Neurocomputing 527 (2023) 36–46

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

Balanced knowledge distillation for long-tailed learning

Shaoyu Zhang [a,b], Chen Chen [a,b,*], Xiyuan Hu [c], Silong Peng [a,b,d]

[a] Institute of Automation, Chinese Academy of Sciences, Beijing, China
[b] University of Chinese Academy of Sciences, Beijing, China
[c] Nanjing University of Science and Technology, Nanjing, China
[d] Beijing Visystem Co. Ltd, Beijing, China

ARTICLE INFO

Article history:
Received 22 May 2022
Revised 1 December 2022
Accepted 9 January 2023
Available online 13 January 2023
Communicated by Zidong Wang

Keywords:
Long-tailed learning
Knowledge distillation
Vision and text classification

ABSTRACT

Deep models trained on long-tailed datasets exhibit unsatisfactory performance on tail classes. Existing methods usually modify the classification loss to increase the learning focus on tail classes, which unexpectedly sacrifice the performance on head classes. In fact, this scheme leads to a contradiction between the two goals of long-tailed learning, i.e., learning generalizable representations and facilitating learning for tail classes. In this work, we explore knowledge distillation in long-tailed scenarios and propose a novel distillation framework, named *Balanced Knowledge Distillation (BKD)*, to disentangle the contradiction between the two goals and achieve both simultaneously. Specifically, given a teacher model, we train the student model by minimizing the combination of an instance-balanced classification loss and a class-balanced distillation loss. The former benefits from the sample diversity and learns generalizable representation, while the latter considers the class priors and facilitates learning for tail classes. We conduct extensive experiments on several long-tailed benchmark datasets and demonstrate that the proposed BKD is an effective knowledge distillation framework in long-tailed scenarios, as well as a competitive method for long-tailed learning. Our source code is available: https://github.com/EricZsy/BalancedKnowledgeDistillation.

1. Introduction

Recent advances in deep neural networks are mainly driven by the use of large-scale datasets, such as ImageNet ILSVRC 2012 [1,2] and Microsoft COCO [3]. Such datasets are often carefully collected, with roughly balanced quantities in each category. However, in practical scenarios, data tends to exhibit long-tailed distribution [4,5], wherein a few classes (head classes) have a significantly larger number of instances than other classes (tail classes). This uneven distribution affects both convergence during the training phase and generalization on the test set [6]. When dealing with such imbalanced data, deep models tend to bias towards head classes, resulting in performance drop on tail classes.

The goals of long-tailed learning are twofold: learning generalizable representations and facilitating learning for tail classes. In the literature, one of the most common practices to facilitate learning for tail classes is to re-balance the class distribution, either by re-sampling the examples [7–9] or re-weighting the classification loss [10–12]. Although effective for tail classes, these methods could damage the overall representation learning [13,14] and sacrifice performance on head classes, as shown in Fig. 1. Therefore, such re-balancing strategies naturally bring along with a contradiction between learning generalizable representations and facilitating learning for tail classes. In Section 3.1, we take the most representative re-weighting method [12] for example and explain where the contradiction comes from by analyzing the effect of gradient. We reveal that this re-weighting method brings overwhelming discouraging gradients to head classes, resulting in suboptimal representations.

In this work, we disentangle the contradiction between the two goals from the perspective of teacher-student learning [15]. Specifically, if trained on imbalanced data, the student tries the best to learn high-quality representations while the teacher should facilitate learning with focus on tail classes. Motivated by this, we explore knowledge distillation in long-tailed scenarios and propose a simple yet effective distillation framework, named *balanced knowledge distillation (BKD)*, to alleviate the long-tailed problem. We first train a teacher model by minimizing vanilla cross-entropy loss. Then the student model with the same size as the teacher model is trained by minimizing the combination of an instance-balanced classification loss and a class-balanced

* Corresponding author at: Institute of Automation, Chinese Academy of Sciences, Beijing, China.
E-mail address: chen.chen@ia.ac.cn (C. Chen).

https://doi.org/10.1016/j.neucom.2023.01.063
0925-2312/

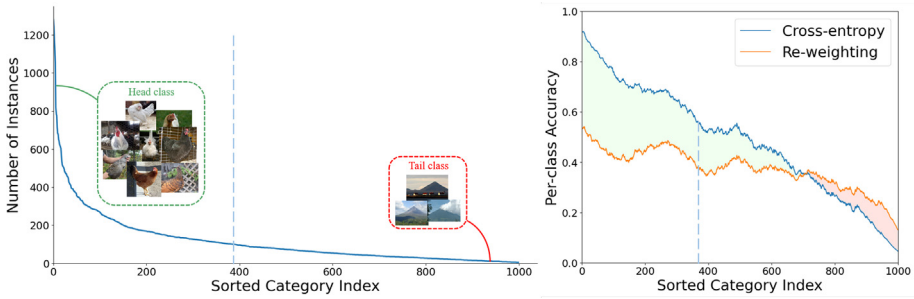


Fig. 1. *Left*: Illustration of the number of instances per class in ImageNet-LT, sorted in descending order. *Right*: Comparisons of class-balanced re-weighting with the cross-entropy training baseline on ImageNet-LT. Although the re-weighting method improves the performance on tail classes, it badly affects the model generalization and degrades the performance on head classes.

distillation loss. Thus, the learning objective of the student model is decoupled into two tasks, where each performs its own duty for learning universal representations and facilitating learning for tail classes, respectively. As a result, the model trained by BKD is able to show significant improvement on tail classes while maintaining performance of head classes simultaneously.

In experiments, we observe that the student model trained with BKD even outperforms the teacher model by a large margin, e.g., more than 10 points accuracy gain on imbalanced CIFAR-10. Furthermore, we conduct extensive experiments on several large-scale image and text benchmarks including iNaturalist 2018 [16], ImageNet-LT [17], Places-LT [17] and long-tailed AG's News [18]. Experimental results justify that our BKD framework is able to achieve state-of-the-art performance.

Our key contributions can be summarized as follows:

- We analyze the underlying cause of the drawback of re-weighting methods. Accordingly, we discuss our motivation and propose to disentangle the contradiction between the two goals of learning generalizable representations and facilitating learning for tail classes.
- We study knowledge distillation on imbalanced data and propose *balanced knowledge distillation (BKD)* as an effective distillation framework in long-tailed scenarios.
- We conduct experiments and demonstrate that our BKD framework achieves the two goals simultaneously and improves the performance of long-tailed classification significantly.

2. Related works

To alleviate the challenge of long-tailed learning, most of pioneer works have been proposed from three aspects:

Data. Re-sampling methods created a roughly balanced distribution by either over-sampling or under-sampling. Over-sampling [9,19–21] repeatedly samples training examples from the minority classes, the downside of which is the high potential risk of overfitting. To overcome this issue, SMOTE [22,23] is proposed to augment synthetic data created by interpolating neighboring data points. As opposed to over-sampling, under-sampling [6,24,25] randomly discards examples from the majority classes. When the imbalance is extreme, under-sampling may lose valuable information in majority classes.

Optimization objective. The key to this line of work is to adjust learning focus by modifying objective functions. Cost-sensitive re-weighting [26–28] assigns different weights to the classification loss terms corresponding to different classes [11–13,29] or different samples [30,31]. The traditional strategy re-weights classes proportionally to the inverse of their frequency of samples [10]. Taking data overlap into consideration, Cui et al. [12] design a class-balanced cross-entropy loss based on effective number of samples in each class. Other important works [29,32,33] propose to adjust output logits or softmax function based on label frequency. Besides those, Tang et al. [34] establish a causal inference framework to remove the harmful effect of SGD momentum via causal intervention.

Recently, [13,14] found that re-weighting on classification loss has a negative effect on representation learning, while vanilla instance-balanced cross-entropy loss gives the most generalizable representations. Motivated by this observation, two-stage [13,35–37] and two-branch [14,38] methods were proposed to take care of both representation learning and classifier learning. Differently, Zhang et al. [39] designed auxiliary tasks for the two parts to solve the dilemma.

Transfer learning. Many approaches [40–44] design additional modules or meta-network to transfer knowledge, e.g., distributions [45,46], memory features [17] or meta-knowledge [47,48], from head to tail classes. Beyond that, as a popular technique of transferring knowledge, knowledge distillation attracts attention in the field of long-tailed learning [49–51]. For example, *Learning From Multiple Experts (LFME)* [52] has been proposed as a multi-teacher framework, in which each teacher learns from a relatively balanced subset and then jointly distills knowledge for the student model. Recently, SSD [51] proposed a self-distillation framework that utilized self-supervision to train the initial feature.

3. Insight and motivation

Notation. Consider a classification problem on long-tailed training data. Let $x \in R^d$ and $y \in \{1, \ldots, C\}$ denote a data point and its label, respectively. Due to the imbalanced distribution, the number of training examples in each class n_i is highly imbalanced. Without loss of generality, we sort the classes in descending order of frequency so that $n_1 > \ldots > n_C$. Our goal is to learn a model $f : R^d \to R^C$ that estimates the probability $p_i = \text{softmax}(z_i)$ from the network output $z = [z_1, \ldots, z_C]^T$.

3.1. A closer look at re-weighting methods

Given a dataset, the most straightforward method minimizes the misclassification error by minimizing the following instance-balanced cross-entropy loss

$$L_{\text{CE}} = -\sum_i y_i \log p_i. \tag{1}$$

On the basis of cross-entropy loss, re-weighting methods assign different weights for different classes. In this paper, we mainly focus on the frequency based re-weighting, where the weight factor is negatively correlated with class frequency. Although effective, Cao et al. [13] and Zhou et al. [14] experimentally show that

these re-weighting methods have an adverse effect on representation learning and cause overfitting on tail classes. We intuitively show this phenomenon by visualizing the embedding space of long-tailed CIFAR-10 via t-Distributed Stochastic Neighbor Embedding (t-SNE) [53]. As shown in Fig. 2(a), the universal representation learned from re-weighting is worse than that learned from vanilla cross-entropy training, and the decision boundaries for head classes are ambiguous. We start by analyzing the underlying cause of such problems from the perspective of gradient.

The key of re-weighting is to balance the classification loss by a weight vector $\omega = [\omega_1, \ldots, \omega_C]^T$. ω_i is typically a decreasing transform of n_i. As a representative work of re-weighting methods, Cui et al. [12] formulate the weight factor inversely proportional to the effective number of samples for class i:

$$\omega_i = \frac{1-\beta}{1-\beta^{n_i}}, \tag{2}$$

where

$\beta \in (0, 1)$ is a hyperparameter to adjust the class-balanced term. In this setting, we have $\omega_1 < \ldots < \omega_C$. For a sample of category t, the corresponding class-balanced re-weighting loss L_{CB} is formulated as

$$\begin{aligned} L_{CB} &= -\sum_i \omega_i y_i \log p_i \\ &= -\omega_t \log p_t. \end{aligned} \tag{3}$$

The derivative of the L_{CB} with respect to the model's class-k output z_k is

$$\frac{\partial L_{CB}}{\partial z_k} = \begin{cases} \omega_t(p_t - 1), & \mathrm{k = t} \\ \omega_t p_k, & \mathrm{k \neq t} \end{cases} \tag{4}$$

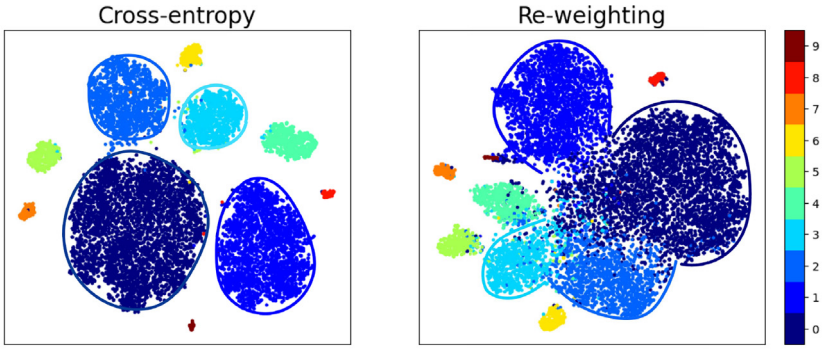
Depending on the frequency of class k, the gradient contributions of examples are far different. If category k is a tail class, for example $k = C$, in which case we have $\omega_k = \max_i \omega_i$, re-weighting seems reasonable:

1. If $k = t$, a large encouraging gradient $|\omega_t(p_t - 1)|$ is produced by a correct prediction for tail classes;
2. If $k \neq t$, the discouraging gradient $\omega_t p_k$ is relatively small as $\omega_t < \omega_k$. This is consistent with the idea of ignoring discouraging gradient for tail classes [54].

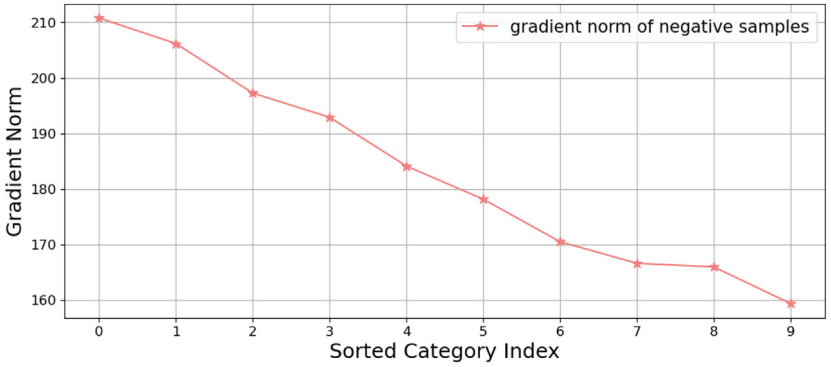
However, when it comes to head classes, for example $k = 1$, in which case we have $\omega_k = \min_i \omega_i$, the learning process is seriously hindered:

1. If $k = t$, the encouraging gradient $|\omega_t(p_t - 1)|$ is small as ω_t is close to zero;
2. If $k \neq t$, the discouraging gradient $\omega_t p_k$ is relatively large as $\omega_t > \omega_k$. This suppression effect is further accumulated because each positive sample of other classes will be treated as a negative sample for the class k.

In general, while tail classes benefit from re-weighting, the universal representation is damaged as the head classes suffer from overwhelmed discouraging gradients. To validate the analysis, we calculate the sum of $\frac{\partial L_{CB}}{\partial z_k}$ from negative samples by re-weighting



(a) Feature visualizations from cross-entropy and re-weighting training.



(b) The sum of gradient norm of negative samples by re-weighting training.

Fig. 2. Feature visualizations and gradient analysis of long-tailed CIFAR-10. Re-weighting is unfavourable for learning generalizable representations and the class boundaries are ambiguous, due to the overwhelmed discouraging gradients.

and plot it in Fig. 2(b). It shows that the head classes receive more discouraging gradients contributed by negative samples.

3.2. Motivation

As aforementioned, the class-balanced re-weighting methods lead to a sub-optimal result, because the focus on tail classes is naturally in conflict with the overall representation learning. Accordingly, our key idea is to decompose the task of long-tailed learning into two separate parts: learning generalizable representations and facilitating learning for tail classes. The motivation is twofold.

Motivation 1. To learn the most generalizable representations, directly applying class-balanced weight on cross-entropy loss may be sub-optimal and should be carefully avoided. It has been proved by experiments [13,14,35] and our analyses in the Section 3.1. It motivates us to keep the instance-balanced cross-entropy loss unchanged, which could exhaustively take advantage of the diversity of dominant data and guarantee the model to learn generalizable representations. Mixup [55], which encourages the model to behave linearly in-between training samples, is also adopted to further enhance the generalization.

Motivation 2. To facilitate the learning process of tail classes, we take advantage of the transfer ability of knowledge distillation. The predictive distributions in knowledge distillation contain informative knowledge which has low risk of overfitting to specific classes or examples. When distilling knowledge from long-tailed data, the emphasized learning for tail classes should be especially focused. To achieve this, we assign class-wise weights to the distillation loss, instead of distilling the knowledge for all classes without distinction.

4. Method

As motivated, we propose balanced knowledge distillation to decompose the two goals of long-tailed learning and achieve both simultaneously. In this section, we firstly revisit the conventional knowledge distillation method, and then describe the proposed method in detail. Furthermore, we discuss our method from the Bayesian view.

4.1. Recall of knowledge distillation

The conventional response-based knowledge distillation [15] consists of two stages. First, a teacher model is trained with cross-entropy loss. Second, the student model is trained together with ground truth targets in addition to the teacher's soft targets.

Formally, as a supplement to notations in the Section 3, we define the network outputs of the teacher model as $\hat{z} = [\hat{z}_1, \dots, \hat{z}_c]^T$ and the class probability $\hat{p}_i$ is calculated as $\hat{p}_i = \text{softmax}(\frac{\hat{z}_i}{T})$, where T is a temperature parameter that controls the *softness* of probability distribution over classes. Similarly, with a slight abuse of the notation, we re-define the student's probability in a more general form, $p_i = \text{softmax}(\frac{z_i}{T})$. The loss for the student is a combination of the cross-entropy loss L_{CE} and a Kullback–Leibler divergence loss L_{KL}:

$$L_{\text{KD}} = L_{\text{CE}} + \lambda L_{\text{KL}}, \quad (5)$$

$$\text{where } L_{\text{KL}} = T^2 \sum_i \hat{p}_i \log \frac{\hat{p}_i}{p_i},$$

and λ is a hyperparameter balancing the weight of the two losses. For the convenience of analysis, we set $\lambda = 1$.

Guided by the dual objectives, the student model could learn high-quality representations by minimizing the softmax cross-entropy loss, and meanwhile learn specific information from teacher model via the KL-divergence loss. This is desirable for achieving our two motivations simultaneously.

4.2. Balanced knowledge distillation

Although effective for model compression, the conventional knowledge distillation framework fails to improve the model performance dramatically by distilling another structured identically model. Particularly, if trained on long-tailed data, the teacher model is naturally biased towards the head classes. In the distillation process, the predictive information for the tail classes is overwhelmed by the head classes. Therefore, the student model guided by such biased model may exhibit even worse performance on tail classes.

Algorithm 1: Balanced Knowledge Distillation

require: A training dataset D, maximum epoch number E and weight vector ω, as well as a teacher model $t_{\hat{\theta}}$ and a student model s_θ with identical structure.

```
Initialize teacher's parameters θ̂ randomly;
for e = 1, ..., E do
    Sample a minibatch B̂ from D;
    Update t_θ̂ by minimizing L_CE (Eq. 1) on B̂;
end for
Initialize student's parameters θ randomly;
for e = 1, ..., E do
    Sample a minibatch B from D;
    Use t_θ̂ to produce predictions p̂ on mixup(B);
    Normalize ω^T p̂ by ω_i p̂_i ← ω_i p̂_i / Σ_i ω_i p̂_i.
    Update s_θ by minimizing L_BKD (Eq. 8) on mixup(B);
end for
```

To solve this problem, we propose balanced knowledge distillation (BKD), which distills knowledge with focus. Our BKD follows the two-stage teacher-student learning pipeline as mentioned above. The key difference is that we take class priors into consideration and control the importance of distilled information for different classes. Concretely, given a teacher model trained with vanilla cross-entropy loss, the student model is trained by minimizing the summation of an instance-balanced cross-entropy loss and a class-balanced distillation loss. The total loss of balanced knowledge distillation for the student model is formulated as

$$\begin{aligned} L &= L_{\text{CE}} + T^2 \sum_i \omega_i \hat{p}_i \log \frac{\hat{p}_i}{p_i} \\ &= L_{\text{CE}} + T^2 \sum_i (\omega_i \hat{p}_i \log \hat{p}_i - \omega_i \hat{p}_i \log p_i). \end{aligned} \quad (6)$$

The weight factor ω_i is defined as Eq. 2. In this way, the knowledge from tail classes is distilled with focus to facilitate learning on tail classes.

Despite the validity from the perspective of distillation with focus, the non-negativity of KL-divergence is damaged because the weighted probabilities of the teacher model do not sum to one any more, i.e., $\sum_i \omega_i \hat{p}_i \neq 1$. To keep the divergence loss non-negative, we consider $\omega^T \hat{p}$ as a whole and normalize it to one. Since the term $\omega_i \hat{p}_i \log \hat{p}_i$ is constant for a fixed teacher model, we first reformulate Eq. 6 to

$$L = L_{\text{CE}} + T^2 \sum_i (\omega_i \hat{p}_i \log \omega_i \hat{p}_i - \omega_i \hat{p}_i \log p_i). \quad (7)$$

Then we normalize $\omega_i \hat{p}_i$ by $\omega_i \hat{p}_i \leftarrow \frac{\omega_i \hat{p}_i}{\sum_i \omega_i \hat{p}_i}$. Accordingly, the loss can be rewritten as

$$L_{\mathrm{BKD}} = L_{\mathrm{CE}} + T^2 \sum_i \omega_i \hat{p}_i \log \frac{\omega_i \hat{p}_i}{p_i}. \tag{8}$$

With the normalization, the non-negativity of KL-divergence is satisfied and a definite lower bound of the loss function is now guaranteed.

During the training of the student model, we adopt Mixup [55] strategy to enhance the generalization of learned representations. The overall BKD framework is summarized in Algorithm 1.

4.3. Discussion

In this subsection, we discuss the proposed method from the Bayesian view and reveal that the weighted probabilities of the teacher model actually approximate the balanced predictions. Let the probability that refers to long-tailed training data and balanced test data be denoted as P_{LT} and P_{Bal}, respectively. The label priors on the training and the test data are denoted as $P_{\mathrm{LT}}(y)$ and $P_{\mathrm{Bal}}(y)$, respectively. According to Bayes' theorem, the prediction of the teacher model implicitly corresponds to

$$P_{\mathrm{LT}}(y_i|x) = \frac{P(x|y_i)P_{\mathrm{LT}}(y_i)}{P(x)}. \tag{9}$$

Following [6], $P(x|y_i)$ and $P(x)$ are the same for training and test data. Then we can obtain a balanced prediction by

$$P_{\mathrm{Bal}}(y_i|x) = \frac{P(x|y_i)P_{\mathrm{Bal}}(y_i)}{P(x)} = P_{\mathrm{LT}}(y_i|x)\frac{P_{\mathrm{Bal}}(y_i)}{P_{\mathrm{LT}}(y_i)}. \tag{10}$$

As $P_{\mathrm{Bal}}(y)$ is uniform, it could be omitted among classes. Then, the weighted teacher prediction $\omega_i \hat{p}_i$ is actually an approximation of $P_{\mathrm{Bal}}(y_i|x)$, where ω_i estimates inverse proportion of label prior $\frac{1}{P_{\mathrm{LT}}(y_i)}$ while considering data overlap. In fact, the BKD implicitly uses a balanced teacher prediction to guide the student model.

5. Experiments

5.1. Datasets

We evaluate the proposed method on five long-tailed image datasets, including imbalanced CIFAR-10/-100, Places-LT, ImageNet-LT and iNaturalist 2018, as well as an artificially created long-tailed version of AG's News for text classification.

Imbalanced CIFAR-10 and CIFAR-100. The original version of CIFAR-10 and CIFAR-100 contains 60,000 images, 50,000 for training and 10,000 for validation with 10 and 100 classes, respectively. Following the prior work [12,13], we use both the long-tailed version and step imbalanced [6] version of both the CIFAR datasets by downsampling examples per class with different ratios. The imbalance ratio ρ denotes the ratio between the number of training examples between the most frequent class and the least frequent class. We use $\rho = 10, 50, 100$ in our experiments.

Places-LT. Places365-Standard [56] is a large-scale image database for scene recognition, with more than 1.8 million training images from 365 categories. We construct Places-LT by the same sampling strategy as [17], with the number of images per class ranging from 4980 to 5.

iNaturalist 2018. The iNaturalist species classification dataset [16] is a large-scale real-world dataset. The iNaturalist 2018 dataset contains 437,513 training images from 8142 classes, with an imbalance ratio of 500. For fair comparisons, we use the official training and validation splits in our experiments.

ImageNet-LT. ImageNet-LT is constructed by sampling a subset of ImageNet-2012 [1] following the Pareto distribution with the power value $\alpha = 6$. It has 115.8 K images from 1000 categories, with the number of images per class ranging from 1280 to 5.

Long-tailed AG's News. The original version of AG's News introduced by [18] contains 4 classes of news articles, with 30,000 training samples and 1900 test examples for each class. We create a long-tailed training set with an imbalance ratio of 100, where the number of training samples per class ranges from 30,000 to 300.

5.2. Implementation details

In our experiments, the teacher model and the student model have identical network architecture in each setting. As summarized in Algorithm 1, we first train the teacher model with vanilla cross-entropy loss and then train the student model with the proposed BKD loss. Following [12], we set $\beta = 0.9999$ in Eq. 2, and the temperature T is set to 2. All networks are trained with SGD. Unless otherwise specified, the base learning rate is set to 0.2, with cosine learning rate decay. Other details are given below.

Implementation details for imbalanced CIFAR. We employ ResNet-32 as the backbone network and follow the training recipe of [57] for the teacher model and the student model. Both models are trained for 200 epochs with a batch size of 128. The learning rate is initialized as 0.1 and decayed by 0.01 at the 160th epoch and again at the 180th epoch. For imbalanced CIFAR-10, we first train the student model with vanilla knowledge distillation before the 160th epoch, and then deploy our BKD, following [13].

Implementation details for Places-LT. We choose pretrained ResNet-152 as the backbone network, following [17]. Both the teacher model and the student model are trained for 30 epochs with a learning rate decay of 0.1 every 10 epochs.

Implementation details for iNaturalist 2018. We use ResNet-50 as our backbone network. Following [35], we train the models for 90 epochs and 200 epochs, with a batch size of 256.

Implementation details for ImageNet-LT. We choose two backbones, ResNet-10 and ResNeXt-50, for ImageNet-LT. Both the teacher model and the student model are trained for 90 epochs with the batch size of 512.

Implementation details for long-tailed AG's News. We use a one-layer TextCNN [58] as our baseline network, with pre-trained word vectors from GloVe [59]. Both the teacher model and the student model are trained for 30 epochs with a batch size of 64. The learning rate is initialized as 0.1 and decayed by 0.01 at the 20th epoch and the 25th epoch. Mixup is not used for the text data.

5.3. Competing methods

We report the results for relevant baselines, such as CE, KD [15], CB [12], and Mixup [55]. In addition, we compare the proposed BKD with decoupled training [35], modified loss [13,29,30,33,61], knowledge transfer [17,49–52], and other state-of-the-art methods [60,36]. Furthermore, we also compare our method with RIDE [62] following its multi-expert framework.

5.4. Evaluation protocol

We validate the effectiveness of our method on balanced test sets and report the Top-1 accuracy. Besides, we also calculate the accuracy of three subsets: *Many-shot classes* (with over 100 training samples), *Medium-shot classes* (with $20 \sim 100$ training samples) and *Few-shot classes* (with under 20 training samples).

5.5. Experimental results

Results on imbalanced CIFAR datasets. We conduct experiments on long-tailed and step imbalanced CIFAR-10/-100 with three different imbalance ratios $\rho = 10, 50, 100$. The validation accuracy is reported in Table 1 and 2. It can be seen that the pro-

Table 1
Top-1 validation accuracy of ResNet-32 on imbalanced CIFAR-100. Best results are marked in bold.

Imbalance type	Long-tailed			Step		
Imbalance ratio	100	50	10	100	50	10
CE	38.3	43.9	55.7	38.8	40.3	54.9
KD [15]	40.4	45.5	59.2	39.3	41.2	56.5
Mixup [55] (ICLR'18)	38.7	43.6	57.5	38.3	39.0	51.1
CB [12] (CVPR'19)	32.7	38.6	54.9	23.3	29.5	53.3
BBN [14] (CVPR'20)	42.6	47.0	59.1	42.8	48.1	59.1
BSCE [29] (NeurIPS'20)	42.1	47.2	57.9	47.0	50.6	59.7
SSP [60] (NeurIPS'20)	43.4	47.1	58.9	45.7	-	59.7
TDE [34] (NeurIPS'20)	44.1	50.3	59.6	44.5	48.6	58.0
DiVE [50] (ICCV'21)	45.4	51.1	62.0	-	-	-
SSD [51] (ICCV'21)	46.0	50.5	**62.3**	-	-	-
LADE [33] (CVPR'21)	45.4	50.5	61.7	-	-	-
Mixup + LA [32] (ICLR'21)	46.2	50.2	59.8	42.2	48.6	59.8
Mixup + VS [61] (NeurIPS'21)	45.4	-	-	45.2	-	-
BKD	**46.5**	**51.7**	62.0	**47.3**	**52.2**	**61.8**

Table 2
Top-1 validation accuracy of ResNet-32 on imbalanced CIFAR-10.

Imbalance type	Long-tailed			Step		
Imbalance ratio	100	50	10	100	50	10
CE	70.4	74.8	86.4	63.1	69.8	84.7
KD [15]	70.7	77.9	87.5	64.6	70.0	85.4
Mixup [55] (ICLR'18)	71.4	77.1	86.6	64.4	70.1	83.9
CB [12] (CVPR'19)	72.1	77.7	86.4	61.0	71.7	85.2
BBN [14] (CVPR'20)	79.8	82.2	88.3	77.9	81.7	87.6
BSCE [29] (NeurIPS'20)	77.6	80.4	87.8	77.2	81.8	88.3
SSP [60] (NeurIPS'20)	77.8	82.1	88.5	77.1	-	88.2
TDE [34] (NeurIPS'20)	80.6	83.6	88.5	78.0	81.7	88.2
Mixup + LA [32] (ICLR'21)	81.6	85.0	88.7	80.6	83.9	88.6
Mixup + VS [61] (NeurIPS'21)	82.1	-	-	79.1	-	-
BKD	**82.5**	**85.1**	**89.5**	**80.9**	**83.9**	**89.5**

posed BKD significantly outperforms the cross-entropy training and conventional knowledge distillation. It also demonstrates that BKD exhibits superior performance compared with existing state-of-the-art methods, including recent DiVE [50] and LADE [33]. Specifically, when the long-tailed imbalance is relatively slight, e.g., CIFAR-100 with $\rho = 10$, BKD is on par with DiVE and LADE. However, when the data distribution is severely skewed, the advantage of BKD is more obvious. Besides, we conduct experiments on combining LA [32] and VS [61] with Mixup, and find that BKD still surpasses their combinations. To intuitively understand the advantage of BKD, we further visualize three confusion matrices respectively by the models of CE, CB and our BKD on long-tailed CIFAR-10. As shown in Fig. 3, the improvement from CB is marginal, while BKD significantly improves the performance on tail classes and meanwhile maintains similar head performance as cross-entropy training.

Results on large-scale image datasets. Evaluation results for ImageNet-LT, iNaturalist 2018 and Places-LT are reported in Table 3–5, respectively. Comparing our BKD with previous methods, we observe consistent improvements. On ImageNet-LT, our method achieves the best overall accuracy for both the ResNet-10 and ResNeXt-50 backbones. Compared with LFME [52] which distills knowledge from multiple experts, we obtain 4.7% accuracy gain for ResNet-10, even without multiple teacher models and complex learning schedules. In Fig. 4, we further plot the per-class accuracy. It is worth noting that, different from CB re-

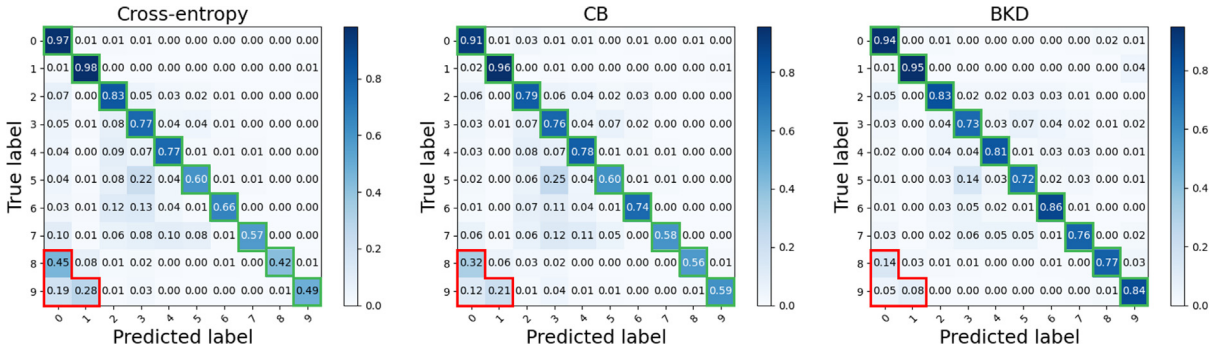


Fig. 3. Confusion matrices by cross-entropy training (*left*), class-balanced re-weighting (*middle*) and our BKD (*right*) on long-tailed CIFAR-10. The model trained by BKD produces more correct predictions as located in the diagonal (framed in green box). In particular, BKD reduces the misclassification from tail classes to head classes (framed in red box).

Table 3
Top-1 accuracy on ImageNet-LT with ResNet-10 and ResNeXt-50 backbones. † denotes results reported in [52].

Backbone	ResNet-10				ResNeXt-50			
Method	Many	Medium	Few	**Overall**	Many	Medium	Few	**Overall**
CE	56.9	25.4	3.6	34.6	65.9	37.5	7.7	44.4
KD [15]	**58.8**	26.6	3.4	35.8	**67.5**	39.3	8.2	45.9
CB [12] (CVPR'19)	39.6	32.7	16.8	33.2	42.3	34.8	17.9	35.4
Focal† [30] (CVPR'17)	36.4	29.9	16.0	30.5	-	-	-	-
OLTR† [17] (CVPR'19)	43.2	35.1	18.5	35.6	-	-	-	-
LFME† [52] (ECCV'19)	47.1	35.0	17.5	37.2	-	-	-	-
NCM [35] (ICLR'20)	42.8	33.1	20.6	35.2	56.6	45.3	23.1	47.3
τ-Norm [35] (ICLR'20)	50.6	37.9	19.2	40.3	59.1	46.9	30.7	49.4
LWS [35] (ICLR'20)	49.9	**38.7**	**23.8**	41.0	60.2	47.2	30.3	49.9
Mixup + cRT [35] (ICLR'20)	50.7	38.3	23.6	41.1	61.2	46.2	27.5	49.4
BKD	55.6	36.7	21.4	**41.9**	64.8	**47.7**	**33.3**	**52.3**

Table 4
Top-1 accuracy on iNaturalist 2018 with ResNet-50 backbone.

Backbone	90 epochs				200 epochs			
Method	Many	Medium	Few	**Overall**	Many	Medium	Few	**Overall**
CE	**72.7**	63.1	57.4	61.8	75.7	66.9	61.7	65.8
KD [15]	72.6	63.6	57.4	62.2	**77.1**	67.6	61.4	66.1
NCM [35] (ICLR'20)	55.5	57.9	59.3	58.2	61.0	63.5	63.3	63.1
τ-Norm [35] (ICLR'20)	65.6	65.3	65.9	65.6	71.1	68.9	69.3	69.3
LWS [35] (ICLR'20)	65.0	66.3	65.5	65.9	71.0	69.8	68.8	69.5
Mixup + cRT [35] (ICLR'20)	68.7	66.1	63.4	65.3	72.8	68.5	65.5	67.8
CBD [49] (BMVC'21)	-	-	-	-	70.5	69.5	66.5	68.4
DisAlign [36] (CVPR'21)	64.1	**68.5**	67.9	**67.8**	69.0	**71.1**	70.2	70.6
BKD	68.9	67.2	**68.3**	**67.8**	72.7	70.3	**71.9**	**71.2**

Table 5
Top-1 accuracy on Places-LT with pre-trained ResNet-152.

Method	Many	Medium	Few	**Overall**
CE	**45.7**	27.3	8.2	30.2
KD [15]	**45.7**	28.1	9.0	30.7
CB [12] (CVPR'19)	36.5	29.7	9.2	28.2
LFME [52] (ECCV'20)	38.4	39.1	21.7	35.2
NCM [35] (ICLR'20)	40.4	37.1	27.3	36.4
τ-Norm [35] (ICLR'20)	37.8	**40.7**	31.8	37.9
LWS [35] (ICLR'20)	40.6	39.1	28.6	37.6
Mixup + cRT [35] (ICLR'20)	43.3	36.4	25.1	36.7
BKD	43.7	37.1	**34.2**	**38.9**

weighting which gains on tail classes at the expense of performance drop on head classes, BKD significantly facilitates learning for tail classes, and meanwhile has negligible harm on head classes. In addition, Our method achieves the overall accuracy of 67.8% and 71.2% on iNaturalist 2018 with 90 and 200 training epochs, surpassing the two-stage decoupled method [35] by 1.9% and 1.7% respectively. Compared with the distribution alignment method [36], BKD promotes the performance on many-shot and few-shot classes simultaneously. On Places-LT dataset with the highest imbalance ratio of 996, the performance on tail class is significantly improved by BKD, which is 2.4% higher than the best of other methods.

Results on long-tailed AG's News. For text classification, we report the overall accuracy Acc_{all}, as well as the results for the four classes. Here Acc_0 denotes the top-1 accuracy for the most frequent class, while Acc_3 corresponds to the least frequent class. As shown in Table 6, the results are consistent with other datasets. KD slightly improves performance of head classes but degrades performance of tail classes, while CB presents just the opposite tendency. By contrast, our method achieves substantial improvements on tail classes while keeping the performance on head classes, and increases the overall accuracy by 13.2% over conventional knowledge distillation.

5.6. Ablation study

The merit of each component. To explore the effectiveness of each component in BKD, we conduct ablation experiments on long-

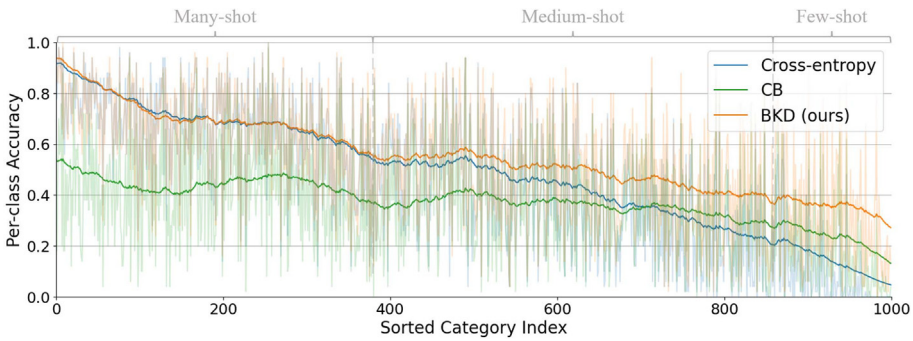


Fig. 4. Comparisons of per-class accuracy between cross-entropy training, CB re-weighting and BKD on ImageNet-LT.

Table 6
Per-class and overall accuracy on long-tailed AG's News.

Method	Acc_0	Acc_1	Acc_2	Acc_3	Acc_{all}
CE	97.5	91.6	58.8	28.7	69.2
KD [15]	**97.7**	**92.4**	57.7	20.6	67.1
CB [12]	89.2	84.1	**73.6**	50.4	74.3
BKD	94.5	92.1	63.3	**71.2**	**80.3**

Table 7
Ablation study of each component on long-tailed CIFAR-10. The absolute improvements over CE baseline is reported in brackets.

Distillation	Mixup	Weight	Imbalance ratio		
			100	50	10
			70.4	74.8	86.4
✓			70.7 (+0.3)	77.9 (+3.1)	87.5 (+1.1)
✓	✓		72.7 (+2.3)	78.6 (+3.8)	87.5 (+1.1)
✓		✓	81.7 (+11.3)	83.8 (+9.0)	89.2 (+2.8)
✓	✓	✓	82.5 (+12.1)	85.1 (+10.3)	89.5 (+3.1)

tailed CIFAR-10. We report the results in Table 7, where "Distillation" means whether or not the knowledge distillation loss is enabled, "Mixup" means whether or not Mixup is used for representation enhancement, and "Weight" means whether or not the re-balance weight is enabled on the knowledge distillation loss. It is shown that the re-balance weight on knowledge distillation loss is the key for improving performance, which gains accuracy by 10.0%, 5.9% and 1.7% respectively for the three imbalance ratios. In addition, it shows that Mixup also brings benefits owing to the enhanced model generalization.

Different temperature T for distillation. In Fig. 5, we study the impact of tuning the temperature parameter T ranging from 1.0 to 4.0 on the long-tailed CIFAR datasets. We find the optimal temperature is 2.0 for both datasets. Accordingly, we fix $T = 2.0$ in all our experiments.

Different β for calculating class weight. Following [12], we take data overlap into consideration and formulate the class weight based on the effective number (EN) of samples, which is controlled by the hyper-parameter β as in Eq. 2. Fig. 6 shows how the different value of β affects the performance on long-tailed CIFAR-10. We find that a larger β tends to be superior. When $\beta = 0.9999$, the result is better than directly weighting by inverse frequency (IF) of classes.

5.7. Model validation

Comprehensive comparison with baselines. We first make a comprehensive comparison of our BKD with the baselines in Fig. 7. It shows that the class-balanced re-weighting method is more

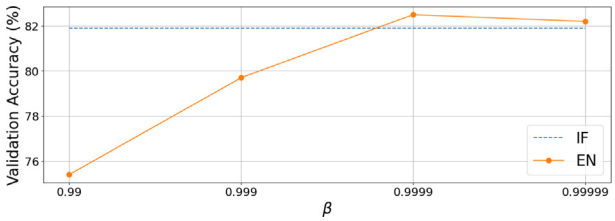


Fig. 6. Impact of β on long-tailed CIFAR-10. *IF* denotes calculating weights by inverse frequency of classes, while *EN* denotes calculating weights based on effective number of samples [12].

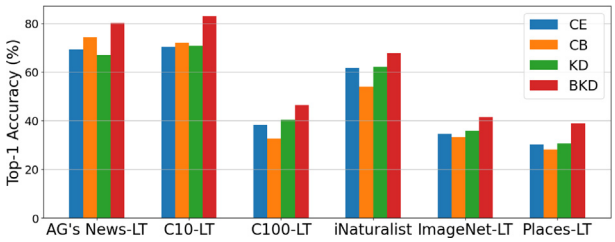


Fig. 7. Comprehensive comparison with baselines.

effective on small datasets, e.g., CIFAR-10. However, when dealing with large-scale and extremely imbalanced datasets, CB may lead to even worse performance. The phenomenon is consistent with our analysis in Section 3.1. KD is slightly superior to vanilla cross-entropy training due to the ability of transferring knowledge, while our BKD further improves the performance by 5.5% ∼ 13.2% over KD on all the reported datasets.

Feature visualization. In order to compare the representations learned from our BKD with the CB re-weighting, we select ten classes from long-tailed CIFAR-100 and sort them in the decreasing order of label frequency. Then we plot the t-SNE [53] visualization of embedding space on the selected classes. As shown in Fig. 8, the CB re-weighting leads to mixed representations and ambiguous boundaries, while BKD provides clearer boundaries to separate different classes. It verifies the effectiveness of the proposed method for simultaneously learning generalizable representations and improving tail performance.

5.8. Results of ensemble models

The ensemble method, which combines several expert models to produce one optimal prediction, has attracted a lot of interest in the field of long-tailed learning [62–64]. To test the performance of BKD on ensemble models, we follow the implementation of RIDE [62] to form a multi-expert framework and deploy BKD in it, denoted as RIDE-BKD. We first train a multi-expert teacher model

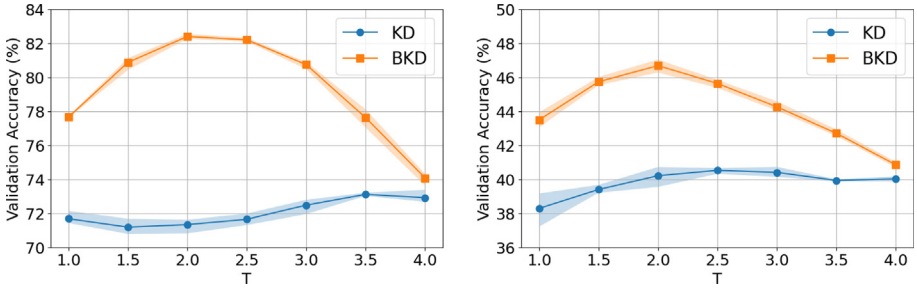
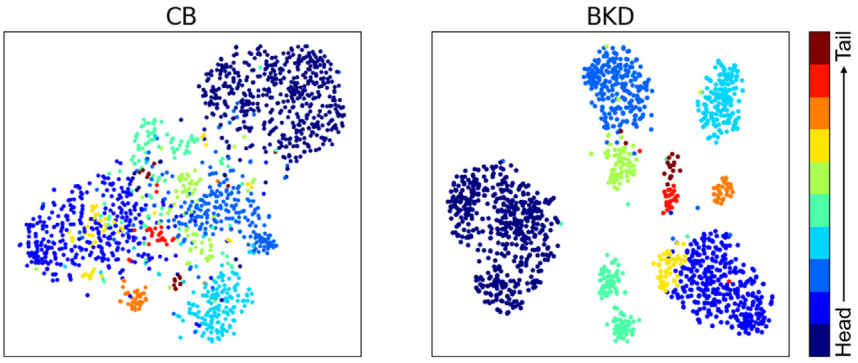
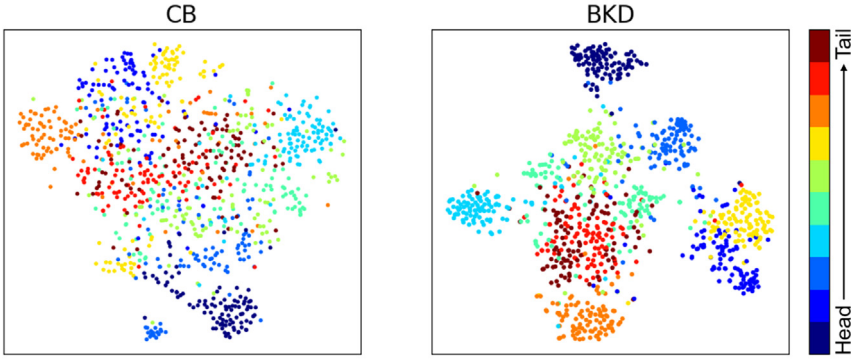


Fig. 5. Impact of the temperature parameter T on long-tailed CIFAR-10 (*left*) and CIFAR-100 (*right*).



(a) Visualization on the training set.



(b) Visualization on the test set.

Fig. 8. The t-SNE visualizations of learned representations on long-tailed CIFAR-100.

Table 8
Top-1 accuracy of ensemble models on long-tailed CIFAR-100. *#expert* indicates the number of experts. The RIDE results are copied from [62].

Method	#expert	Many	Medium	Few	Overall
Baseline	1	66.1	37.3	10.6	39.3
RIDE [62]	2	67.9	48.4	21.8	47.0
RIDE [62]	3	**68.1**	**49.2**	23.9	48.0
RIDE-BKD	2	62.3	47.8	**34.8**	48.9
RIDE-BKD	3	64.0	49.0	33.6	**49.6**

without re-balance, and then train a structurally identical student model by the BKD. We test on long-tailed CIFAR-100 ($\rho = 100$) with $2 \sim 3$ experts. The results are reported in Table 8. Although RIDE surpasses the CE baseline a lot due to the ensemble of multiple experts, the results on few-shot classes are still weak. However, RIDE-BKD consistently improve the performance over RIDE, especially for the few-shot classes.

6. Conclusion

In this work, we propose a novel knowledge distillation framework to address long-tailed classification problem. We first analyze that the class-balanced re-weighting method cares for tail classes but fails to learn high-quality representations due to overwhelming discouraging gradients on head classes. Instead, we propose balanced knowledge distillation, which consists of an instance-balanced classification loss and a class-balanced distillation loss. With the BKD, the goals of learning generalizable representations and facilitating learning for tail classes could be achieved simultaneously. We conduct experiments on several image and text benchmarks, and prove BKD to be a simple and effective framework on long-tailed data.

However, there are still some limitations in this work. For example, a teacher model has to be trained in advance, which may increase the time for training. We also note that the performance on head classes may drop slightly in some extremely imbalanced cases, e.g. iNaturalist 2018. Future work should be designed to overcome these limitations.

CRediT authorship contribution statement

Shaoyu Zhang: Conceptualization, Methodology, Software, Formal analysis, Writing – original draft, Visualization. **Chen Chen:** Conceptualization, Methodology, Validation, Writing – review & editing. **Xiyuan Hu:** Conceptualization, Methodology, Validation. **Silong Peng:** Conceptualization, Validation, Formal analysis, Supervision, Funding acquisition.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This work is supported by the National Key R&D Program of China under Grant 2021YFF0602101 and the National Science Foundation of China under Grant NSFC 61906194.

References

[1] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, L. Fei-Fei, Imagenet: A large-scale hierarchical image database, in: IEEE Conference on Computer Vision and Pattern Recognition, 2009, pp. 248–255.

[2] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, et al., Imagenet large scale visual recognition challenge, Int. J. Comput. Vision 115 (3) (2015) 211–252.

[3] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, C.L. Zitnick, Microsoft coco: Common objects in context, in: European Conference on Computer Vision, Springer, 2014, pp. 740–755.

[4] X. Zhang, Z. Fang, Y. Wen, Z. Li, Y. Qiao, Range loss for deep face recognition with long-tailed training data, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 5409–5418.

[5] Y. Zhang, B. Kang, B. Hooi, S. Yan, J. Feng, Deep long-tailed learning: A survey, arXiv preprint arXiv:2110.04596.

[6] M. Buda, A. Maki, M.A. Mazurowski, A systematic study of the class imbalance problem in convolutional neural networks, Neural Networks 106 (2018) 249–259.

[7] C. Drummond, R.C. Holte, et al., C4. 5, class imbalance, and cost sensitivity: why under-sampling beats over-sampling, in: Workshop on learning from imbalanced datasets II, Vol. 11, Citeseer, 2003, pp. 1–8.

[8] M. Kubat, S. Matwin, et al., Addressing the curse of imbalanced training sets: one-sided selection, in: International Conference on Machine Learning, 1997, pp. 179–186.

[9] J. Byrd, Z. Lipton, What is the effect of importance weighting in deep learning?, in: International Conference on Machine Learning, PMLR, 2019, pp. 872–881.

[10] C. Huang, Y. Li, C.C. Loy, X. Tang, Learning deep representation for imbalanced classification, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 5375–5384.

[11] S.H. Khan, M. Hayat, M. Bennamoun, F.A. Sohel, R. Togneri, Cost-sensitive learning of deep feature representations from imbalanced data, IEEE Trans. Neural Networks Learn. Syst. 29 (8) (2017) 3573–3587.

[12] Y. Cui, M. Jia, T.-Y. Lin, Y. Song, S. Belongie, Class-balanced loss based on effective number of samples, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 9268–9277.

[13] K. Cao, C. Wei, A. Gaidon, N. Arechiga, T. Ma, Learning imbalanced datasets with label-distribution-aware margin loss, Advances in Neural Information Processing Systems 32.

[14] B. Zhou, Q. Cui, X.-S. Wei, Z.-M. 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, pp. 9719–9728.

[15] G. Hinton, O. Vinyals, J. Dean, Distilling the knowledge in a neural network, in: NIPS Deep Learning and Representation Learning Workshop, 2015.

[16] G. Van Horn, O. Mac Aodha, Y. Song, Y. Cui, C. Sun, A. Shepard, H. Adam, P. Perona, S. Belongie, The inaturalist species classification and detection dataset, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 8769–8778.

[17] Z. Liu, Z. Miao, X. Zhan, J. Wang, B. Gong, S.X. Yu, Large-scale long-tailed recognition in an open world, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 2537–2546.

[18] X. Zhang, J. Zhao, Y. LeCun, Character-level convolutional networks for text classification, Advances in Neural Information Processing Systems 28.

[19] L. Shen, Z. Lin, Q. Huang, Relay backpropagation for effective learning of deep convolutional neural networks, in: European Conference on Computer Vision, Springer, 2016, pp. 467–482.

[20] X. Ye, H. Li, A. Imakura, T. Sakurai, An oversampling framework for imbalanced classification based on laplacian eigenmaps, Neurocomputing 399 (2020) 107–116.

[21] X. Wang, J. Xu, T. Zeng, L. Jing, Local distribution-based adaptive minority oversampling for imbalanced data classification, Neurocomputing 422 (2021) 200–213.

[22] N.V. Chawla, K.W. Bowyer, L.O. Hall, W.P. Kegelmeyer, Smote: synthetic minority over-sampling technique, J. Artif. Intell. Res. 16 (2002) 321–357.

[23] H. Han, W.-Y. Wang, B.-H. Mao, Borderline-smote: a new over-sampling method in imbalanced data sets learning, in: International Conference on Intelligent Computing, Springer, 2005, pp. 878–887.

[24] H. He, E.A. Garcia, Learning from imbalanced data, IEEE Trans. Knowl. Data Eng. 21 (9) (2009) 1263–1284.

[25] N. Japkowicz, S. Stephen, The class imbalance problem: A systematic study, Intell. Data Anal. 6 (5) (2002) 429–449.

[26] C. Elkan, The foundations of cost-sensitive learning, in: International Joint Conference on Artificial Intelligence, Vol. 17, 2001, pp. 973–978.

[27] K.M. Ting, A comparative study of cost-sensitive boosting algorithms, in: International Conference on Machine Learning, 2000.

[28] N. Sarafianos, X. Xu, I.A. Kakadiaris, Deep imbalanced attribute classification using visual attention aggregation, in: European Conference on Computer Vision, Springer, 2018, pp. 708–725.

[29] J. Ren, C. Yu, S. Sheng, X. Ma, H. Zhao, S. Yi, H. Li, Balanced meta-softmax for long-tailed visual recognition, in: Advances in Neural Information Processing Systems, Vol. 33, 2020, pp. 4175–4186.

[30] T.-Y. Lin, P. Goyal, R. Girshick, K. He, P. Dollár, Focal loss for dense object detection, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 2980–2988.

[31] S. Khan, M. Hayat, S.W. Zamir, J. Shen, L. Shao, Striking the right balance with uncertainty, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 103–112.

[32] A.K. Menon, A. Veit, A.S. Rawat, H. Jain, S. Jayasumana, S. Kumar, Long-tail learning via logit adjustment, in: International Conference on Learning Representations (ICLR), 2021.

[33] Y. Hong, S. Han, K. Choi, S. Seo, B. Kim, B. Chang, Disentangling label distribution for long-tailed visual recognition, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 6626–6636.

[34] K. Tang, J. Huang, H. Zhang, Long-tailed classification by keeping the good and removing the bad momentum causal effect, Advances in Neural Information Processing Systems 33.

[35] B. Kang, S. Xie, M. Rohrbach, Z. Yan, A. Gordo, J. Feng, Y. Kalantidis, Decoupling representation and classifier for long-tailed recognition, in: Eighth International Conference on Learning Representations (ICLR), 2020.

[36] S. Zhang, Z. Li, S. Yan, X. He, J. Sun, Distribution alignment: A unified framework for long-tail visual recognition, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 2361–2370.

[37] Z. Zhong, J. Cui, S. Liu, J. Jia, Improving calibration for long-tailed recognition, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 16489–16498.

[38] P. Wang, K. Han, X.-S. Wei, L. Zhang, L. Wang, Contrastive learning based hybrid networks for long-tailed image classification, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 943–952.

[39] J. Zhang, L. Liu, P. Wang, J. Zhang, Exploring the auxiliary learning for long-tailed visual recognition, Neurocomputing 449 (2021) 303–314.

[40] M.A. Jamal, M. Brown, M.-H. Yang, L. Wang, B. Gong, Rethinking class-balanced methods for long-tailed visual recognition from a domain adaptation perspective, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 7610–7619.

[41] M. Ren, W. Zeng, B. Yang, R. Urtasun, Learning to reweight examples for robust deep learning, in: International Conference on Machine Learning, PMLR, 2018, pp. 4334–4343.

[42] J. Kim, J. Jeong, J. Shin, M2m: Imbalanced classification via major-to-minor translation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 13896–13905.

[43] P. Chu, X. Bian, S. Liu, H. Ling, Feature space augmentation for long-tailed data, in: European Conference on Computer Vision, Springer, 2020, pp. 694–710.

[44] Q. Chen, Q. Liu, E. Lin, A knowledge-guide hierarchical learning method for long-tailed image classification, Neurocomputing 459 (2021) 408–418.

[45] X. Yin, X. Yu, K. Sohn, X. Liu, M. Chandraker, Feature transfer learning for face recognition with under-represented data, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 5704–5713.

[46] J. Liu, Y. Sun, C. Han, Z. Dou, W. Li, Deep representation learning on long-tailed data: A learnable embedding augmentation perspective, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 2970–2979.

[47] Y.-X. Wang, D. Ramanan, M. Hebert, Learning to model the tail, Advances in Neural Information Processing Systems 30.

[48] S. Li, K. Gong, C.H. Liu, Y. Wang, F. Qiao, X. Cheng, Metasaug: Meta semantic augmentation for long-tailed visual recognition, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 5212–5221.

[49] A. Iscen, A. Araujo, B. Gong, C. Schmid, Class-balanced distillation for long-tailed visual recognition, arXiv preprint arXiv:2104.05279.

[50] Y.-Y. He, J. Wu, X.-S. Wei, Distilling virtual examples for long-tailed recognition, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 235–244.

[51] T. Li, L. Wang, G. Wu, Self supervision to distillation for long-tailed visual recognition, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 630–639.

[52] L. Xiang, G. Ding, J. Han, Learning from multiple experts: Self-paced knowledge distillation for long-tailed classification, in: European Conference on Computer Vision, Springer, 2020, pp. 247–263.

[53] L. Van der Maaten, G. Hinton, Visualizing data using t-sne, J. Mach. Learn. Res. 9(11).

[54] J. Tan, C. Wang, B. Li, Q. Li, W. Ouyang, C. Yin, J. Yan, Equalization loss for long-tailed object recognition, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 11662–11671.

[55] H. Zhang, M. Cisse, Y.N. Dauphin, D. Lopez-Paz, mixup: Beyond empirical risk minimization, in: International Conference on Learning Representations, 2018.

[56] B. Zhou, A. Lapedriza, A. Khosla, A. Oliva, A. Torralba, Places: A 10 million image database for scene recognition, IEEE Trans. Pattern Anal. Mach. Intell. 40 (6) (2017) 1452–1464.

[57] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 770–778.

[58] Y. Kim, Convolutional neural networks for sentence classification, in: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2014, pp. 1746–1751.

[59] J. Pennington, R. Socher, C.D. Manning, Glove: Global vectors for word representation, in: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2014, pp. 1532–1543.
[60] Y. Yang, Z. Xu, Rethinking the value of labels for improving class-imbalanced learning, Adv. Neural Inf. Process. Syst. 33 (2020) 19290–19301.
[61] G.R. Kini, O. Paraskevas, S. Oymak, C. Thrampoulidis, Label-imbalanced and group-sensitive classification under overparameterization, Advances in Neural Information Processing Systems 34.
[62] X. Wang, L. Lian, Z. Miao, Z. Liu, S. Yu, Long-tailed recognition by routing diverse distribution-aware experts, in: International Conference on Learning Representations, 2020.
[63] Y. Zhang, B. Hooi, L. Hong, J. Feng, Test-agnostic long-tailed recognition by test-time aggregating diverse experts with self-supervision, arXiv preprint arXiv:2107.09249.
[64] J. Cai, Y. Wang, J.-N. Hwang, Ace: Ally complementary experts for solving long-tailed recognition in one-shot, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 112–121.

Shaoyu Zhang received the B.S. degree in Sichuan University in 2019, He is currently a Ph.D. candidate in the Institute of Automation, Chinese Academy of Sciences (CASIA), Beijing, China. His research interests include computer vision and machine learning.

Chen Chen received her M.Sc. and Ph.D degree in Computer Science from University of Copenhagen, Denmark in 2011 and 2013. She is currently an associate professor in Institute of Automation, Chinese Academy of Sciences (CASIA), China. She was a visiting scholar at Stanford University in 2012. Her research interests focus on pattern recognition and machine learning.

Xiyuan Hu received Ph.D degree in Institute of Automation, Chinese Academy of Sciences in 2011. After graduation, he became a member of High Technology Innovation Center (HITIC) at the Institute of Automation, CAS. Since June 2020, he became a full professor at the school of Computer Science and Engineering in Nanjing University of Science and Technology. He was a visiting scholar at Academia Sinica and Harvard University in 2012 and 2014, respectively. His research focus on adaptive signal separation theory, image and video processing. He has published more than 80 journal and conference papers (such as IEEE T-SP, IEEE T-MM, IEEE SPL, CVPR, ECCV, etc.) in these areas. He is coauthor of two books (Chinese Science Press and Tsinghua University Press). In 2020, he won First prize of science and technology award of the Ministry of Public Security of China. In 2019, he won Youth Science and Technology Award of Chinese Society for Imaging Science and Technology (CSIST) and Scientific and Technology Progress Award of Chinese Computer Federation (CCF).

Silong Peng received the B.S. degree in mathematics from the Anhui University in 1993, and the M.S. and Ph. D. degrees in mathematics from Institute of Mathematics, Chinese Academy of Sciences (CAS), in 1995 and 1998, respectively. From 1998 to 2000, he worked as a postdoctoral researcher in the Institute of Automation, CAS. During this period, he was also a visiting scholar with Department of Mechanics and Mathematics, Lomonosov Moscow State University, Russia. In 2000, he became a full professor of signal processing and pattern recognition in Institute of Automation, CAS. His research interests include wavelets, multi-rate signal processing, and digital image processing.