Balanced Contrastive Learning for Long-Tailed Visual Recognition

Jianggang Zhu[1,2]*, Zheng Wang[1,2]*, Jingjing Chen[1,2]†, Yi-Ping Phoebe Chen[3] and Yu-Gang Jiang[1,2]

[1]Shanghai Key Lab of Intelligent Information Processing, School of Computer Science, Fudan University
[2]Shanghai Collaborative Innovation Center on Intelligent Visual Computing
[3]Department of Computer Science and Information Technology, La Trobe University

{jgzhu20,zhengwang17,chengjingjing,ygj}@fudan.edu.cn, phoebe.chen@latrobe.edu.au

Abstract

Real-world data typically follow a long-tailed distribution, where a few majority categories occupy most of the data while most minority categories contain a limited number of samples. Classification models minimizing cross-entropy struggle to represent and classify the tail classes. Although the problem of learning unbiased classifiers has been well studied, methods for representing imbalanced data are under-explored. In this paper, we focus on representation learning for imbalanced data. Recently, supervised contrastive learning has shown promising performance on balanced data recently. However, through our theoretical analysis, we find that for long-tailed data, it fails to form a regular simplex which is an ideal geometric configuration for representation learning. To correct the optimization behavior of SCL and further improve the performance of long-tailed visual recognition, we propose a novel loss for balanced contrastive learning (BCL). Compared with SCL, we have two improvements in BCL: class-averaging, which balances the gradient contribution of negative classes; class-complement, which allows all classes to appear in every mini-batch. The proposed balanced contrastive learning (BCL) method satisfies the condition of forming a regular simplex and assists the optimization of cross-entropy. Equipped with BCL, the proposed two-branch framework can obtain a stronger feature representation and achieve competitive performance on long-tailed benchmark datasets such as CIFAR-10-LT, CIFAR-100-LT, ImageNet-LT, and iNaturalist2018.

1. Introduction

Deep neural networks have achieved remarkable success in a series of computer vision tasks, such as image recognition [4,12,25], video analysis [37,49], object detection [35],
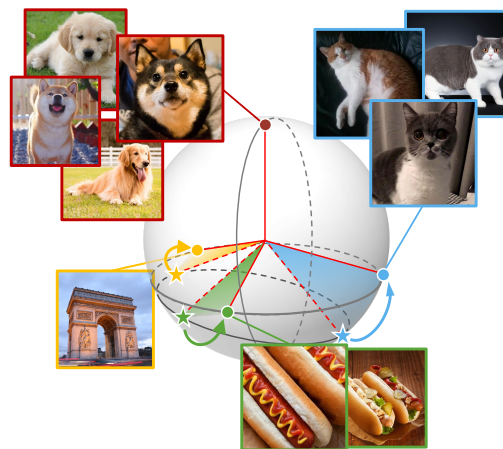

Figure 1. Illustration of Balanced Contrastive Learning. Head classes dominant the training procedure of SCL and compress the representation space of tail classes on the hypersphere (denoted by ★). BCL learns an embedding space that treats all classes equally and forms a regular simplex (denoted by ●).

etc. These achievements are owing largely to the availability of large-scale dataset such as ImageNet [11], where each class has sufficient and equal amount of training samples. However, real-world datasets are often imbalanced, where many classes have only a few samples and few classes have a great number of samples. Deep models trained with such unbalanced data usually generalize badly on balanced testing data, especially for rare classes. Improving recognition performance with unbalanced data poses a huge challenge to modern deep learning methods.

To tackle the problem of learning with imbalanced data, early methods mainly focus on re-sampling the training data [1, 2, 14, 33] or re-weighting the loss functions [9, 21, 45] to pay more attention to rare classes. Recently, diverse methods have emerged. For example, Logit compensation methods [3, 31, 38] calibrate distribution between the training data and the test data. Decoupling [23] adopts a two-stage training scheme where the classifier is re-

---

*Indicates equal contribution.
†Jingjing-Chen is the corresponding author.

balanced in the second stage. The work in [44] has multiple distribution-aware experts for responding to samples of different class frequencies. Nevertheless, contrastive learning approaches are less explored before, not until contrastive learning [8, 22, 42] are introduced. We attach great importance to representation learning because it's the most remarkable capability of deep models.

In this paper, we focus on using supervised contrastive learning (SCL) [24] to assist representation learning. Supervised contrastive loss has achieved better performance than supervised cross-entropy loss on large-scale classification problems. The work in [16] then has explained in detail the reason for the excellent performance of SCL on the balanced datasets. Despite the great success, some recent work [8, 22] indicate that high-frequency classes dominate SCL for representing imbalanced data, which results in unsatisfactory performance across all classes. To analyze the optimizing behavior of SCL in learning the representations for long-tailed data, we depict the geometric arrangement of representations of training instances when the lower bound of loss is achieved. Specifically, we decouple the lower bound of loss by deriving two competing dynamics: an attraction term and a repulsion term as in [16]. We reveal that the long-tailed distribution mainly affects the repulsion term. At the minimal supervised contrastive loss, the representations of classes of long-tailed data no longer attain a regular simplex configuration. In other words, when all instances with the same label collapse to points, these points are not equidistant from each other. A regular simplex configuration empirically confers important benefits, such as better generalization performance [32]. Besides, it has been proved to be the target geometric configuration of SCL on balanced data [16], hence forming a regular simplex configuration will benefit recognition on long-tailed data [15].

Inspired by our analysis, we urge the model learning on imbalanced data to form a regular simplex, and propose a balanced contrastive learning (BCL) method (illustrated in Fig. 1). We have two modifications in BCL that distinguish it from SCL. First, class-complement introduces the class-center embeddings, i.e., prototypes, as instances for comparison in every mini-batch. Second, class-averaging has the gradient contributions of all negatives of each class averaged for every mini-batch. Through these two improvements, BCL ensures that the overall lower bound of the loss is a class-independent constant, and alleviates the imbalance problem of SCL when representing long-tailed data. Furthermore, We adopt a cross-entropy loss with logit compensation to obtain a balanced classifier. Logit compensation can effectively alleviate overlooking tail classes in the classifier learning [3, 20, 31, 34]. Overall, we propose a two-branch framework to implement the mentioned techniques, i.e., a contrastive learning branch with BCL and a classification branch with logit compensated cross-entropy.

Our main contributions are as follows:

- We present a theoretical analysis showing that supervised contrastive learning forms an undesired asymmetric geometry configuration for long-tailed data due to the overwhelming numerical dominance of the head classes.

- Motivated by our analysis, we extend supervised contrastive learning to balanced contrastive learning, which overcomes the imbalance problem and remains a regular simplex configuration of long-tailed data.

- The proposed two-branch framework combines the classification module and the balanced contrastive learning module, achieving competitive results on several popular long-tailed datasets.

2. Related Work

Long-tailed Recognition Early solutions to address the long-tailed problem comprise two main ideas: re-sampling and re-weighting. Re-sampling methods undersample [1, 14] high-frequency classes or oversample [1, 2, 33] low-frequency classes. Re-weighting methods [9, 21, 45] assign different losses to different training samples for each class or each example. Both BBN [51] and Decoupling [23] indicate that the re-balancing method is detrimental to representation learning. BBN dynamically adjusts the weights between features from the instance-balanced sampling branch and the reversed sampling branch. While Decoupling proposes a two-stage learning strategy that firstly obtains a good feature extractor and secondly fixes the feature extractor and fine-tunes the classifier. Recently proposed logit compensation methods [3, 23, 31, 38] learn relatively larger margins between different classes based on the prior of class frequencies. For example, logit adjustment [31] derives the general form of the compensation value based on the optimal Bayesian classifier. Our proposed framework simultaneously improves the representation learning with BCL and strengthens the classifier learning with logit compensation in an end-to-end manner.

Supervised Contrastive Learning Contrastive learning (CL) trains the model in a pairwise way by aggregating semantically similar samples while excluding semantically dissimilar ones, which has been employed for feature representation learning in varies tasks [17, 43, 46]. SimCLR [5] and MoCo [18] are two typical types of self-supervised contrastive learning. SCL [24] leverages label information for fully-supervised representation learning, leading to state-of-the-art performance for image classification.

Contrastive Learning for Long-tailed Recognition Trained on long-tailed data, conventional contrastive learning can pose potential problems. SSP [48] boosts long-tailed learning with self-supervised and semi-supervised

contrastive learning. Hybrid-SC [42] designs a two-branch network, using a supervised contrastive learning branch for learning better representations and a classifier branch for eliminating the bias of the classifier towards head classes. Despite our framework shares a similar two-branch architecture with Hybrid-SC, our framework differs from Hybrid-SC as we introduce BCL loss in the SCL branch for dealing with the domination of majority classes problem. Hybrid-PSC [42] is proposed to overcome the memory bottleneck problem of SCL with a prototype for each class for contrast. PaCo [8] overcomes the performance degradation of SCL by introducing a set of class-wise learnable centers. KCL [22] adopts the two-stage learning paradigm and uses the same number of positives for all classes in every batch. The recently proposed TSC [28] is the most relevant work to ours, which urges the features of classes closer to the target features on the vertices of a regular simplex. Targets in TSC are learned without class semantics, while our BCL uses the class prototypes as extra samples.

3. Method

3.1. Preliminaries

In the image classification task, we aim to learn a complex function φ mapping from an input space $\mathcal{X}$ to the target space $\mathcal{Y} = [K] = \{1, 2, \ldots, K\}$. The function φ is usually implemented as the composition of an encoder $f : \mathcal{X} \to \mathcal{Z} \in \mathbb{R}^h$ and a linear classifier $W : \mathcal{Z} \to \mathcal{Y}$. The final classification accuracy strongly depends on the quality of the representations $\mathcal{Z}$. Therefore, we aim to learn a good encoder f to improve long-tailed learning. Further, we have the following definitions to facilitate later analysis:

Supervised contrastive loss. For an instance x_i of representation z_i in a batch B, supervised contrastive loss has the following expression:

$$\mathcal{L}_i = -\frac{1}{|B_y| - 1} \sum_{p \in B_y \setminus \{i\}} \log \frac{\exp(z_i \cdot z_p / \tau)}{\sum\limits_{k \in B \setminus \{i\}} \exp(z_i \cdot z_k / \tau)}$$

(1)

where B_y is a subset of B that contains all samples of class y, and we further define B_y^C as the complement set of B_y. $|\cdot|$ stands for the number of samples in the set. $\tau > 0$ is a scalar temperature hyper parameter that controls tolerance to similar samples, and a small temperature tends to be less tolerant to similar samples [41]. Note that we omit τ in the following contrastive losses for simplicity.

Similar to [16], we also introduce the class-specific batch-wise loss:

$$\mathcal{L}_{SCL}(Z; Y, B, y) = \begin{cases} \sum_{i \in B_y} \mathcal{L}_i & if \; |B_y| > 1 \\ 0 & else \end{cases}$$

(2)

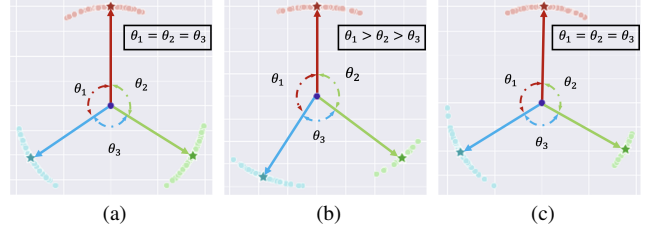**Regular simplex.** A set of points $\zeta_1, \ldots, \zeta_K \in \mathbb{R}^h$ form the

Figure 2. Illustration of the geometry configuration of data points with regard to (a)SCL on balanced data, (b) SCL on long-tailed data, and (c) BCL on long-tailed data in a plane. The ★ represents the class-mean of each class. Different colors represent different classes. SCL enlarges the distances between high-frequency classes and reduces the distances between low-frequency classes as in (b), causing an asymmetrical geometry configuration of long-tailed data.

vertices of a regular simplex inscribed in the hypersphere of radius $\rho > 0$, if and only if the following conditions hold:

(1) $\sum_{i \in [K]} \zeta_i = 0$

(2) $\|\zeta_i\| = \rho$, for $i \in [K]$

(3) $\exists d \in \mathbb{R} : d = \langle \zeta_i, \zeta_j \rangle \quad$ for $1 \leq i < j \leq K$

where h, $K \in \mathbb{N}$ with $K \leq h + 1$, and $\langle \cdot \rangle$ stands for the inner product operation. Regular simplex has a highly symmetric structure that all vertices are equally spaced. Given a balanced dataset, it's worth mentioning that when supervised contrastive loss attain the minimum, representations of each class collapse to the vertices of a regular simplex spontaneously [16, 32]. See the illustration in Fig. 2(a).

3.2. Analysis

Drawbacks of SCL. To clearly show the optimization behaviour of SCL on long-tailed data, we mainly focus on the variation of the geometry configuration formed by the representations of each class. Although representations of each class collapse to the vertices of a regular simplex when supervised contrastive loss attain its minimum on a balanced dataset, SCL forms an asymmetrical configuration on long-tailed data as shown in Fig. 2(b). In the following, we will give an in-depth analysis on the loss function to show why the geometry configuration changes for imbalanced data. In particular, we analysis the lower-bound of the loss. Since directly computing the lower bound on the whole long-tailed dataset is often intractable, we pay attention to the loss of a specific mini-batch instead.

Theorem 1. *Assuming the normalization function is applied for feature embedding, let $Z = (z_1, \ldots, z_N) \in \mathcal{Z}^N$ be an N point configuration with labels $Y = (y_1, \ldots, y_N) \in [K]^N$, where $\mathcal{Z} = \{z \in \mathbb{R}^h : \|z\| = 1\}$. The class-specific batch-*
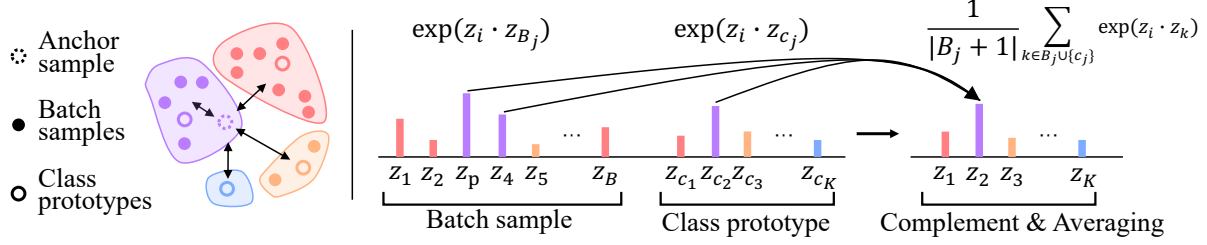
Figure 3. **Left:** Contrasting an anchor sample with others. **Right:** Applying both class-averaging and class-complement for a specific class. We average the similarity between the anchor sample and the batch sample as well as the class prototype. Note that the class colored blue does not appear in the mini-batch, so the similarity between the anchor and its prototype can be directly taken as the result.

wise loss is bounded by

$$\mathcal{L}_{SCL}(Z;Y,B,y) \geq \sum_{i \in B_y} \log((|B_y|-1)+$$

$$|B_y^C| \exp(\underbrace{\frac{1}{|B_y^C|}\sum_{k \in B_y^C} z_i \cdot z_k}_{repulsion\ term} - \underbrace{\frac{1}{|B_y|-1}\sum_{j \in B_y \setminus \{i\}} z_i \cdot z_j}_{attraction\ term})) \quad (3)$$

Proof. See Lemma S1 in [16].

The above lower bound of SCL loss is derived by [16], which consists of a repulsion term and an attraction term. The attraction term leads to variability collapse [32] as training progresses, and all the within-class representations collapse to their class means in the end. The attraction term only relates to samples within a specific class. It means that whether the dataset is balanced or not, samples within the same class should be as close as possible.

The attraction term leads to intra-class feature collapse regardless of the class frequency. While the repulsion term affects inter-class uniformity and is dominated by classes with higher frequency, thereby features in SCL are less separable. We indicate that data imbalance mainly affects the repulsion term. Obviously, the repulsion term is strongly related to the data distribution of the classes appeared within a mini-batch. When the dataset is long-tailed, almost every mini-batch we sampled is long-tailed. This leads to the dominance of the head classes in the repulsion term and makes each sample farther away from the heads. However, due to the number of samples in each class being different, the distance between the head classes will be larger compared to the others. Additionally, for each sample, the gradients from negative head classes will be much larger than negative tail classes. This unavoidably causes the loss to focus more on optimizing the head classes and leads to an asymmetrical geometry as shown in Fig. 2(b).

Solution. We modify the supervised contrastive loss with two straightforward ideas, i.e., *class-averaging* and *class-complement*. The modified loss will lead to a regular simplex configuration of long-tailed data as we will show below. To avoid excessive concentration on head classes, an intuitive approach is to equilibrate the gradients con-

tributed by different negative classes. We call this operation as *class-averaging*. The gradients from the negative samples of head classes are reduced. Below, we give the lower bound after performing class-averaging.

Theorem 2. *Let Z, Y be defined as in Theorem 1, $\mathcal{Y}_B \subseteq [K]$ denotes the set of classes that appear within batch. the class-specific batch-wise loss after performing class-averaging is bounded by*

$$\mathcal{L}_{BCL}(Z;Y,B,y) \geq \sum_{i \in B_y} \log(1+ (|\mathcal{Y}_B|-1) \times \exp($$

$$\underbrace{\frac{1}{|\mathcal{Y}_B|-1}\sum_{q \in \mathcal{Y}_B \setminus \{y\}} \frac{1}{|B_q|}\sum_{k \in B_q} z_i \cdot z_k}_{repulsion\ term} - \underbrace{\frac{1}{|B_y|-1}\sum_{j \in B_y \setminus \{i\}} z_i \cdot z_j}_{attraction\ term}))$$

$$(4)$$

Proof. See the Supplementary Material.

Consequently, head classes no longer dominate the repulsion term. Since each class is not sampled with equal probability, this may still lead to an unstable optimization and fail to form a regular simplex. To address this problem, we make all classes appear in every mini-batch and name this operation as *class-complement*. Below, we give the overall lower bound after performing class-complement.

Theorem 3. *Let Z, Y be defined as in Theorem 1, if we have $\mathcal{Y}_B = \mathcal{Y}$ for every B, the overall loss is given by*

$$\mathcal{L}_{BCL}(Z;Y) \geq |\mathcal{D}| \log(1+(K-1)\exp(-\frac{K}{K-1})) \quad (5)$$

where $\mathcal{D}$ represents the dataset. Here, the normalization term is neglected. Recall that Z is an N point configuration with labels Y, the equality of Eq. 5 is attained if and only if the following conditions hold. There are $\zeta_1, \ldots, \zeta_K \in \mathbb{R}^h$ such that:

(1) $\forall n \in [N] : z_n = \zeta_{y_n}$

(2) $\zeta_1, \ldots, \zeta_K\ form\ a\ regular\ simplex$

Proof. See the Supplementary Material.

Note that condition (1) implies variability collapse, and condition (2) demonstrates the regular simplex structure. When balanced contrastive loss attains its lower bound,
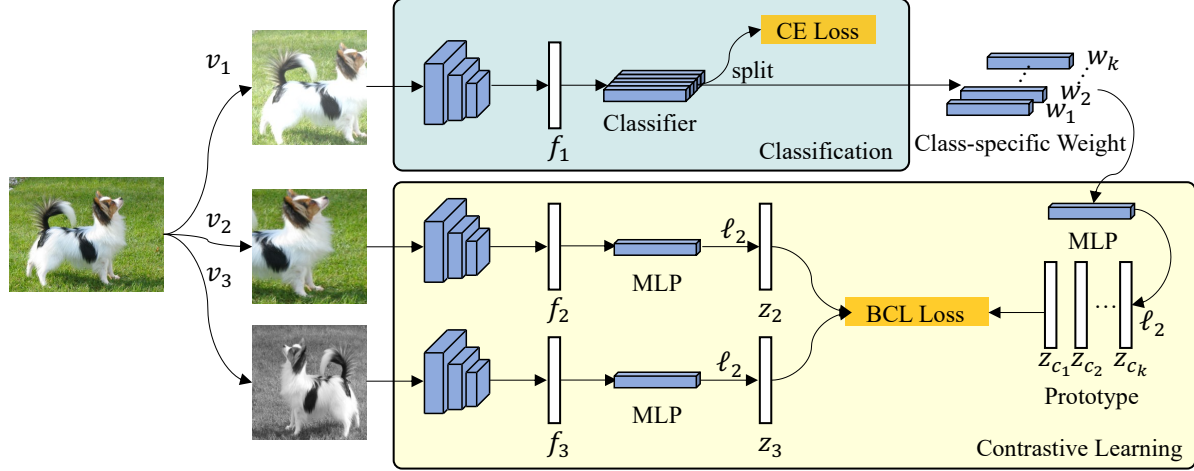
Figure 4. Overview of the proposed framework. The framework consists of a classification branch and a balanced contrastive learning branch. v_2 and v_3 adopt the same augmentation method different from v_1. The backbone is shared between two branches. The classifier weights are separately transformed by a MLP to be used as prototypes. All representations are ℓ_2-normalized for balanced contrastive loss.

each negative class contributes to the gradient equally. Additionally, BCL ensures that the loss of each sample is consistent and class-independent when attaining its lower bound, which implies that the learning will be less biased towards the head classes (Fig. 2(c)).

3.3. Balanced Contrastive Learning

Class-averaging The key idea is to average the instances of each class in a mini-batch so that each class has an approximate contribution for optimizing. Intuitively, it reduces the proportion of head classes in the denominator and emphasizes the importance of tails. In Section 3.2, we take the loss form as $\mathcal{L}_1$ for analysis. However, there are other ways to implement class-averaging. Here, we give three loss forms as follows:

$$\mathcal{L}_1 = -\frac{1}{|B_y|-1} \sum_{p \in B_y \setminus \{i\}} \log \frac{\exp(z_i \cdot z_p)}{\sum_{j \in \mathcal{Y}_B} \frac{1}{|B_j|} \sum_{k \in B_j} \exp(z_i \cdot z_k)} \quad (6)$$

$$\mathcal{L}_2 = -\frac{1}{|B_y|-1} \sum_{p \in B_y \setminus \{i\}} \log \frac{\exp(z_i \cdot z_p)}{\sum_{j \in \mathcal{Y}_B} \exp(\frac{1}{|B_j|} \sum_{k \in B_j} z_i \cdot z_k)} \quad (7)$$

$$\mathcal{L}_3 = -\log \frac{\exp(z_i \cdot z_{c_y})}{\sum_{j \in \mathcal{Y}} \exp(z_i \cdot z_{c_j})} \quad (8)$$

where the $|B_j|$ term will minus one when the positive class is averaged. The only difference between our $\mathcal{L}_1$ and $\mathcal{L}_2$ is that the averaging operation takes place in different positions. $\mathcal{L}_1$ performs averaging outside the exponential function, while $\mathcal{L}_2$ has averaging inside the exponential function. Since -log and exp are convex functions, which implies that $\mathcal{L}_1 \geq \mathcal{L}_2$ by Jensen's inequality. $\mathcal{L}_3$ take the form proposed in other prototype-based contrastive learning

methods [26, 36, 42], where each sample is pulled towards its class prototype and pushed away from others. Note that $z \in \mathcal{Z} \subseteq \mathbb{R}^h$ is ℓ_2-normalized for the inner product, so that $\|z\|_2 = 1$. A comparison of their performance is in the experiment section. Below, we choose $\mathcal{L}_1$ for optimization.

Class-complement To have all classes appear in every mini-batch, we introduce class-center representations, i.e., prototypes for balanced contrastive learning. Now we have the formulation of balanced contrastive loss as follows:

$$\mathcal{L}_{BCL} = -\frac{1}{|B_y|} \times$$
$$\sum_{p \in \{B_y \setminus \{i\}\} \cup \{c_y\}} \log \frac{\exp(z_i \cdot z_p)}{\sum_{j \in \mathcal{Y}} \frac{1}{|B_j|+1} \sum_{k \in B_j \cup \{c_j\}} \exp(z_i \cdot z_k)} \quad (9)$$

where c_j is the index of the prototype. In practice, we perform a nonlinear mapping of the classifier weights and regard the output as the prototype of each class.

By applying both class-averaging and class-complement, the lower bound is a class-independent constant, avoiding the model's preference for head classes. Note that in practice, we perform class-complement before class-averaging, as formulated in Eq. 9 and illustrated in Fig. 3.

Framework The overview of the proposed framework is shown in Fig. 4. It consists of two main components: a classification branch and a contrastive learning branch. Both branches are trained simultaneously and share the same feature extractor. BCL is a unified end-to-end model, which is different from conventional contrastive learning methods that follow a two-stage training strategy. We have different augmentation methods for the two branches. Three different views are generated in total, where v_1 is the view used for the classification task, v_2 and v_3 are the pairwise views for

the contrastive learning task. Following the work in [5,24], we utilize a symmetric architecture for the contrastive learning branch. We use a MLP with one hidden layer to obtain the representation z_i for contrastive learning where $z_i = W_1\sigma(W_2 f_i)$ and σ is a ReLU function. Instead of using mean embeddings [36] or learnable parameters [8,42] as prototypes, we are motivated by that the weights of the linear classifier are co-linear with these simplex vertices to which the classes collapse [16,32]. Specifically, we have class-specific weights $w_1, w_2, \ldots, w_K$ after a non-linear transformation MLP as prototypes $z_{c_1}, z_{c_2}, \ldots, z_{c_K}$. Ablations of using different configurations of views and different forms of prototypes are given in the Supplementary Material. All representations for contrastive learning is ℓ_2 normalized to ensure the feature space is a unit hypersphere.

Optimization with Logit Compensation For long-tailed learning tasks, due to the imbalance of data, the output logit of the last classification layer usually exists bias. Logit compensation aims to eliminate the bias caused by the imbalance of data and learn the rectification of the boundary [3,31,38]. The compensation can be applied during either training or testing. Previous work [3,9,20,31,38] illustrates the desirability of logit compensation in long-tailed visual tasks and it can be summarized as the following form

$$\mathcal{L}_{LC}(y, \varphi(x)) = -\alpha_y \log \frac{\exp(\varphi_y(x) + \delta_y)}{\sum\limits_{y' \in [\mathcal{Y}]} \exp(\varphi_{y'}(x) + \delta_{y'})} \quad (10)$$

Here, α_y is the factor that controls the importance of class y, δ_y is the compensation for class y and its value is related to class-frequency. We define $\alpha_y = 1$, $\delta_y = \log \mathbb{P}_y$ as in [31] and perform logit compensation along with training, where $\mathbb{P}_y$ denotes the class prior of label y.

Finally, we have the following loss for training:

$$\mathcal{L} = \lambda \mathcal{L}_{LC} + \mu \mathcal{L}_{BCL} \quad (11)$$

where λ and μ are hyperparameters that control the impact of $\mathcal{L}_{LC}$ and $\mathcal{L}_{BCL}$, respectively. In addition, the contrastive branch only intends for the backbone to learn the desired feature embeddings.

4. Experiment

4.1. Dataset

Long-Tailed CIFAR-10 and CIFAR-100 CIFAR-10-LT and CIFAR-100-LT are the subsets of CIFAR-10 and CIFAR-100, respectively. Both CIFAR-10 and CIFAR-100 contain 50,000 images for training and 10,000 images for the validation of size 32×32 with 10 and 100 classes respectively. Following [3,9,51], we use the same long-tailed version for a fair comparison. The imbalanced factor β is defined by $\beta = N_{max}/N_{min}$, and this reflects the degree

Methods	Top-1 Acc.
$\mathcal{L}_1$	51.9
$\mathcal{L}_2$	50.2
$\mathcal{L}_3$	51.0

Table 1. Ablation study for different class-averaging methods. All models run for 200 epochs with the same training scheme.

LC	SC	Complement	Averaging	Top-1 Acc.
✓	✗	✗	✗	50.8
✓	✓	✗	✗	52.4
✓	✓	✓	✗	52.3
✓	✓	✗	✓	52.0
✓	✓	✓	✓	53.9

Table 2. Ablation study for the primary components of BCL. LC and SC denote logit compensation and supervised contrastive loss. Complement and averaging stand for class-complement and class-averaging, respectively. All models run for 400 epochs.

of imbalance in the data. The imbalance factors used in the experiment are set to 100, 50, and 10.

ImageNet-LT ImageNet-LT is proposed in [30], which is a long-tailed version of vanilla ImageNet by sampling a subset following the Pareto distribution with power value $\alpha = 0.6$. It consists of 115.8K images of 1000 classes in total with 1280 to 5 images per class.

iNaturalist 2018 iNaturalist 2018 [40] is a large-scale dataset containing 437.5K images from 8,142 classes. It is long-tailed by nature with an extremely imbalanced distribution. In addition to long-tailed recognition, this dataset is also used for evaluating the fine-grained classification task.

4.2. Implementation details

For both CIFAR-10-LT and CIFAR-100-LT, we use the ResNet-32 as the backbone. Same to [8], we use AutoAugment [6] and Cutout [13] as data augmentation strategies for the classification branch and SimAugment [5] for the contrastive learning branch. To control the impact of $\mathcal{L}_{LC}$ and $\mathcal{L}_{BCL}$, λ is set to 2.0, μ is 0.6, and the temperature τ is set to 0.1. We set the batch size as 256 and the weight decay as 5e−4. The dimension of the hidden layer and the output layer of MLP are set to 512 and 128, respectively. We run BCL for 200 epochs with the learning rate warms up to 0.15 within the first 5 epochs and decays at epoch 160 and 180 with a step size of 0.1. Following [8], we also run the model for 400 epochs, where the learning rate warms up to 0.15 within the first 10 epochs and decays at epoch 360 and 380 with a step size of 0.1. We train the above models with one Nvidia GeForce 1080Ti GPU.

For ImageNet-LT, we use ResNet-50 [19] and ResNeXt-50-32x4d [47] as our backbone. We run BCL for 90 epochs with an initial learning rate of 0.1 and the weight decay is 5e−4. For iNaturalist, we use ResNet-50 as our backbone

Method	CIFAR-100-LT			CIFAR-10-LT		
Imbalance Factor	100	50	10	100	50	10
SSP [48]	43.43	47.11	58.91	77.83	82.13	88.53
Focal loss[†] [29]	38.41	44.32	55.78	70.38	76.72	86.66
CB-Focal [10]	39.60	45.17	57.99	74.57	79.27	87.10
BBN [51]	42.56	47.02	59.12	79.82	81.18	88.32
Casual model [39]	44.10	50.30	59.60	80.60	83.60	88.50
LDAM-DRW [3]	42.04	46.62	58.71	77.03	81.03	88.16
ResLT [7]	48.21	52.71	62.01	82.40	85.17	89.70
Hybrid-SC [42]	46.72	51.87	63.05	81.40	85.36	**91.12**
MetaSAug-LDAM [27]	48.01	52.27	61.28	80.66	84.34	89.68
BCL(ours)	**51.93**	**56.59**	**64.87**	**84.32**	**87.24**	**91.12**

Table 3. Top-1 accuracy of ResNet-32 on CIFAR-100-LT and CIFAR-10-LT. The best results are marked in bold. † denotes results borrowed from [42]. We report the results of 200 epochs.

Methods	Many	Medium	Few	All
200 epochs				
τ-norm[†] [23]	61.4	42.5	15.7	41.4
Hybrid-SC [42]	-	-	-	46.7
MetaSAug-LDAM [27]	-	-	-	48.0
DRO-LT [36]	64.7	50.0	23.8	47.3
RIDE(3 experts) [44]	**68.1**	49.2	23.9	48.0
BCL(Ours)	67.2	**53.1**	**32.9**	**51.9**
400 epochs				
Balanced Softmax[‡] [34]	-	-	-	50.8
PaCo [8]	-	-	-	52.0
BCL(Ours)	**69.7**	**53.8**	**35.5**	**53.9**

Table 4. Top-1 accuracy of ResNet-32 on CIFAR-100-LT with an imbalance factor of 100. We report the results of 200 epochs and 400 epochs. † and ‡ denote results borrowed from [36] and [8].

and run BCL for 100 epochs using an initial learning rate of 0.2 and the weight decay is 1e−4. For both ImageNet-LT and iNaturalist 2018, we use cosine scheduling for learning rate, λ is set to 1.0 and μ is set to 0.35. The batch size is set to 256. We use the RandAug augmentation strategy for the classification branch, and SimAug for the contrastive learning branch. The performances of different augmentation strategies are in the Supplementary Material. To reduce memory consumption, the dimension of the output layer of MLP is set to 1024 for both datasets. We use the cosine classifier. All models are trained using SGD optimizer with a momentum set to 0.9. For a fair comparison, we reproduce PaCo of ResNext-50 on ImageNet-LT for 180 epochs.

4.3. Ablation study

We perform several ablation studies to characterize the proposed BCL method. All experiments are performed on CIFAR-100 with an imbalance factor of 100. First, we compare the performances of different class-averaging implementations (i.e., $\mathcal{L}_1$, $\mathcal{L}_2$ and $\mathcal{L}_3$) mentioned in Section 3.3.

The main difference between $\mathcal{L}_1$ and $\mathcal{L}_2$ is the order in which the averaging operations are implemented. For $\mathcal{L}_3$, we use the prototype implemented in our work instead of the average of all embeddings of the same class. As shown in Table 1, $\mathcal{L}_1$ achieves the best performance, which is consistent with our previous analysis. Surprisingly, $\mathcal{L}_3$ achieves better performance than $\mathcal{L}_2$, which may be attributed to the well-represented characteristics of prototypes.

To demonstrate the superiority of the balanced contrastive loss, we compare the performance of the primary components of the loss in Table 2. We use the cross-entropy loss with logit compensation (LC) as the vanilla baseline. SC denotes a baseline that adds the contrastive learning branch with conventional supervised contrastive loss. Class-complement and class-averaging are the main techniques of the proposed balanced contrastive loss. We show that using either class-complement or class-averaging alone cannot improve the overall accuracy. In contrast, a significant performance boost can be obtained when both of them are applied, which indicates both components are indispensable components to achieve stronger performance.

4.4. Main results

Long-tailed CIFAR The comparison results between the proposed BCL and other existing methods on long-tailed CIFAR are shown in Table 3. As can be seen from the table, BCL consistently outperforms the other methods. Furthermore, BCL achieves better performance on long-tailed CIFAR datasets with large imbalance factors. We note that the accuracy gap between BCL and Hybrid-SC decreases as the degree of data imbalance reduces. This result is mainly attributed to the fact that the conventional supervised contrastive loss leads to more serious bias in representation learning when unbalance problem is more severe.

Further, we report the accuracy on three groups of classes, including Many-shot(>100 images), Medium-shot(2∼100 images), Few-shot(<20 images), on CIFAR-100-LT with imbalance factor as 100. Additionally, for a

Methods	ImageNet-LT	iNaturalist 2018
τ-norm [23]	46.7	65.6
cRT [23]	49.6	65.2
LWS [23]	49.9	65.9
BBN [51]	-	66.3
Hybrid-SC [42]	-	66.7
Hybrid-PSC [42]	-	68.1
SSP [48]	51.3	68.1
KCL [22]	51.5	68.6
DisAlign [50]	52.9	69.5
RIDE(2 experts) [44]	54.4	71.4
BCL(Ours)	**56.0**	**71.8**

Table 5. Top-1 accuracy of ResNet-50 on ImageNet-LT and iNaturalist 2018. All methods are trained for up to 100 epochs.

Methods	Many	Medium	Few	All
90 epochs				
Focal Loss[†] [29]	64.3	37.1	8.2	43.7
τ-norm [23]	59.1	46.9	30.7	49.4
Balanced Softmax[†] [34]	62.2	48.8	29.8	51.4
LWS [23]	60.2	47.2	30.3	49.9
LADE [20]	62.3	49.3	31.2	51.9
Casual model [39]	62.7	48.8	31.6	51.8
DisAlign [50]	62.7	52.1	31.4	53.4
RIDE(2 experts) [44]	-	-	-	55.9
BCL(Ours)	**67.2**	**53.9**	**36.5**	**56.7**
180 epochs				
LADE [20]	65.1	48.9	33.4	53.0
Balanced Softmax[†]	65.8	53.2	34.1	55.4
PaCo[†] [8]	64.4	**55.7**	33.7	56.0
BCL(Ours)	**67.9**	54.2	**36.6**	**57.1**

Table 6. Top-1 accuracy of ResNext-50 on ImageNet-LT. We report the results of 90 epochs and 180 epochs. † denotes results reproduced using the code released by authors, both with RandAug.

fair comparison with [8], we report the accuracy on 200 and 400 epochs. DRO-LT [36] is also a contrastive learning method which extends prototypical contrastive learning by introducing distributional robustness. As shown in Table 4, BCL surpasses DRO-LT by 4.6 and PaCo by 1.9 with 200 training epochs and 400 training epochs, respectively. It's worthwhile to mention that different to most of the previous methods which compromise the performance of the head classes, our BCL further improves the performance of the head classes while simultaneously improving the tails.

ImageNet-LT Table 5 and Table 6 list the results on ImageNet-LT. We report the overall Top-1 accuracy as well as the Top-1 accuracy on Many-shot, Medium-shot, and Few-shot groups. Compared with Balanced Softmax [34], which proposes a logit compensation by adjusting the predictions according to the class frequency. BCL significantly outperforms Balanced Softmax on all groups, confirming the well-learned representations can boost the overall performance. LWS [23], τ-norm [23], and DisAlign [50] adopt the two-stage learning strategy. These methods focus on fine-tuning the classifier in the second stage, while they neglect the bias implied in the representation learning stage. PaCo [8] uses a set of parametric centers in supervised contrastive learning. These centers are assigned with a much greater weight, which can be regarded as the weights of a classifier. However, the prototypes used in BCL complement the samples of each class to make sure that all classes appear in every mini-batch. Compared with PaCo, BCL achieves a better overall accuracy of 57.1% with remarkable accuracy improvements on the head and few classes.

iNaturalist 2018 Table 5 shows the experimental results on iNaturalist 2018. Since BCL is a contrastive learning method, it benefits more from a longer training time. However, for a fair comparison, we report the results of various models trained for up to 100 epochs. Hybrid-SC [42] and Hybrid-PSC [42] are contrastive learning approaches, and their performances are inferior to BCL due to the potential

bias produced in representation learning. RIDE [44] trains model from an ensemble-based manner. RIDE obtains high performance by routing diverse experts. However, it's different from our motivation of boosting long-tailed learning via training a sufficient encoder and a balanced feature space. Compared with the ensemble-based model RIDE, BCL consistently demonstrates better recognition performance and achieves the overall accuracy at 71.8%.

5. Conclusion

In this work, we investigated the problem of long-tailed recognition from the perspective of representation learning. We provided in-depth analysis to demonstrate that existing supervised contrastive learning forms an undesired asymmetric geometry configuration for long-tailed data. To tackle the imbalanced data representation learning problem, we developed a balanced contrastive loss, so that all classes are optimized for a regular simplex configuration that yields a balanced feature space. In addition to BCL, we employ a classification branch with logit compensation to tackle the biased classifier. Overall we have presented a framework unifying both branches. We conducted extensive experiments on the long-tailed benchmarks of long-tailed CIFAR, ImageNet-LT, and iNaturalist 2018. The experimental results adequately demonstrate the superiority of BCL compared to existing long-tailed learning methods.

6. Acknowledgements

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. 1, 2

[2] Jonathon Byrd and Zachary Lipton. What is the effect of importance weighting in deep learning? In *International Conference on Machine Learning*, pages 872–881. PMLR, 2019. 1, 2

[3] Kaidi Cao, Colin Wei, Adrien Gaidon, Nikos Arechiga, and Tengyu Ma. Learning imbalanced datasets with label-distribution-aware margin loss. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, pages 1567–1578, 2019. 1, 2, 6, 7

[4] Jingjing Chen, Bin Zhu, Chong-Wah Ngo, Tat-Seng Chua, and Yu-Gang Jiang. A study of multi-task and region-wise deep learning for food ingredient recognition. *IEEE Transactions on Image Processing*, 30:1514–1526, 2020. 1

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

[6] Ekin D Cubuk, Barret Zoph, Dandelion Mane, Vijay Vasudevan, and Quoc V Le. Autoaugment: Learning augmentation strategies from data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 113–123, 2019. 6

[7] Jiequan Cui, Shu Liu, Zhuotao Tian, Zhisheng Zhong, and Jiaya Jia. Reslt: Residual learning for long-tailed recognition. *arXiv preprint arXiv:2101.10633*, 2021. 7

[8] Jiequan Cui, Zhisheng Zhong, Shu Liu, Bei Yu, and Jiaya Jia. Parametric contrastive learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 715–724, 2021. 2, 3, 6, 7, 8

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

[10] Yin Cui, Menglin Jia, Tsung-Yi Lin, Yang Song, and Serge Belongie. Class-balanced loss based on effective number of samples. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9268–9277, 2019. 7

[11] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 1

[12] Lixi Deng, Jingjing Chen, Qianru Sun, Xiangnan He, Sheng Tang, Zhaoyan Ming, Yongdong Zhang, and Tat Seng Chua. Mixed-dish recognition with contextual relation networks. In *Proceedings of the 27th ACM International Conference on Multimedia*, pages 112–120, 2019. 1

[13] Terrance DeVries and Graham W Taylor. Improved regularization of convolutional neural networks with cutout. *arXiv preprint arXiv:1708.04552*, 2017. 6

[14] Chris Drumnond and Robert C Holte. Class imbalance and cost sensitivity: Why undersampling beats oversampling. In *ICML-KDD 2003 Workshop: Learning from Imbalanced Datasets*, volume 3, 2003. 1, 2

[15] Cong Fang, Hangfeng He, Qi Long, and Weijie J Su. Exploring deep neural networks via layer-peeled model: Minority collapse in imbalanced training. *Proceedings of the National Academy of Sciences*, 118(43), 2021. 2

[16] Florian Graf, Christoph Hofer, Marc Niethammer, and Roland Kwitt. Dissecting supervised constrastive learning. In *International Conference on Machine Learning*, pages 3821–3830. PMLR, 2021. 2, 3, 4, 6

[17] Tengda Han, Weidi Xie, and Andrew Zisserman. Self-supervised co-training for video representation learning. *Advances in Neural Information Processing Systems*, 33:5679–5690, 2020. 2

[18] 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*, pages 9729–9738, 2020. 2

[19] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 6

[20] Youngkyu Hong, Seungju Han, Kwanghee Choi, Seokjun Seo, Beomsu Kim, and Buru Chang. Disentangling label distribution for long-tailed visual recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6626–6636, 2021. 2, 6, 8

[21] Chen Huang, Yining Li, Chen Change Loy, and Xiaoou Tang. Learning deep representation for imbalanced classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5375–5384, 2016. 1, 2

[22] Bingyi Kang, Yu Li, Sa Xie, Zehuan Yuan, and Jiashi Feng. Exploring balanced feature spaces for representation learning. In *International Conference on Learning Representations*, 2020. 2, 3, 8

[23] Bingyi Kang, Saining Xie, Marcus Rohrbach, Zhicheng Yan, Albert Gordo, Jiashi Feng, and Yannis Kalantidis. Decoupling representation and classifier for long-tailed recognition. In *International Conference on Learning Representations*, 2019. 1, 2, 7, 8

[24] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. *Advances in Neural Information Processing Systems*, 33, 2020. 2, 6

[25] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25:1097–1105, 2012. 1

[26] Junnan Li, Pan Zhou, Caiming Xiong, and Steven C.H. Hoi. Prototypical contrastive learning of unsupervised representations. In *ICLR*, 2021. 5

[27] Shuang Li, Kaixiong Gong, Chi Harold Liu, Yulin Wang, Feng Qiao, and Xinjing Cheng. Metasaug: Meta semantic augmentation for long-tailed visual recognition. In *Proceed-*

ings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 5212–5221, 2021. 7

[28] Tianhong Li, Peng Cao, Yuan Yuan, Lijie Fan, Yuzhe Yang, Rogerio Feris, Piotr Indyk, and Dina Katabi. Targeted supervised contrastive learning for long-tailed recognition. arXiv preprint arXiv:2111.13998, 2021. 3

[29] 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, pages 2980–2988, 2017. 7, 8

[30] Ziwei Liu, Zhongqi Miao, Xiaohang Zhan, Jiayun Wang, Boqing Gong, and Stella X Yu. Large-scale long-tailed recognition in an open world. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 2537–2546, 2019. 6

[31] Aditya Krishna Menon, Sadeep Jayasumana, Ankit Singh Rawat, Himanshu Jain, Andreas Veit, and Sanjiv Kumar. Long-tail learning via logit adjustment. In International Conference on Learning Representations, 2020. 1, 2, 6

[32] Vardan Papyan, XY Han, and David L Donoho. Prevalence of neural collapse during the terminal phase of deep learning training. Proceedings of the National Academy of Sciences, 117(40):24652–24663, 2020. 2, 3, 4, 6

[33] Samira Pouyanfar, Yudong Tao, Anup Mohan, Haiman Tian, Ahmed S Kaseb, Kent Gauen, Ryan Dailey, Sarah Aghajanzadeh, Yung-Hsiang Lu, Shu-Ching Chen, et al. Dynamic sampling in convolutional neural networks for imbalanced data classification. In 2018 IEEE conference on multimedia information processing and retrieval (MIPR), pages 112–117. IEEE, 2018. 1, 2

[34] Jiawei Ren, Cunjun Yu, Shunan Sheng, Xiao Ma, Haiyu Zhao, Shuai Yi, and Hongsheng Li. Balanced meta-softmax for long-tailed visual recognition. In Proceedings of Neural Information Processing Systems(NeurIPS), Dec 2020. 2, 7, 8

[35] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. Advances in neural information processing systems, 28:91–99, 2015. 1

[36] Dvir Samuel and Gal Chechik. Distributional robustness loss for long-tail learning. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), pages 9495–9504, October 2021. 5, 6, 7, 8

[37] Zixuan Su, Xindi Shang, Jingjing Chen, Yu-Gang Jiang, Zhiyong Qiu, and Tat-Seng Chua. Video relation detection via multiple hypothesis association. In Proceedings of the 28th ACM International Conference on Multimedia, pages 3127–3135, 2020. 1

[38] 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, pages 11662–11671, 2020. 1, 2, 6

[39] Kaihua Tang, Jianqiang Huang, and Hanwang Zhang. Long-tailed classification by keeping the good and removing the bad momentum causal effect. Advances in Neural Information Processing Systems, 33, 2020. 7, 8

[40] 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, pages 8769–8778, 2018. 6

[41] Feng Wang and Huaping Liu. Understanding the behaviour of contrastive loss. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 2495–2504, 2021. 3

[42] Peng Wang, Kai Han, Xiu-Shen Wei, Lei Zhang, and Lei Wang. Contrastive learning based hybrid networks for long-tailed image classification. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 943–952, 2021. 2, 3, 5, 6, 7, 8

[43] Rui Wang, Zuxuan Wu, Zejia Weng, Jingjing Chen, Guo-Jun Qi, and Yu-Gang Jiang. Cross-domain contrastive learning for unsupervised domain adaptation. IEEE Transactions on Multimedia, 2022. 2

[44] 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, 2020. 2, 7, 8

[45] Yu-Xiong Wang, Deva Ramanan, and Martial Hebert. Learning to model the tail. In Proceedings of the 31st International Conference on Neural Information Processing Systems, pages 7032–7042, 2017. 1, 2

[46] Zheng Wang, Jingjing Chen, and Yu-Gang Jiang. Visual co-occurrence alignment learning for weakly-supervised video moment retrieval. In Proceedings of the 29th ACM International Conference on Multimedia, pages 1459–1468, 2021. 2

[47] 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, pages 1492–1500, 2017. 6

[48] Yuzhe Yang and Zhi Xu. Rethinking the value of labels for improving class-imbalanced learning. In Conference on Neural Information Processing Systems (NeurIPS), 2020. 2, 7, 8

[49] Hao Zhang, Yanbin Hao, and Chong-Wah Ngo. Token shift transformer for video classification. In Proceedings of the 29th ACM International Conference on Multimedia, pages 917–925, 2021. 1

[50] Songyang Zhang, Zeming Li, Shipeng Yan, Xuming He, and Jian Sun. Distribution alignment: A unified framework for long-tail visual recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 2361–2370, 2021. 8

[51] 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, pages 9719–9728, 2020. 2, 6, 7, 8