS by from $3.7\%$ to $4.5\%$ for different imbalance factors, which demonstrates that our method is robust to different imbalanced situations. We achieve the state-of-the-art performance across all imbalance factors.

#### 4.2.3 Experimental results on iNaturalist 2018

We also investigate our method on the naturally long-tailed dataset of iNaturalist 2018. Following the common practice, we utilize ResNet-50 [15] as backbone for fair comparison and follow the same training strategy in [20] except for changing batch size to 256 due to GPU memory limitation and linearly decreasing the learning rate from 0.2 to

| Methods | 1.5× | I | II | III-hard (test) | III-soft (test) | IV-LWS | Many | Medium | Few | Overall |
|---|---|---|---|---|---|---|---|---|---|---|
| CE | | | | | | | 66.9 | 38.0 | 8.1 | 45.1 |
| | ✓ | | | | | | 67.9 | 39.5 | 9.5 | 46.3 |
| LWS | | | | | | | 61.1 | 48.0 | 31.5 | 50.7 |
| | ✓ | | | | | | 63.4 | 48.6 | 32.3 | 52.1 |
| Our SSD | ✓ | ✓ | | | | | 69.8 | 42.8 | 11.0 | 48.9 |
| | ✓ | ✓ | ✓ | | | | 64.9 | 51.1 | 34.0 | 54.1 |
| | ✓ | | ✓ | | | ✓ | 66.0 | 50.8 | 34.2 | 54.4 |
| | ✓ | ✓ | ✓ | ✓ | | | **71.1** | 46.1 | 15.6 | 51.6 |
| | ✓ | ✓ | ✓ | | ✓ | | 67.1 | 52.8 | 33.3 | 55.7 |
| | ✓ | ✓ | ✓ | | | ✓ | 66.8 | **53.1** | **35.4** | **56.0** |

Table 4. Ablation study on ImageNet-LT. We investigate the effectiveness of each stage of our proposed SSD method. Different stage are marked by Roman numerals I, II, III. The outputs of hard classifier and soft classifier are termed as III-hard and III-soft. IV-LWS means an extra classifier fine-tuning stage by LWS after self-distillation.

0.1. The training setting of the self-supervised task is the same with our recipe for ImageNet-LT. Following [20], we train our SSD with two different schedulers, 90 and 200 epochs (termed as $1\times$ and $2\times$ schedulers) for converging sufficiently.

Table 3 reports the Top-1 accuracy of various methods. We also re-implement baseline model LWS with our batch size and learning rate, which obtains $0.7\%$ improvement for $1\times$ scheduler but shows the same performance for longer training epochs. For $1\times$ training scheduler, we obtain $2.7\%$ improvement against the result of LWS. Even for the stronger baseline of $69.5\%$ with a longer training scheduler, our SSD can still outperform it by $2.7\%$.

## 4.3. Ablation studies

### 4.3.1 Effectiveness of each training stage

In this study, we investigate the contribution of each stage in our proposed SSD framework on the ImageNet-LT dataset, which is shown in Table 4. We set several baselines for reference: the plain model with cross-entropy loss (CE) using instance-balanced sampling with or without $1.5\times$ training scheduler, and decoupled methods using LWS classifier with or without $1.5\times$ training scheduler. Also, we tear our SSD method into three stages, marked by Roman numerals. For two classification heads of self-distillation, III-hard (test) is for using the output of classifier $\mathcal{G}_{hard}$ supervised by the hard label for recognition and III-soft (test) is for using the output of classifier $\mathcal{G}_{soft}$ supervised by the soft label for recognition. In respect that the classifier $\mathcal{G}_{hard}$ is still biased to head classes, after self-distillation, we propose run another classifier adjustment stage using LWS for further improvement, termed as IV-LWS.

**Self-supervision guided feature learning.** The effectiveness of the self-supervision guided feature can be verified through this study. Compared with CE baseline with $1.5\times$ scheduler ($46.3\%$), feature learning with instance discrimination task as self-supervision brings $+2.6$ improvement for the overall performance. Since self-supervised tasks treat each image equally, the improvements come from all three splits of the test set, e.g., $+1.9\%$ for many-shot classes, $+3.3\%$ for medium-shot classes, and $+1.6\%$ for few-shot classes. Also, the result of IV-LWS without self-supervision demonstrates the effectiveness of self-supervision in teacher training, which could help to generate better soft labels, making the distilled student achieve higher performance ($56.0\%$ vs. $54.4\%$). We randomly select features of 15 classes from the ImageNet-LT dataset and visualize them via t-SNE in Figure. 3. Compared with baseline, features trained with self-supervision (SS) are more separable, especially for tailed classes. For example, when training without SS, the test samples of purple class cover larger space and can not be well separated from others. Instead, using SS is more stable and does not interfere with other classes.

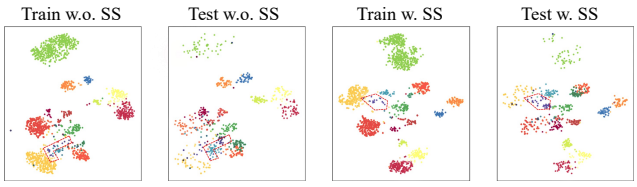


Figure 3. Visualization of self-supervision guided feature learning.

**Long-tailed recognition via self-distillation.** We also investigate the effectiveness of the self-supervision guided distilled label. We term the distilled label as the teacher model for less confusion, whose performance is $54.1\%$ for the overall test set. As shown in Table 4, III-soft (test) outperforms the teacher model by $1.6\%$. The teacher model is adjusted by LWS, which sacrifices the performance of head classes ($69.8\%$ to $64.9\%$) to improve tail classes. In contrast, our self-distillation improves the performance of many- and medium-shot classes with a slight accuracy drop of tail classes thanks to the goodness of balanced soft supervision and instance-balanced sampling. Also, III-hard (test) exhibits the best performance among results using hard labels and instance-balanced sampling ($+2.7\%$ than Stage-I and $+5.3\%$ against CE baseline). Fine-tuning the biased classifier III-hard using LWS (IV-LWS) can achieve further improvement of $0.3\%$, which is adopted by default.

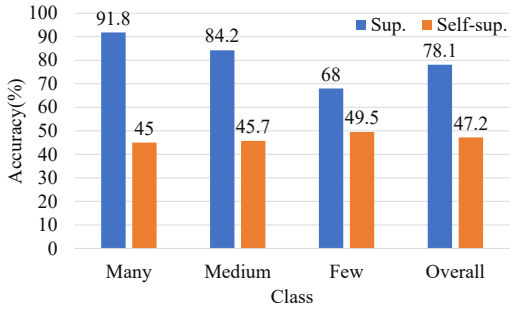


Figure 4. Training top-1 accuracy for supervised and self-supervised tasks for many-shot, medium-shot, few-shot and overall classes on the ImageNet-LT dataset.

| Methods | Many | Medium | Few | Overall |
|---|---|---|---|---|
| Plain | 67.9 | 39.5 | 9.5 | 46.3 |
| Teacher model | 64.9 | 51.1 | **34.0** | 54.1 |
| Coupled | 68.6 | 49.1 | 23.8 | 53.2 |
| Single | 67.4 | 52.0 | 31.3 | 55.1 |
| **Our III-hard** | **71.1** | 46.1 | 15.6 | 51.6 |
| **Our III-soft** | 67.1 | **52.8** | 33.3 | **55.7** |

Table 5. Top-1 accuracy of different self-distillation strategies on the test set of ImageNet-LT.

#### 4.3.2 Study on different self-distillation strategies

To demonstrate the effectiveness of our self-distillation module, we evaluate several distillation strategies as baselines: (1) *Coupled* self-distillation which is the conventional way of knowledge distillation and trains a single classifier using both hard and soft labels; (2) *Single* self-distillation, which only use soft labels to train the classifier. The numerical results are provided in Table 5. The accuracy of the teacher model is $54.1\%$ which has the highest performance of few-shot classes. The coupled method surpasses the plain model due to the abundant knowledge in soft labels. However, it does not reach the performance of teacher model because there is interference between hard and soft labels, resulting in limited improvement in medium- and few-shots classes. Also, the soft classifier of our proposed hybrid supervision strategy outperforms the single one, which indicates that the hard labels might be able to provide complementary knowledge for feature learning.

#### 4.3.3 Evaluation on self-supervised task

For a better understanding of the self-supervised task on long-tailed data, we visualize the training top-1 accuracy of both supervised classification and self-supervised instance discrimination on the ImageNet-LT dataset in Figure 4. As we can see, even for images from the training set, the accuracy of tailed categories is still very low. However, results of instance discrimination are more stable among different splits, which proves our motivation that self-supervised learning can treat each image equally during the training procedure, thus relieving the effect of imbalanced label distribution.

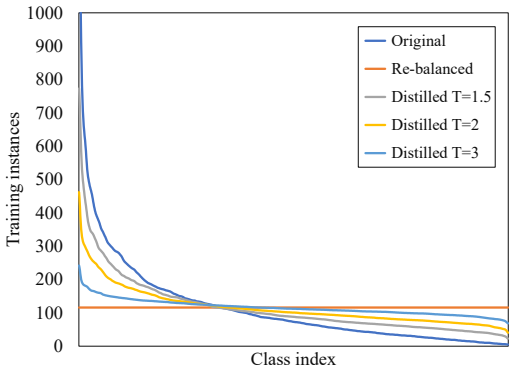


Figure 5. Visualization for different training strategy on ImageNet-LT dataset. *Original*, *Re-balanced* and *Distilled* denote distribution for original long-tailed data, after class-balanced sampling and distilled label.

### 4.4. Visualization

We visualize the distribution of training samples for the distilled label on ImageNet-LT in Figure 5. We sum up the softmax of logits divided by temperature $T$ to calculate the training sample distribution of distilled labels. Unlike conventional knowledge distillation that uses temperature to smooth the label distribution of a single image, we consider taking it to flatten the data distribution of the entire dataset by suppressing the frequency of head classes and hope to transfer knowledge from head classes to tail classes. A larger temperature will result in a more flat distribution. We choose $T = 2$ in our experiments for self-distillation by cross-validation.

## 5. Conclusion

In this paper we have introduced a simple yet effective multi-stage training framework for long-tailed visual recognition by leveraging distilled labels (SSD). Training with distilled supervision can overcome the over-fitting issue of re-balanced methods and endow the network with relatively-balanced information for feature learning. In addition, we propose to generate soft labels guided by self-supervised learning, by leveraging both top-down semantics and bottom-up data structure. Our SSD achieves the state-of-the-art performance on three long-tailed recognition benchmarks, ImageNet-LT, CIFAR100-LT and iNaturalist 2018. We hope our SSD opens a new direction in long-tailed visual recognition via knowledge transfer to learn more powerful representation.

**Acknowledgements.** This work is supported by National Natural Science Foundation of China (No. 62076119, No. 61921006), Program for Innovative Talents and Entrepreneur in Jiangsu Province, and Collaborative Innovation Center of Novel Software Technology and Industrialization.

## References


[1] Mateusz Buda, Atsuto Maki, and Maciej A. Mazurowski. A systematic study of the class imbalance problem in convolutional neural networks. *Neural Networks*, 106:249–259, 2018. 3

[2] Kaidi Cao, Colin Wei, Adrien Gaidon, Nikos Arechiga, and Tengyu Ma. Learning imbalanced datasets with label-distribution-aware margin loss. In *Advances in Neural Information Processing Systems*, 2019. 1, 2, 3, 5, 6

[3] Nitesh V. Chawla, Kevin W. Bowyer, Lawrence O. Hall, and W. Philip Kegelmeyer. SMOTE: synthetic minority oversampling technique. *J. Artif. Intell. Res.*, 16:321–357, 2002. 3

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

[5] Xinlei Chen, Haoqi Fan, Ross B. Girshick, and Kaiming He. Improved baselines with momentum contrastive learning. *CoRR*, abs/2003.04297, 2020. 5

[6] Peng Chu, Xiao Bian, Shaopeng Liu, and Haibin Ling. Feature space augmentation for long-tailed data. In *ECCV 2020*. 1, 3, 6

[7] Yin Cui, Menglin Jia, Tsung-Yi Lin, Yang Song, and Serge J. Belongie. Class-balanced loss based on effective number of samples. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR*, 2019. 3

[8] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *IEEE conference on computer vision and pattern recognition*, 2009. 1, 5

[9] Carl Doersch, Abhinav Gupta, and Alexei A Efros. Unsupervised visual representation learning by context prediction. In *Proceedings of the IEEE international conference on computer vision*, 2015. 3

[10] Chris Drummond, Robert C Holte, et al. C4. 5, class imbalance, and cost sensitivity: why under-sampling beats over-sampling. Citeseer, 2003. 1, 2, 3

[11] Tommaso Furlanello, Zachary Chase Lipton, Michael Tschannen, Laurent Itti, and Anima Anandkumar. Born-again neural networks. In *Proceedings of the 35th International Conference on Machine Learning, ICML*, 2018. 3

[12] Spyros Gidaris, Andrei Bursuc, Nikos Komodakis, Patrick Pérez, and Matthieu Cord. Boosting few-shot visual learning with self-supervision. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019. 3

[13] Hui Han, Wenyuan Wang, and Binghuan Mao. Borderline-smote: A new over-sampling method in imbalanced data sets learning. In *Advances in Intelligent Computing, International Conference on Intelligent Computing, ICIC 2005*. 1, 2, 3

[14] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020. 3, 5, 6

[15] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2016. 6

[16] Geoffrey E. Hinton, Oriol Vinyals, and Jeffrey Dean. Distilling the knowledge in a neural network. *NIPS Deep Learning and Representation Learning Workshop*, 2015. 1, 3, 5

[17] R Devon Hjelm, Alex Fedorov, Samuel Lavoie-Marchildon, Karan Grewal, Phil Bachman, Adam Trischler, and Yoshua Bengio. Learning deep representations by mutual information estimation and maximization. *arXiv preprint arXiv:1808.06670*, 2018. 3

[18] Chen Huang, Yining Li, Chen Change Loy, and Xiaoou Tang. Learning deep representation for imbalanced classification. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016*. 3

[19] 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 *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR*, 2020. 2, 6

[20] Bingyi Kang, Saining Xie, Marcus Rohrbach, Zhicheng Yan, Albert Gordo, Jiashi Feng, and Yannis Kalantidis. Decoupling representation and classifier for long-tailed recognition. In *International Conference on Learning Representations*, 2020. 1, 3, 4, 5, 6, 7

[21] Will Kay, João Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, Mustafa Suleyman, and Andrew Zisserman. The kinetics human action video dataset. *CoRR*, abs/1705.06950, 2017. 1

[22] Salman H. Khan, Munawar Hayat, Mohammed Bennamoun, Ferdous Ahmed Sohel, and Roberto Togneri. Cost-sensitive learning of deep feature representations from imbalanced data. *IEEE Trans. Neural Networks Learn. Syst.*, 29(8):3573–3587, 2018. 1, 2, 3

[23] Jaehyung Kim, Jongheon Jeong, and Jinwoo Shin. M2m: Imbalanced classification via major-to-minor translation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020. 3, 6

[24] Nikos Komodakis and Spyros Gidaris. Unsupervised representation learning by predicting image rotations. In *International Conference on Learning Representations (ICLR)*, 2018. 3

[25] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009. 5

[26] Yu Li, Tao Wang, Bingyi Kang, Sheng Tang, Chunfeng Wang, Jintao Li, and Jiashi Feng. Overcoming classifier imbalance for long-tail object detection with balanced group softmax. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020*. 3

[27] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, 2017. 6

[28] Ziwei Liu, Zhongqi Miao, Xiaohang Zhan, Jiayun Wang, Boqing Gong, and Stella X. Yu. Large-scale long-tailed

recognition in an open world. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR*, 2019. 2, 3, 5, 6

[29] Dhruv Mahajan, Ross Girshick, Vignesh Ramanathan, Kaiming He, Manohar Paluri, Yixuan Li, Ashwin Bharambe, and Laurens Van Der Maaten. Exploring the limits of weakly supervised pretraining. In *Proceedings of the European conference on computer vision (ECCV)*, 2018. 1

[30] Mark EJ Newman. Power laws, pareto distributions and zipf's law. *Contemporary physics*, 46(5):323–351, 2005. 1

[31] Mehdi Noroozi and Paolo Favaro. Unsupervised learning of visual representations by solving jigsaw puzzles. In *European conference on computer vision*, 2016. 3

[32] Deepak Pathak, Philipp Krahenbuhl, Jeff Donahue, Trevor Darrell, and Alexei A Efros. Context encoders: Feature learning by inpainting. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016. 3

[33] Li Shen, Zhouchen Lin, and Qingming Huang. Relay backpropagation for effective learning of deep convolutional neural networks. In *European conference on computer vision*, 2016. 1, 3

[34] Jingru Tan, Changbao Wang, Buyu Li, Quanquan Li, Wanli Ouyang, Changqing Yin, and Junjie Yan. Equalization loss for long-tailed object recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020. 3

[35] Kaihua Tang, Jianqiang Huang, and Hanwang Zhang. Long-tailed classification by keeping the good and removing the bad momentum causal effect. In *NeurIPS*, 2020. 6

[36] Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive multiview coding. *CoRR*, abs/1906.05849, 2019. 3

[37] Grant Van Horn, Oisin Mac Aodha, Yang Song, Yin Cui, Chen Sun, Alex Shepard, Hartwig Adam, Pietro Perona, and Serge Belongie. The inaturalist species classification and detection dataset. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018. 2, 5

[38] Xudong Wang, Long Lian, Zhongqi Miao, Ziwei Liu, and Stella Yu. Long-tailed recognition by routing diverse distribution-aware experts. In *International Conference on Learning Representations*, 2021. 3

[39] Yu-Xiong Wang, Deva Ramanan, and Martial Hebert. Learning to model the tail. In *Advances in Neural Information Processing Systems*, 2017. 1, 2, 3

[40] Zhirong Wu, Yuanjun Xiong, Stella X Yu, and Dahua Lin. Unsupervised feature learning via non-parametric instance discrimination. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018. 3, 5

[41] Liuyu Xiang, Guiguang Ding, and Jungong Han. Learning from multiple experts: Self-paced knowledge distillation for long-tailed classification. In *European Conference on Computer Vision*, 2020. 3, 6

[42] Qizhe Xie, Minh-Thang Luong, Eduard Hovy, and Quoc V. Le. Self-training with noisy student improves imagenet classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 3, 4

[43] Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. Aggregated residual transformations for deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017. 5

[44] Guodong Xu, Ziwei Liu, Xiaoxiao Li, and Chen Change Loy. Knowledge distillation meets self-supervision. In *European Conference on Computer Vision*, 2020. 3

[45] Yuzhe Yang and Zhi Xu. Rethinking the value of labels for improving class-imbalanced learning. *CoRR*, abs/2006.07529, 2020. 3

[46] Xi Yin, Xiang Yu, Kihyuk Sohn, Xiaoming Liu, and Manmohan Chandraker. Feature transfer learning for face recognition with under-represented data. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019. 3

[47] Richard Zhang, Phillip Isola, and Alexei A Efros. Colorful image colorization. In *European conference on computer vision*, 2016. 3

[48] Boyan Zhou, Quan Cui, Xiu-Shen Wei, and Zhao-Min Chen. Bbn: Bilateral-branch network with cumulative learning for long-tailed visual recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020. 1, 3, 6

[49] Linchao Zhu and Yi Yang. Inflated episodic memory with region self-attention for long-tailed visual recognition. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020*. 3