Contents lists available at ScienceDirect

Pattern Recognition

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

NCL++: Nested Collaborative Learning for long-tailed visual recognition

Zichang Tan [a],[*],[1], Jun Li [b],[c],[1], Jinhao Du [a], Jun Wan [b],[c], Zhen Lei [b],[c], Guodong Guo [a]

[a] Baidu Inc., Beijing, China
[b] Institute of Automation, Chinese Academy of Sciences (CASIA), Beijing, China
[c] School of Artificial Intelligence, University of Chinese Academy of Sciences (UCAS), Beijing, China

A B S T R A C T

Long-tailed visual recognition has received increasing attention in recent years. Due to the extremely imbalanced data distribution in long-tailed learning, the learning process shows great uncertainties. For example, the predictions of different experts on the same image vary remarkably despite the same training settings. To alleviate the uncertainty, we propose a Nested Collaborative Learning (NCL++) which tackles the long-tailed learning problem by a collaborative learning. To be specific, the collaborative learning consists of two folds, namely inter-expert collaborative learning (InterCL) and intra-expert collaborative learning (IntraCL). InterCL learns multiple experts collaboratively and concurrently, aiming to transfer the knowledge among different experts. IntraCL is similar to InterCL, but it aims to conduct the collaborative learning on multiple augmented copies of the same image within the single expert. To achieve the collaborative learning in long-tailed learning, the balanced online distillation is proposed to force the consistent predictions among different experts and augmented copies, which reduces the learning uncertainties. Moreover, in order to improve the meticulous distinguishing ability on the confusing categories, we further propose a Hard Category Mining (HCM), which selects the negative categories with high predicted scores as the hard categories. Then, the collaborative learning is formulated in a nested way, in which the learning is conducted on not just all categories from a full perspective but some hard categories from a partial perspective. Extensive experiments manifest the superiority of our method with outperforming the state-of-the-art whether with using a single model or an ensemble. The code will be publicly released.

1. Introduction

In recent years, deep neural networks have achieved resounding success in various visual tasks, e.g., image classification [1,2], object detection [3,4], semantic segmentation [5,6] and so on. Despite the advances in deep technologies and computing capability, the huge success also highly depends on large well-designed datasets of having a roughly balanced distribution, such as ImageNet [7], MS COCO [8] and Places [9]. This differs notably from real-world datasets, which usually exhibits long-tailed data distributions [10–18] where few head classes occupy most of the data while many tail classes only have few samples as shown in Fig. 1. In such scenarios, the model is easily dominated by those few head classes, whereas low accuracy rates are usually achieved for many other tail classes. Undoubtedly, the long-tailed characteristics challenge deep visual recognition, and also immensely hinder the practical use of deep models.

In long-tailed visual recognition, most early works focus on designing the class re-balancing strategies [10,19–23] and decoupled learning [24,25]. Those methods can accomplish some accuracy improvements but still cannot deal with the long-tailed class imbalance problem well. For example, class re-balancing methods are often confronted with risk of overfitting. More recent efforts aim to improve the long-tailed learning by using multiple experts [26–30]. The multi-expert algorithms follow a straightforward idea of making them diverse from each other. To achieve this, some works [26,30] force different experts to focus on different aspects. For example, LFME [26] formulates a network with three experts and it forces each expert to learn samples from one of head, middle and tail classes. Besides, there are some works [27] adopt some constraint losses to diversify the multiple experts. For example, RIDE [27] proposes a distribution-aware diversity loss to achieve this. The previous multi-expert methods can achieve good performance mainly due to the diversity among them. Also because of this, these methods obtain the predictions through the ensemble rather than a single expert due to that only employing a single expert hardly achieves reliable recognition.
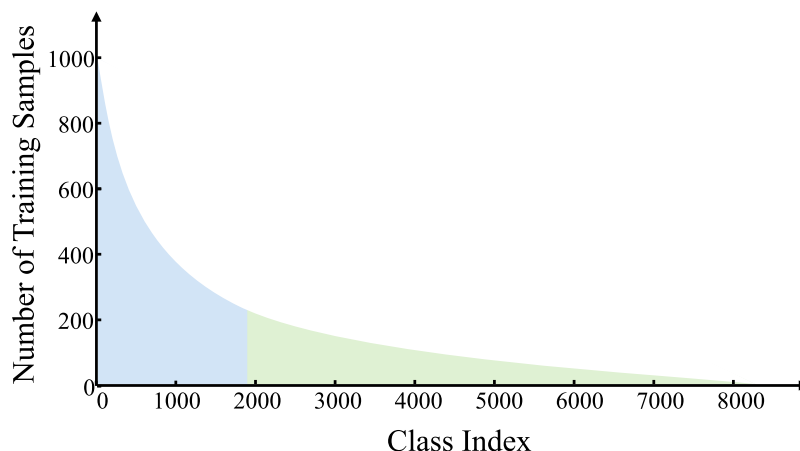

Fig. 1. An illustration of long-tailed data distribution. In the distribution, few head classes occupy most of the data while many tail classes have only a few samples.

However, we found that there are great uncertainties in the long-tailed learning. The learning uncertainties mainly come from two aspects. One is that the same expert performs differently for the same image with different augmentations. The other is that, for two experts with the same configuration and the same training settings, they still vary greatly on predictions with respect to the same input image. To analyze this, we visualize the Kullback–Leibler (KL) distance[2] over the predictions generated by two experts with the same training settings, and the predictions generated by two augmented image copies with respect to the same expert. As shown in Fig. 2(b), the green color shows the KL distance distribution for two different experts of the same structure and training settings with taking the same image as the input, and the blue color shows that of the same expert with taking different augmentations of the same image. Specifically, to analyze the uncertainties across all categories, the Kullback–Leibler (KL) distance is specifically calculated for all categories with reporting the average distance over all images for each class. As we can see, the predictions vary greatly especially in tail classes, which signifies the great uncertainty in the learning process as well as the diversity in models. To show the prediction differences in details, we visualize the detailed predictions of the two experts with the same structure and training settings on a randomly selected sample (corresponding to the green color in Fig. 2(b)). According to previous works [26–30], a simple ensemble like averaging the predictions could be employed to reduce the uncertainty in the prediction stage, which can achieve a certain performance improvement. However, an ensemble of multiple experts also brings huge amount of calculation and parameters, which limits its practical use. This brings up a question, *can we utilize these uncertainties to improve the performance of single model?*

The goal of this paper is to enhance the performance of a single model via reducing the learning uncertainty by the collaborative learning. In the multi-expert framework, owing to the uncertainty in the learning process (e.g., the parameter initialization), different experts capture various knowledge. Besides, the learning uncertainty of the model itself can also lead to various predictions when adding some noises or augmentations to the input. To reduce the uncertainty in the learning process, the Inter-expert Collaborative Learning (InterCL) is employed to collaboratively learn multiple experts concurrently, where each expert serves as a teacher to teach others and also as a student to learn extra knowledge from others. In this way, all experts tend to learn the consistent and right knowledge. In addition to the uncertainty among different experts, the predictions of the same expert with taking different augmentations of the same image as the input also show great uncertainties. Grounded in this, we augment the image by multiple times and then input them to the same expert. Later, the Intra-expert Collaborative Learning (IntraCL) is further employed to distill knowledge among different augmentations and also reduce the learning

uncertainty among them. Specifically, we employ the online distillation to achieve the goal of collaborative learning. In order to better cope with the long-tailed data distribution, we further formulate the online distillation in a balanced way, i.e., balanced online distillation. Specifically, in the proposed balanced online distillation, the contributions of tail classes would be strengthened while that of head classes would be suppressed.

Previous works [31–33] simply conduct the classification and distillation from a full perspective on all categories. It helps the network to obtain the global discrimination on all categories, but lacks of meticulous distinguishing ability on the confusing categories. In the classification problem, the network should pay attention to the hard negative categories with high predictive scores rather than simply treating all categories equally, which helps to reduce the confusion between the target category and the confusing categories. To achieve this, we first propose a Hard Category Mining (HCM) to select the hard categories for each sample, in which the hard category is defined as the negative category with a high predictive score. Then, we formulate a nested learning for the supervised individual learning of a single network as well as the collaborative learning (including IntraCL and InterCL). To the specific, the nested learning conducts the supervised learning or distillation from two perspectives, where one is the full perspective on all categories and the other is the partial perspective only on the selected hard categories. In this way, not only the global discriminative capability but also the meticulous distinguishing ability can be captured.

The proposed method utilizes the collaborative learning for long-tailed visual recognition. The proposed collaborative learning is two folds, where one is inter-expert collaborative learning and the other is intra-expert collaborative learning. The collaborative learning allows the knowledge transferring among different experts and image augmentations, which aims to reduce the learning uncertainties and promotes each expert model to achieve better performance. Our method is inherently a multi-expert framework. However, unlike previous works that need an ensemble of all experts to obtain the final predictions, our method can achieve the state-of-the-art performance only based on a single expert (better performance can be achieved based on the ensemble of all). This is because we collaboratively learn multiple experts by reducing the diversity of models rather than increasing their diversity as in previous works [26–30]. Our contributions can be summarized as follows:

- We propose a Nested Collaborative Learning (NCL++) to improve long-tailed visual recognition via collaborative learning. The collaborative learning consists of two folds, namely intra-expert and inter-expert collaborative learning, which allow the knowledge transferring among different image augmented copies
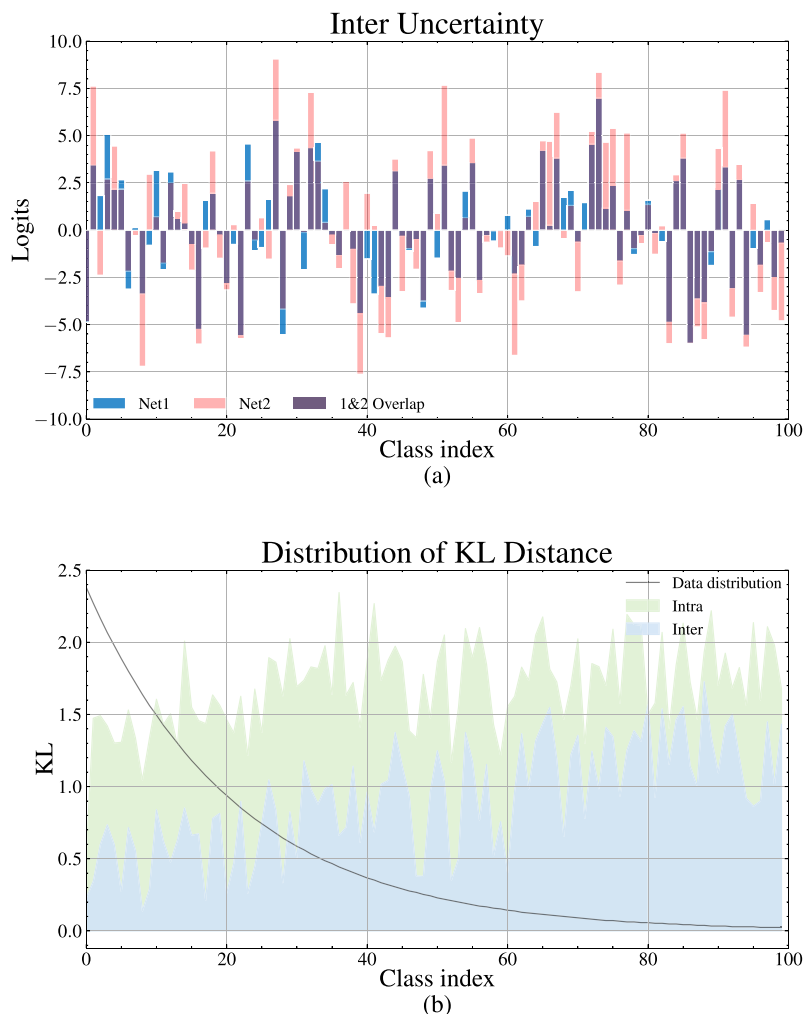
Fig. 2. (a) The Kullback–Leibler (KL) distance calculated from two aspects. The green color indicates the KL distance of two experts with taking the same image as the input. The blue color indicates the KL distance of an expert with taking two augmented image copies with respect to the same image as the inputs. (b) An illustration of the predictions produced by two experts with respect to the same input. The two experts have the same structure and are trained with the same settings. The analysis is conducted on CIFAR100-LT dataset with an Imbalanced Factor (IF) of 100. The predictions are visualized on the basis of a random selected example, and the KL distance is computed based on the whole test set and then the average results of each category are counted and reported. The predictions differ largely from each other between different networks and different augmented images. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

and experts, respectively. To our best knowledge, **it is the first work to adopt the collaborative learning to address the problem of long-tailed visual recognition**.

- We propose a Nested Feature Learning (NFL) to conduct the learning from both a full perspective on all categories and a partial perspective of focusing on hard categories. This helps the model to capture meticulous distinguishing ability on confusing categories.
- We propose a Hard Category Mining (HCM) to select hard categories for each sample.
- The proposed method gains significant performance over the state-of-the-art on five popular datasets including CIFAR-10/100-LT, Places-LT, ImageNet-LT and iNaturalist 2018.

2. Related work

2.1. Long-tailed visual recognition

To alleviate the long-tailed class imbalance, lots of studies [26, 27,34–37] have been conducted in recent years. The existing methods for long-tailed visual recognition can be roughly divided into three categories: class-rebalancing [10,19–23], multi-stage training [24,25]

and multi-expert methods [26–30]. In the following, we will review the methods in the above three categories.

Class re-balancing, which aims to re-balance the contribution of each class during training, is a classic and widely used method for long-tailed learning. Usually, there are two types for class re-balancing. The first one is data re-sampling [24,37–40], including over-sampling [38], under-sampling, square-root sampling [24] and progressively-balanced sampling [24]. The goal of data re-sampling is to increase the sampling probability of samples in tail classes while weakening that of samples in head classes. The second one is loss re-weighting [36,37,41–44], that is, re-weighting of loss function by the numbers of different classes. The popular methods in loss re-weighting include Focal loss [41], Seesaw loss [42], Balanced Softmax Cross-Entropy (BSCE) [36], Class-Dependent Temperatures (CDT) [45], LDAM loss [25] and Equalization loss [43]. Class re-balancing improves the overall performance but usually at the sacrifice of the accuracy on head classes.

Multi-stage training methods divide the training process into several stages [24,46–49]. For example, Kang et al. [24] decouple the training procedure into representation learning and classifier learning, where the representation learning adopts the instance-balanced sampling to learn a good feature extractor and the classifier learning adopts the class-balanced sampling to re-adjust the classifier. Besides, some other

works [46–48,50] tend to improve performance via a post-process of shifting model logits. Li et al. [49] propose a four-stage training strategy on basis of knowledge distillation, self-supervision and decoupled learning. To be specific, the network is first trained with classification and self-supervised losses, and then the feature extractor is frozen and the classifier is re-adjusted by the class-balanced sampling. Later, an additional network is employed to distill the knowledge from the previously trained network, and finally, the network is further trained by re-adjusting the classifier. As we can see, multi-stage training methods may rely on heuristic design.

More recently, multi-expert frameworks [26–30,34,39,51] receive increasing concern, e.g., Learning From Multiple Expert (LFME) [26], Bilateral-Branch Network (BBN) [34], RoutIng Diverse Experts (RIDE) [27], Test-time Aggregating Diverse Experts (TADE) [30] and Ally Complementary Experts (ACE) [29]. For example, LFME [26] formulates three experts with each corresponding to one of head, medium and tail classes. RIDE [27] proposes a distribution-aware diversity loss to the multi-expert network and it encourages the experts to be diverse from each other. Multi-expert methods indeed improve the recognition accuracy for long-tailed learning, but those methods still need to be further exploited. For example, most current multi-expert methods employ different models to learn knowledge from different aspects, while the mutual supervision among them is deficient. Moreover, they often employ an ensemble to produce predictions, which leads to an increase in complexity.

2.2. Knowledge distillation

Knowledge distillation is a prevalent technology in knowledge transferring. One typical manner of knowledge distillation is teacher–student learning [52,53], which transfers knowledge from a large teacher model to a small student model. Current methods in knowledge distillation can be divided into three categories: offline distillation [52, 54–57], online distillation [31,33,58,58–60] and self-distillation [61–65]. Early methods [52,55] often adopt an offline learning strategy, which transfers the knowledge from a pretrained teacher model to a student model. Most works [52,55,66] distill the knowledge from the output distributions, while some works achieve the knowledge transferring by matching feature representations [54] or attention maps [57]. The offline distillation is very popular in the early stage. However, the offline way only considers transferring the knowledge from the teacher to the student, and therefore, the teacher normally should be a more complex high-capacity model than the student. In recent years, knowledge distillation has been extended to an online way [31,33,58,58,59,67], where the whole knowledge distillation is conducted in an one-phase and end-to-end training scheme. For example, in Deep Mutual Learning [33], any one model can be a student and can distill knowledge from all other models. Zhu et al. [32] propose a multi-branch architecture with treating each branch as a student to further reduce computational cost. Compared with offline distillation, online distillation is more efficient by taking the learning in an one-phase end-to-end scheme. For self-distillation [61–65], it can be regarded as a special case in online distillation, where the teacher and the student refer to the same network. In other words, self-distillation means that the model distills the knowledge from itself. For example, Zhang et al. [62] divide the network into several sections according to their depth, and allow the low sections to distill the knowledge from high sections. Our work aims to reduce the predictive uncertainty in long-tailed learning by online distillation and self-distillation. However, different from previous works, we formulate the distillation in a nested way to focus on all categories as well as some important categories, which facilitates the network to obtain the meticulous distinguishing ability.

3. Methodology

The proposed NCL++ aims to adopt the collaborative learning to reduce the learning uncertainties in long-tailed visual recognition as shown in Fig. 3. We first introduce the Balanced Individual Learning (BIL), Intra-expert Collaborative Learning (IntraCL) and Inter-expert Collaborative Learning (InterCL). Then, we present the Hard Category Mining (HCM) of selecting hard categories, and further give a detailed introduction of the Nested Feature Learning (NFL). Finally, we show the overall loss of how to aggregate them together.

3.1. Balanced individual learning

We denote the training set with n samples as $\mathcal{D} = \{\mathbf{x}_i, y_i\}$, where $\mathbf{x}_i$ indicates the ith image sample and y_i denotes the corresponding label. Assume a total of K experts are employed and the kth expert model is parameterized with θ_k. Given image $\mathbf{x}_i$, the predicted probability of class-j in the kth expert is computed as:

$$\tilde{\mathbf{p}}_j(\mathbf{x}_i; \theta_k) = \frac{exp(z_{ij}^k)}{\sum_{l=1}^{C} exp(z_{il}^k)} \tag{1}$$

where z_{ij}^k is the model's class-j output and C is the number of classes. This is a widely used way to compute the predicted probability, and some losses like Cross Entropy (CE) loss is computed based on it. However, this way does not consider the data distribution, and is not suitable for long-tailed visual recognition, where a vanilla model based on $\tilde{\mathbf{p}}(\mathbf{x}_i; \theta_k)$ would be largely dominated by head classes. Therefore, some researchers [36] proposed to compute predicted probability of class-j in a balanced way:

$$\mathbf{p}_j(\mathbf{x}_i; \theta_k) = \frac{n_j exp(z_{ij}^k)}{\sum_{l=1}^{C} n_l exp(z_{il}^k)} \tag{2}$$

where n_j is the total number of samples of class j. In this way, contributions of tail classes are strengthened while contributions of head classes are suppressed. Based on such balanced probabilities, Ren et al. [36] further proposed a Balanced Softmax Cross Entropy (BSCE) loss to alleviate long-tailed class imbalance in model training. We also adopt the BSCE to conduct the individual learning for each expert, which ensures that each one can achieve the strong discrimination ability. Mathematically, the loss of the balanced individual learning for expert-k can be denoted as:

$$L_{bil}^{\mathcal{G}} = -\sum_{k} log(\mathbf{p}_{y_i}(\mathbf{x}_i; \theta_k)) \tag{3}$$

where θ_k indicates the parameters of the kth expert. The superscript $\mathcal{G}$ indicates the loss is conducted from a global view on all categories.

3.2. Inter-expert collaborative learning

To collaboratively learn multiple experts from each other, we employ the Inter-expert Collaborative Learning (InterCL) to allow each model to learn extra knowledge from others. We employ the knowledge distillation to allow the knowledge transferring among them, and following previous works [31,33], the Kullback Leibler (KL) divergence is employed to calculate the loss, which can be denoted as:

$$L_{inter}^{\mathcal{G}} = \frac{2}{K(K-1)} \sum_{k}^{K} \sum_{q \neq k}^{K} KL(\mathbf{p}(\mathbf{x}_i; \theta_k) || \mathbf{p}(\mathbf{x}_i; \theta_q)) \tag{4}$$

where K denotes the number of experts and $KL(\mathbf{p}||\mathbf{q})$ indicates the KL divergence between the distributions $\mathbf{p}$ and $\mathbf{q}$, which can be specifically denoted as $KL(\mathbf{p}||\mathbf{q}) = \sum_j \mathbf{p}_j log(\frac{\mathbf{p}_j}{\mathbf{q}_j})$. Note that we adopt the balanced probabilities $\mathbf{p}$ for distillation, which increases the contributions of tail classes while suppressing that of head classes. In this way, each expert can be the teacher to teach others, and each expert also can be a student to learn knowledge from others. Such collaborative learning aims to
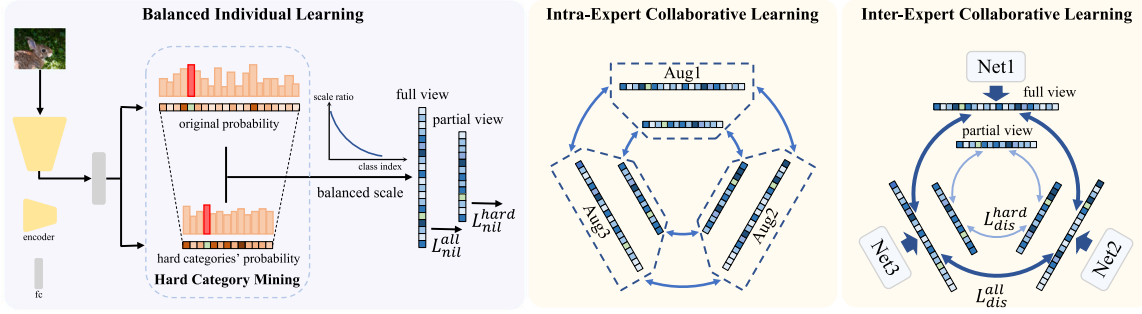
Fig. 3. An illustration of our proposed NCL++ of containing three experts and three augmented image copies for conducting collaborative learning. The proposed NCL++ contains five core components, namely Balanced Individual Learning (BIL), Intra-expert Collaborative Learning (IntraCL), Inter-expert Collaborative Learning (InterCL), Hard Category Mining (HCM) and Nested Feature Learning. The BIL aims to enhance discriminative ability of a single expert. Both IntraCL and InterCL aim to reduce the learning uncertainty by collaborative learning and thus improve the discriminative capability. For the proposed HCM, it selects the hard negative categories, which are used for improving the meticulous distinguishing ability of the model. Based on the selected hard categories, the NFL is further employed to learning features from a global view on all categories and also a partial view on hard categories.

reduce the prediction uncertainties among multiple experts through minimizing the KL divergence. Note that the distillation is also set in a balanced way, which helps the model to learn better for long-tailed visual recognition.

3.3. Intra-expert collaborative learning

The learning uncertainty exists not only in different experts, but also in the same expert which takes different augmented images as the input. In other words, given an input image, if we augment it by several times and then feed them into the network, the corresponding outputs also vary greatly. Inspired by this, we further propose the Intra-Expert Collaborative Learning (IntraCL), which helps the network to alleviate the uncertainty within the expert. To be specific, for an input image $\mathbf{x}_i$, assume it will be augmented for T times (like using RandAugment [68] or other data augmentation strategies), and the augmented images are denoted as $\{\mathbf{x}_i^t\}_{t=1}^T$. All the augmented images are fed to the expert-k and produce the corresponding balanced probabilities $\{\mathbf{p}(\mathbf{x}_i^t; \theta_k)\}_{t=1}^T$. Then, the knowledge distillation is performed to guide the network learning reliable and consistent outputs for different augmented copies. Moreover, the distillation also allows the knowledge transferring among the different augmented copies, which helps them to learn extra knowledge from each other. Similarly, KL divergence is adopted for loss calculation, and the corresponding loss of expert-k can be represented as:

$$L_{intra}^{\mathcal{G},k} = \frac{2}{T(T-1)} \sum_{t}^{T} \sum_{\tau \neq t}^{T} KL(\mathbf{p}(\mathbf{x}_i^t; \theta_k) || \mathbf{p}(\mathbf{x}_i^\tau; \theta_k)) \tag{5}$$

Since the proposed method is a multi-expert framework, the proposed IntraCL would be applied to all experts. The summed IntraCL loss on all experts is represented as:

$$L_{intra}^{\mathcal{G}} = \sum_{k} L_{intra}^{\mathcal{G},k} \tag{6}$$

3.4. Hard category mining

In representation learning, one well-known and effective strategy to boost performance is Hard Example Mining (HEM) [69]. HEM claims that different examples have different importance, specifically, hard samples are more important than easy samples. Therefore, HEM selects hard samples for training, while discarding easy samples which contribute very little and are even detrimental to features learning. However, directly applying HEM to long-tailed visual recognition may distort the data distribution and make it more skewed in long-tailed learning. Although it cannot be directly used for long-tailed learning,
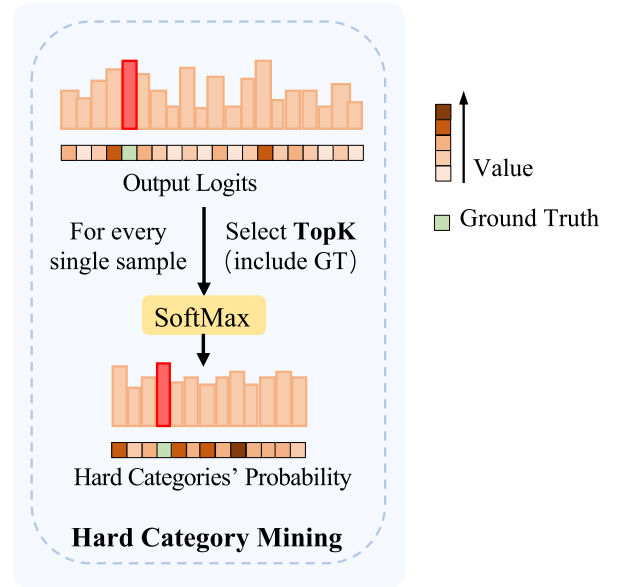

Fig. 4. An illustration of the proposed HCM module.

it still gives us an important inspiration, that is, *do different categories have different degrees of importance in the training process?*

When conducting the classification, the image may be wrongly classified to some confusing categories, like the categories similar to the target category. Therefore, the network should pay close attention to those confusing categories. The confusing category is also called as the hard category in this paper. To be specific, the hard categories are defined as the categories that are not the ground-truth category but with high predicted scores. With the help of focusing on hard categories, the ability of distinguishing the sample from the negative hard categories can be explicitly improved. As shown in Fig. 4, we propose a Hard Category Mining (HCM) to select the hard categories and the target category out, and re-calculate the probabilities over those categories, which are used to be trained in the NFL to increase the ability of distinguishing the sample from the negative hard categories. To be specific, assume we have C categories in total and suppose C_{hard} categories are selected to focus on. For the sample $\mathbf{x}_i$ and expert k, the corresponding set $\boldsymbol{\Psi}_i^k$ containing the outputs of selected categories is denoted as:

$$\boldsymbol{\Psi}_i^k = TopHard\{z_{ij}^k | j \neq y_i\} \cup \{z_{iy_i}^k\} \tag{7}$$

where $TopHard$ means selecting C_{hard} examples with largest values. In order to better adapt to long-tailed learning, we compute the probabilities of the selected categories in a balanced way, which is shown as:

$$\mathbf{p}^*(\mathbf{x}_i; \theta_k) = \{ \frac{n_j exp(z_{ij}^k)}{\sum_{z_{il}^k \in \Psi_i^k} n_l exp(z_{il}^k)} | z_{ij}^k \in \Psi_i^k \}. \tag{8}$$

3.5. Nested feature learning

The nested feature learning means that the feature learning is conducted on both the global view of all categories and the partial view of some important categories. Following this idea, the nested feature learning is conducted on the individual learning as well as the intra-expert and inter-expert collaborative learning. For the individual learning, formulating the learning in a nested way helps the network achieve a global and robust learning on all categories, and also enhances the meticulous distinguishing ability to distinguish the input sample from the most confusing categories. To be specific, the loss on the selected important categories for the balanced individual learning is denoted as:

$$L_{bil}^P = -\sum_k log(\mathbf{p}_{y_i}^*(\mathbf{x}_i; \theta_k)). \tag{9}$$

The whole loss of the balanced individual learning is the sum of the global part L_{bil}^G and partial part L_{bil}^P, which is illustrated as the following formula:

$$L_{bil} = L_{bil}^G + L_{bil}^P. \tag{10}$$

The intra-expert and inter-expert collaborative learning also can be conducted from the global view on all categories and a partial view on the selected important categories. For InterCL, the corresponding distillation on the selected important categories can be written as:

$$L_{inter}^P = \frac{2}{K(K-1)} \sum_k^K \sum_{q \neq k}^K KL(\mathbf{p}^*(\mathbf{x}_i; \theta_k) || \mathbf{p}^*(\mathbf{x}_i; \theta_q)). \tag{11}$$

Similarly, the whole loss of the InterCL is as following:

$$L_{inter} = L_{inter}^G + L_{inter}^P. \tag{12}$$

The term L_{inter}^G transfers the global and structured knowledge from each all categories, and the other term L_{inter}^P focuses on some important categories in which the experts vary greatly in predictions. For the IntraCL, the distillation on some selected important categories can be formulated in a similar way, and we denote it as L_{intra}^P. Therefore, the IntraCL with a nested form can be formulated as:

$$L_{intra} = L_{intra}^G + L_{intra}^P. \tag{13}$$

3.6. Model training

The overall loss in our proposed method consists of three parts: the loss L_{bil} for learning each expert individually to enhance model's discriminative capability, the loss L_{intra} and L_{inter} for cooperation among different augmented images and different experts, respectively. The overall loss L is formulated as:

$$L = L_{bil} + \lambda_1 L_{intra} + \lambda_2 L_{inter} \tag{14}$$

where λ_1 and λ_2 denote the weighting coefficients for IntraCL and InterCL, respectively. Considering both IntraCL and InterCL are collaborative learning, we set their coefficients to the same and denote it as λ (i.e., $\lambda = \lambda_1 = \lambda_2$). Those losses are optimized cooperatively and concurrently, which enhances the discriminative capability of the network and reduces the uncertainties in predictions. Moreover, our method is essentially a multi-expert framework, and therefore, the prediction in the test stage can be obtained by each single network or an ensemble of all of them, one for high efficiency and one for high performance.

4. Experiments

4.1. Datasets and protocols

We conduct experiments on five widely used datasets, including CIFAR10-LT [21], CIFAR100-LT [21], ImageNet-LT [11], Places-LT [9], and iNaturalist 2018 [70].

CIFAR10-LT and CIFAR100-LT [21] are created from the original balanced CIFAR datasets [71]. Specifically, the degree of data imbalance in datasets is controlled by the use of an Imbalance Factor (IF), which is defined by dividing the number of the most frequent category by that of the least frequent category. The imbalance factors of 100 and 50 are employed in these two datasets. **ImageNet-LT** [11] is sampled from the popular ImageNet dataset [7] under long-tailed setting following the Pareto distribution with power value α=6. ImageNet-LT contains 115.8 K images from 1,000 categories. **Places-LT** is created from the large-scale dataset Places [9]. This dataset contains 184.5 K images from 365 categories. **iNaturalist 2018** [70] is the largest dataset for long-tailed visual recognition. iNaturalist 2018 contains 437.5 K images from 8,142 categories, and it is extremely imbalanced with an imbalanced factor of 512.

According to previous works [21,24] the top-1 accuracy is employed for evaluation. Moreover, for iNaturalist 2018 dataset, we follow the works [24,29] to divide classes into many (with more than 100 images), medium (with $20 \sim 100$ images) and few (with less than 20 images) splits, and further report the results on each split.

4.2. Implementation details

For CIFAR10/100-LT, following [25,37], we adopt ResNet-32 [1] as our backbone network and liner classifier for all the experiments. Input images are randomly cropped with size 32×32. We utilize ResNet-50 [1], ResNeXt-50 [72] as our backbone network for ImageNet-LT, ResNet-50 for iNaturalist 2018 and pretrained ResNet-152 for Places-LT respectively, based on [11,24,73]. Following [48], the cosine classifier is utilized for these models. Following previous works [24,34,73], we resize the input image to 256×256 pixels and take a 224×224 crop from the original image or its horizontal flip. We use the SGD with a momentum of 0.9 and a weight decay of 2×10^{-4} as the optimizer to train all the models. As for experiments on CIFAR10/100-LT, the initial learning rate is 0.1 and decreases by 0.1 at epoch 320 and 360, respectively. The learning rate for Places-LT is 0.02 and decreases by 0.1 at epoch 10 and 20. For the rest datasets, the initial learning rate is set to 0.2 and decays by a cosine scheduler to 1×10^{-4}. Unlike the previous works PaCo [73] and NCL [74], we utilize less training epochs for ImageNet-LT and iNaturalist 2018, which is 200. And the training epochs for Places-LT is 30, same as previous works [24,73]. Cumulative gradient [75] is utilized for ImageNet-LT and iNaturalist 2018. Specifically, the training batch size is set to 128, but the gradient keeps accumulating and the parameters are updated every two epochs. This trick uses less GPU memory to achieve similar training results as batch size of 256. Two experts are utilized for all datasets. Two augmented copies are utilized for ImageNet-LT and four for the rest. In addition, for fair comparison, following [73], RandAugment [68] is also used for all the experiments. The influence of RandAugment will be discussed in detail in the Section 4.4. These models are trained on 8 NVIDIA Tesla V100 GPUs. The $\beta = C_{hard}/C$ in HCM is set to 0.3, and the coefficient of InterCL and IntraCL loss λ is set to 0.6. The influence of β and λ will be discussed in detail in Section 4.4.

4.3. Comparisons to prior arts

We compare the proposed NCL++ with previous state-of-art methods, like BBN [34], RIDE [27], NCL [74]. Besides, the baseline result of single network with using BSCE loss is also reported. Two experts is the default setting for NCL++, which has smaller model size than

Table 1
Comparisons on CIFAR100-LT and CIFAR10-LT datasets with the IF of 100 and 50.

Method	Ref.	CIFAR100-LT		CIFAR10-LT	
		100	50	100	50
CB Focal loss [21]	CVPR'19	38.7	46.2	74.6	79.3
LDAM+DRW [25]	NeurIPS'19	42.0	45.1	77.0	79.3
LDAM+DAP [76]	CVPR'20	44.1	49.2	80.0	82.2
BBN [34]	CVPR'20	39.4	47.0	79.8	82.2
LFME [26]	ECCV'20	42.3	–	–	–
CAM [37]	AAAI'21	47.8	51.7	80.0	83.6
Logit Adj. [47]	ICLR'21	43.9	–	77.7	–
Xu et al. [77]	NeurIPS'21	45.5	51.1	82.8	84.3
LDAM+M2m [78]	CVPR'21	43.5	–	79.1	–
MiSLAS [79]	CVPR'21	47.0	52.3	82.1	85.7
LADE [46]	CVPR'21	45.4	50.5	–	–
Hybrid-SC [80]	CVPR'21	46.7	51.9	81.4	85.4
DiVE [81]	ICCV'21	45.4	51.3	–	–
SSD [82]	ICCV'21	46.0	50.5	–	–
PaCo [73]	ICCV'21	52.0	56.0	–	–
xERM [83]	AAAI'22	46.9	52.8	–	–
RISDA [84]	AAAI'22	50.2	53.8	79.9	84.2
Batchformer [85]	CVPR'22	52.4	–	–	–
RIDE (4 experts)[a] [27]	ICLR'21	49.1	–	–	–
ACE (4 experts)[a] [29]	ICCV'21	49.6	51.9	81.4	84.9
TLC (4 experts)[a] [51]	CVPR'22	49.8	–	80.4	–
NCL [74]	CVPR'22	53.3	56.8	84.7	86.8
NCL (3 experts)[a] [74]	CVPR'22	54.2	58.2	85.5	87.3
BSCE (baseline)	–	50.6	55.0	84.0	85.8
NCL++	–	**54.8**	**58.2**	**86.1**	**88.0**
NCL++ (2 experts)[a]	–	**56.3**	**59.8**	**87.2**	**88.8**

[a] Indicates the ensemble performance is reported.

Table 2
Comparisons on ImageNet-LT and Places-LT datasets.

Method	Ref.	ImageNet-LT			Places-LT
		Res10	Res50	ResX50	Res152
OLTR [11]	CVPR'19	34.1	–	–	35.9
BBN [34]	CVPR'20		48.3	49.3	–
NCM [24]	ICLR'20	35.5	44.3	47.3	36.4
cRT [24]	ICLR'20	41.8	47.3	49.6	36.7
τ-norm [24]	ICLR'20	40.6	46.7	49.4	37.9
LWS [24]	ICLR'20	41.4	47.7	49.9	37.6
BSCE [36]	NeurIPS'20	–	–	–	38.7
Xu et al. [77]	NeurIPS'21	42.9	48.4	–	–
DisAlign [48]	CVPR'21	–	52.9	–	–
DiVE [81]	ICCV'21	–	53.1	–	–
SSD [82]	ICCV'21	–	–	56.0	–
PaCo [73]	ICCV'21	–	57.0	58.2	41.2
ALA Loss [50]	AAAI'22	–	52.4	53.3	40.1
xERM [83]	AAAI'22	–	–	54.1	39.3
RISDA [84]	AAAI'22	–	50.7	–	–
MBJ [86]	AAAI'22	–	–	52.1	38.1
WD [87]	CVPR'22	–	53.9	–	–
BatchFormer [85]	CVPR'22	47.6	57.4	–	41.6
RIDE (4 experts)[a] [27]	ICLR'21	–	55.4	56.8	–
MBJ+RIDE (4 experts) [86]	AAAI'22	–	–	57.7	–
ACE (3 experts)[a] [29]	ICCV'21	44.0	54.7	56.6	–
TLC (4 experts)[a] [51]	CVPR'22	–	55.1	–	–
NCL [74]	CVPR'22	46.8	57.4	58.4	41.5
NCL (3 experts)[a] [74]	CVPR'22	47.7	59.5	60.5	41.8
BSCE (baseline)	–	45.7	53.9	53.6	40.2
NCL++	–	**49.0**	**58.0**	**59.1**	42.0
NCL++ (2 experts)[a]	–	**49.9**	**59.6**	**60.9**	**42.4**

[a] Indicates the ensemble performance is reported.

Table 3
Comparisons on iNaturalist 2018 dataset with ResNet-50.

Method	Ref.	iNaturalist 2018			
		Many	Medium	Few	All
OLTR [11]	CVPR'19	59.0	64.1	64.9	63.9
BBN [34]	CVPR'20	49.4	70.8	65.3	66.3
DAP [76]	CVPR'20	–	–	–	67.6
NCM [24]	ICLR'20				
cRT [24]	ICLR'20	69.0	66.0	63.2	65.2
τ-norm [24]	ICLR'20	65.6	65.3	65.9	65.6
LWS [24]	ICLR'20	65.0	66.3	65.5	65.9
LDAM+DRW [25]	NeurIPS'19	–	–	–	68.0
Logit Adj. [47]	ICLR'21	–	–	–	66.4
CAM [37]	AAAI'21	–	–	–	70.9
SSD [82]	ICCV'21				
PaCo [73]	ICCV'21	–	–	–	73.2
RIDE (3 experts)[a] [27]	ICLR'21	70.9	72.4	73.1	72.6
ACE (3 experts)[a] [29]	ICCV'21	–	–	–	72.9
ALA Loss [50]	AAAI'22	71.3	70.8	70.4	70.7
xERM [83]	AAAI'22	–	–	–	67.3
RISDA [84]	AAAI'22	–	–	–	69.1
MBJ [86]	AAAI'22	–	–	–	70.0
MBJ+RIDE [86]	AAAI'22	–	–	–	73.2
WD [87]	CVPR'22	71.2	70.4	69.7	70.2
BatchFormer [85]	CVPR'22	65.5	74.5	75.8	74.1
NCL [74]	CVPR'22	**72.0**	**74.9**	73.8	**74.2**
NCL (3 experts)[a] [74]	CVPR'22	**72.7**	**75.6**	74.5	74.9
BSCE (baseline)	–	67.5	72.0	71.5	71.4
NCL++	–	71.6	73.9	**74.7**	74.0
NCL++ (2 experts)[a]	–	72.2	75.3	**75.7**	**75.2**

[a] Indicates the ensemble performance is reported.

NCL, but achieves better performance. Comparisons on CIFAR10/100-LT are shown in Table 1, comparisons on ImageNet-LT and Places-LT are shown in Table 2, and comparisons on iNaturalist 2018 are shown in Table 3. Our proposed method achieves the state-of-the-art performance on all datasets. For only using a single expert for evaluation,

our NCL++ outperforms previous methods on CIFAR10-LT, CIFAR100-LT, ImageNet-LT and Places-LT with accuracies of 86.1% (IF of 50), 54.8% (IF of 100), 58.0% (with ResNet-50) and 42.0%, respectively. Compared with NCL [74], the proposed NCL++ achieves significant improvements on CIFAR-LT (54.8% vs. 53.3%) and ImageNet-LT with ResNet10 (49.0% vs. 46.8%). When further using an ensemble for evaluation, the performance on CIFAR10-LT, CIFAR100-LT, ImageNet-LT, Places-LT and iNaturalist2018 can be further improved to 87.2% (IF of 50), 56.3% (IF of 100), 59.6% (with ResNet-50), 42.4% and 75.2%, respectively. **NCL++ uses fewer experts than NCL (2 experts vs. 3 experts), but achieves better ensemble performance on all the datasets.** Similar to NCL, a single network can be used for evaluation, which will not bring extra computation but still outperforms previous multi-experts method, like ACE, TLC, etc. When using ensemble of two experts, the performance has been further significantly improved.

4.4. Component analysis

Influence of the ratio of hard categories. The ratio of selected hard categories is defined as $\beta = C_{hard}/C$. Experiments on our BIL model are conducted within the range of β from 0 to 1 as shown in Fig. 5(a). The highest performance is achieved when setting β to 0.3. Setting β with a small and large values brings limited gains due to the under and over explorations on hard categories.

Effect of loss weight. To search an appropriate value for λ, experiments on the proposed NCL++ with a series of λ are conducted as shown in Fig. 5(b). λ controls the contribution of knowledge distillation among multiple experts and augmented images in total loss. The best performance is achieved when $\lambda = 0.6$, which shows a balance is achieved between individual learning and collaborative learning. Intuitively, the network achieves marginal improvements with a small or a large λ, which shows that both ignoring and over emphasizing the collaborative learning are not optimal.

Impact of different number of augmented copies in IntraCL. As shown in Fig. 6(a), we take experiments for IntraCL with different
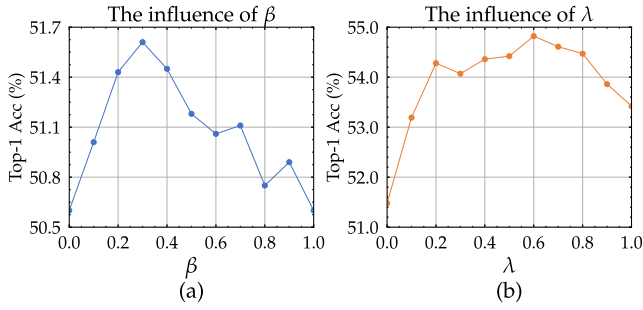
Fig. 5. Parameter analysis of (a) the ratio β and (b) the loss weight λ on CIFAR100-LT dataset with IF of 100.
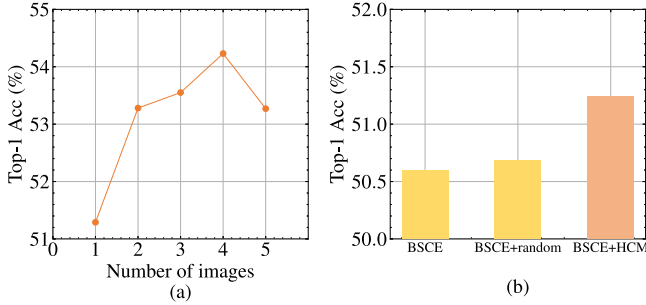

Fig. 6. (a) Comparisons of using different number of augmented image copies in IntraCL with a single network. (b) Comparisons of using hard categories selected by HCM or random categories. All experiments are conducted on CIFAR100-LT with an IF of 100.

Fig. 7. Comparisons of using different expert numbers of InterCL on CIFAR100-LT with an IF of 100. We report the performance on both a single network and an ensemble. Specifically, the performance on a single network is reported as the average accuracy on all experts, and the ensemble performance is computed based on the averaging logits over all experts.

Table 4

Comparisons of training the network with simple augmentation (SimAug), RandAugment (RandAug) and Mixup. Experiments are conducted on CIFAR100-LT dataset with an IF of 100.

Method	SimAug	RandAug	Mixup
Current SOTA	49.1[a] [27]	52.4 [85]	49.4[a] [29]
CE	41.68	44.79	41.85
BSCE	45.69	50.60	49.84
BSCE+NCL++	48.76	54.82	51.37
BSCE+NCL++[a]	51.57	56.33	53.38

[a] Indicates the ensemble performance is reported.

number T of augmented copies. The experiments are taken without using InterCL and thus only a single expert is employed. As we can see, the performance achieves the highest when taking four augmented copies for an image. Besides, $T = 2$ is also a nice choice where the performance can be dramatically improved without a lot of extra computation.

Hard categories vs. random categories. To further verify the effectiveness of the proposed HCM, we also take the experiments with random categories for comparisons as shown in Fig. 6(b) (denoted by 'BSCE + random'). 'BSCE + random' performs slightly better than the baseline method 'BSCE', but its performance is still far worse than 'BSCE + HCM'. It shows that training on hard categories selected by the proposed HCM really helps to improve the discriminative capability of the model.

Impact of different number of experts in InterCL. As shown in Fig. 7, experiments using different number of experts are conducted. The ensemble performance is improved steadily as the number of experts increases, while for only using a single expert for evaluation, its performance can be greatly improved when only using a small number of expert networks, e.g., two experts. Therefore, two experts are mostly employed in our multi-expert framework for a balance between complexity and performance.

Single expert vs. multi-expert. Our method is essentially a multi-expert framework, and the comparison among using a single expert or an ensemble of multi-expert is a matter of great concern. As shown in Fig. 7, the ensemble can perform better than a single network on the performance over all classes and many splits.

Influence of data augmentations. Data augmentation is a common tool to improve performance. For example, previous works [29,37, 79] use Mixup [88] and RandAugment [68] to obtain richer feature representations. Our method follows PaCo [73] to employ RandAugment [68] for experiments. As shown in Table 4, the performance is improved by about 3% to 5% when employing RandAugment for
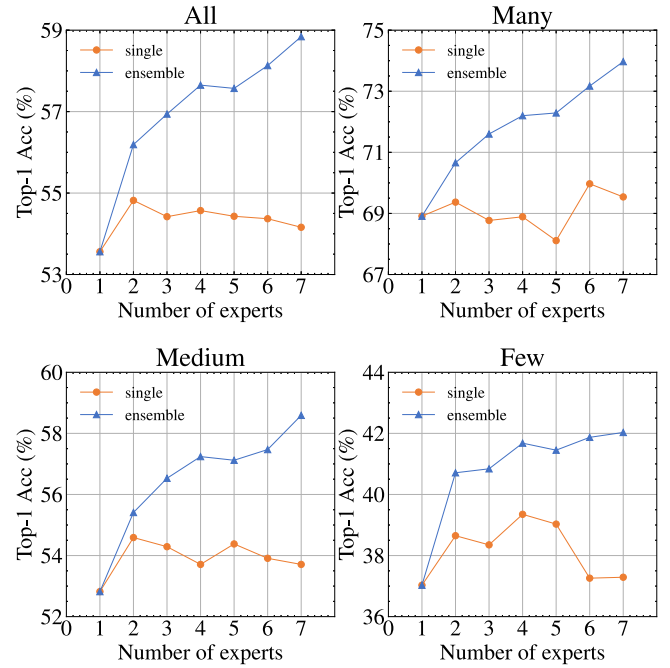
training. However, our high performance depends not entirely on RandAugment. When using Mixup for augmentation, our method achieves the performance of 51.37% and 53.38% for single expert and ensemble of them respectively. Even with the simple data augmentation (random crop, random horizontal flip), our method still performs very well, the ensemble model reaches an amazing performance of 51.57%. The results show the superiority of our method over state-of-the-art methods for both simple and complex data augmentation.

Analysis on computation cost. We compare the computation cost between the proposed method and other state-of-the-art (SOTA) methods as shown in Table 5. For a fair comparison, we set two experts equally for both NCL and NCL++. Due to using multiple networks for collaborative learning, NCL and NCL++ have more computation costs in the training stage. In the test stage, the NCL and its improved version, NCL++, only uses a single network for evaluation and it contains less computation cost but with a higher accuracy compared with other SOTA methods, like NCM [24] and TCL [51]. Moreover, compared with NCL++, NCL contains more computations (with higher GFLOPs) while a lower performance accuracy (57.4% vs. 58.0%). As we know, the only difference between NCL and NCL++ is that the self-supervision module is replaced with the newly proposed IntraCL module. Obviously, by replacing the self-supervision part with IntraCL, the computation cost is largely reduced while the performance is improved.
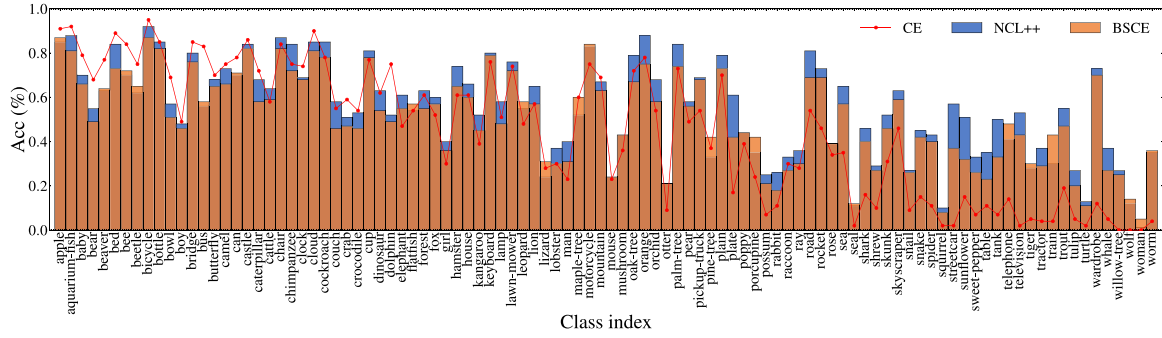
Fig. 8. The accuracy of all categories on CIFAR100-LT with an IF of 100. The horizontal axis represents the details of categories, and the vertical axis represents their accuracies. Note that we sort the categories according to the number of images, where the number of images decreases from the left to right. In other words, the leftmost and the rightmost indicate the category with the maximum and minimum number of images, respectively.

Table 5

Comparisons of computation cost between the proposed method and other state-of-the-art methods. The analysis is conducted on ImageNet-LT dataset with the network of ResNet-50.

Method	FLOPs (G)		Accuracy (%)
	Train	Test	
NCM [24]	4.11	4.11	44.3
TLC [51] (3 experts)	6.55	6.55	55.1
NCL [74] (3 experts)	17.03	4.11[a]	57.4
NCL++ (2 experts)	8.24	4.11[a]	58.0

[a] Evaluating the model with a single expert.

Table 6

Comparisons of using fixed and learnable loss weights. The experiments are conducted on CIFAR100-LT dataset with an IF of 100.

Method	Acc. on train set	Acc. on test set
Fixed	76.7	54.8
Learnable	82.3	54.5

Fixed vs. learnable loss weights. In this paper, we fixedly set the loss weights λ_1 and λ_2 and select the appropriate values for them by a cross-validation (see Fig. 5). The loss weights also can be learn automatically as opposed to the fixed values. In this sub-section, we set loss weights λ_1 and λ_2 as learnable parameters and compare it with the fixed manner. Following the work [61], we normalize the two learnable values λ_1 and λ_2 by a softmax function before using them to aggregate the losses, which ensures that λ_1 and λ_2 are non-negative values and $\lambda_1 + \lambda_2 = 1$. The experimental results are shown in Table 6. As we can see, compared to the fixed weights, using the learnable loss weights can obtain a higher accuracy on the train set but receive a lower accuracy on the test set. It shows that the learnable loss weights may lead to overfitting and setting the parameters as fixed values may achieve better generalizations.

Ablations studies on all components. In this sub-section, we perform detailed ablation studies for our NCL++ on CIFAR100-LT dataset, which is shown in Table 7. To conduct a comprehensive analysis, we evaluate the proposed components including Balanced Individual Learning (BIL), IntraCL, InterCL and the Nested Feature Learning on hard categories (denoted as 'NFL$_{hard}$'). The ablation studies start from a naive method, which does not take any one of above components. To be specific, the naive method is indeed a plain network with taking the Cross-Entropy (CE) loss as the loss function. Compared with the naive method, the BIL improve the performance from 44.79% to 50.60%, which is a considerable improvement. This is because the naive method does not consider anything about the imbalanced data distribution. For the BIL, it dramatically increases the contributions of tail classes while suppressing that of head classes, which could greatly improve the performance by picking low hanging fruit. For other components

Table 7

Ablation studies on CIFAR100-LT dataset with an IF of 100. 'BIL', 'InterCL' and 'IntraCL' represent the individual learning or collaborative learning of the global view on all categories. For 'NFL$_{hard}$', it means adding the nested feature learning of the partial view on hard categories. 'Ensemble' means that the final predictions are obtained by averaging the results on multiple experts. In the proposed NCL++, two experts are employed.

BIL	InterCL	IntraCL	NFL$_{hard}$	Ensemble	Acc.
					44.79
✓					50.60
✓			✓		51.24
✓	✓				53.13
✓		✓			53.94
✓	✓	✓			54.36
✓	✓	✓	✓		54.82
✓	✓	✓	✓	✓	56.33

like IntraCL, InterCL and NFL$_{hard}$, all of them can steadily improve the performance. For example, based on BIL, the proposed InterCL and IntraCL improve the performance by 2.53% and 3.34%, respectively. When both InterCL and IntraCL are employed, this improvement can be enlarged to 3.76%. Then, the NFL$_{hard}$ also gains the improvements from 54.36% to 54.82%. Finally, we take the model ensemble of all experts (two experts are employed) to produce the final predictions, and the performance is further improved to 56.33%. The steadily performance improvements clearly show the effectiveness of the proposed NCL++.

4.5. Discussion and further analysis

KL distance of pre/post collaborative learning. To analyze the uncertainty of pre/post collaborative learning, we visualize the average KL distance of two different augmented copies (see Fig. 9(a)) or two experts (see Fig. 9(b)) with CE, BSCE, NCL and NCL++. As we can see, both CE and BSCE do not consider the uncertainty in long-tailed learning, and thus the corresponding learned experts still show large KL distances among different experts and different image augmented copies. For the NCL, it only considers the learning uncertainty among different experts. Therefore, the KL distance between different image augmented copies is still large in NCL, which shows the intra-expert uncertainty has not been reduced. For the proposed NCL++, the KL distances between different experts and different image augmented copies are greatly reduced, which shows that both the intra- and inter-expert uncertainties are effectively alleviated.

Score distribution of hardest negative category. Deep models normally confuse the target sample with the hardest negative category. Here we visualize the score distribution for the baseline method ('BSCE') and our method ('NCL++') as shown in Fig. 11(a). The higher the score of the hardest negative category, the more likely it is to produce false recognition. The scores in our proposed method are mainly concentrated in the range of 0–0.4, while the scores in the baseline
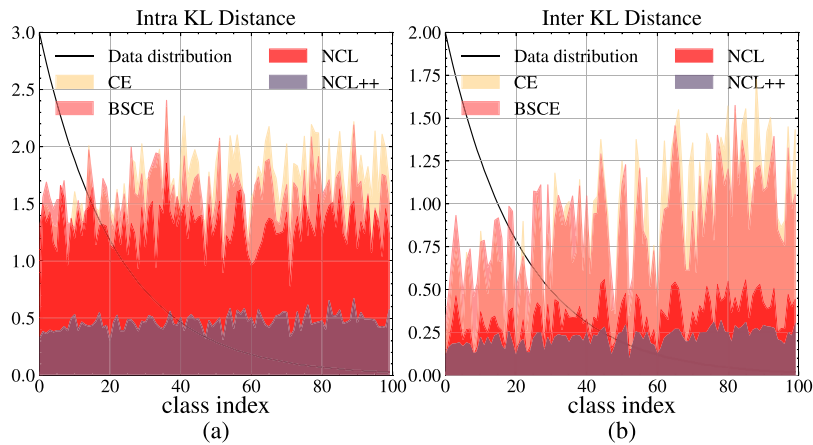
Fig. 9. The average KL distance between (a) the output probabilities of two experts with respect to the same image and (b) the output probabilities of two augmented image copies with respect to the same expert. Analysis is conducted on CIFAR100-LT with an IF of 100.
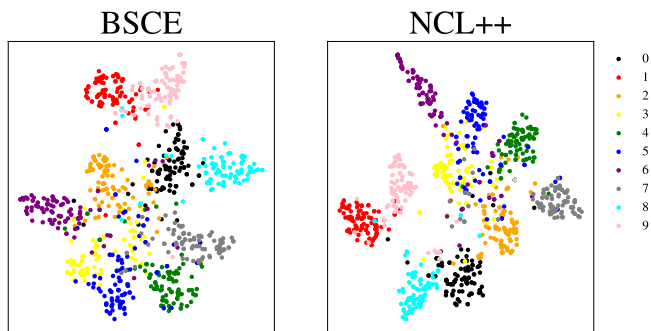

Fig. 10. t-SNE visualization between BSCE and NCL++ on CIFAR10-LT with an IF of 100. Different colors indicate different categories.

Table 8
Comparisons on CIFAR100 dataset.

Method	ResNet-10	ResNet-32
CE	64.45	73.68
NCL	64.79	74.24
NCL++	67.06	76.88

model are distributed in the whole interval (including the interval with large values). This shows that our NCL++ can considerably reduce the confusion with the hardest negative category.

t-SNE visualizations. We use the t-SNE [89] to visualize the features of the baseline method BSCE and the proposed NCL++. Compared to the baseline BSCE, the two-dimensional t-SNE map of the proposed NCL++ seems to be better clustered as shown in Fig. 10, (e.g., the red color). It shows that the proposed NCL++ generate more compact feature representation than the baseline BSCE, where the features of the same category extracted by the network stay closer. The nicely clustered and compact features indicate the model are well-trained for long-tailed visual recognition, which can achieve better classification and recognition performance.

Experiments on balanced datasets. Although NCL++ is proposed based on the phenomenon of long-tailed datasets, except for the re-balanced part, the rest components of NCL++ can still be applied to classification tasks on balanced datasets, such as CIFAR [71]. We conduct experiments on CIFAR100 as shown in Table 8. Cross Entropy (denoted as 'CE') is utilized as the baseline method and the result of the single expert is reported. As we can see from the results, the proposed NCL++ achieves considerable improvement over the baseline. The performance is improved by 2.61% with ResNet-10 and 3.20% with ResNet-32. It is undeniable that even in balanced datasets there are still confusing categories and the uncertainty of model training will still exist. In these scenarios, NCL++ also shows its advantages.

Analysis on all categories. As shown in Fig. 8, we present the accuracy on all categories for three methods, namely the proposed method NCL++, the naive method CE and the baseline method BSCE. Compared with the naive method CE, the proposed method NCL++ can dramatically improve the performance of tail classes, which mainly

benefits from the balanced learning (including balanced individual learning and balanced online distillation). For example, for the sixth category from the last, the accuracy is improved from about 13% to 74%, where over 60% is improved. Moreover, compared with the baseline method BSCE, our NCL++ can improve the performance on almost all categories, which clearly shows the effectiveness of the proposed collaborative learning and NFL.

Visual comparisons to prior arts. As shown in Fig. 11, we compare the proposed method with prior arts on many (with more than 100 images) and few (with less than 20 images) splits. The comparisons on CIFAR100-LT are shown in Fig. 11(b). As we can see, the proposed method achieves remarkable improvements on both many and few splits compared with previous SOTAs (e.g., RIDE and ACE). We take the comparisons on CIFAR100-LT as an example. Many previous methods could perform well on many split (69.2% for RIDE) but perform poorly on few split (only 26.3% for RIDE). For the proposed NCL++, it dramatically improves the accuracy of the few split to 38.7% with a single network through collaborative learning.

InterCL without balancing probability. As shown in Fig. 12(a), when removing the balancing probability in InterCL (denoted as 'w/o balanced scale') both the performance of a single expert and an ensemble decline about 1%, which manifests the importance of employing the balanced probability for the distillation in long-tailed learning.

Offline distillation vs. InterCL. To further verify the effectiveness of our InterCL, we employ an offline distillation for comparisons. The offline distillation (denoted as 'BIL+OffDis') first employs three teacher networks of NIL to train individually, and then produces the teacher labels by using the averaging outputs over three teacher models. The comparisons are shown in Fig. 12(b). Although BIL+OffDis gains some improvements via an offline distillation, but its performance still 1.5% worse than that of BIL+InterCL. It shows that our InterCL of the collaborative learning can learn more knowledge than offline distillation.

Performance vs. data scale. We have conducted more experiments on ImageNet-LT to explore the relationship between performance and dataset size. As shown in Table 9, the performance will increase as the dataset scale increases. Compared to baseline method (BSCE), NCL++ consistently shows improvements across different dada scales.
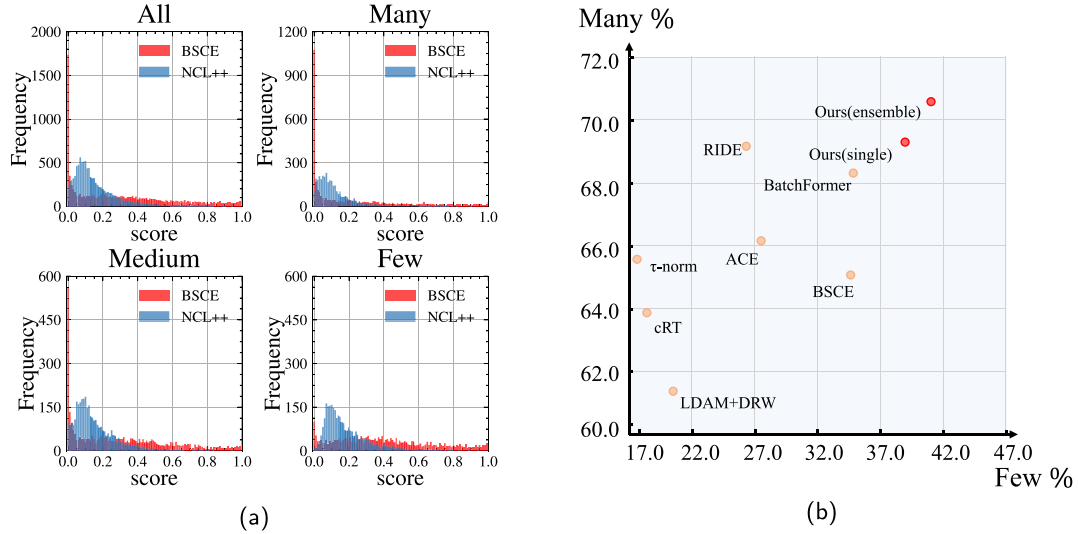
Fig. 11. (a) The distribution of the probability of hardest negative category. (b) Comparisons of our proposed method and some representative methods over many and few splits. Experiments are conducted on CIFAR100-LT with an IF of 100.
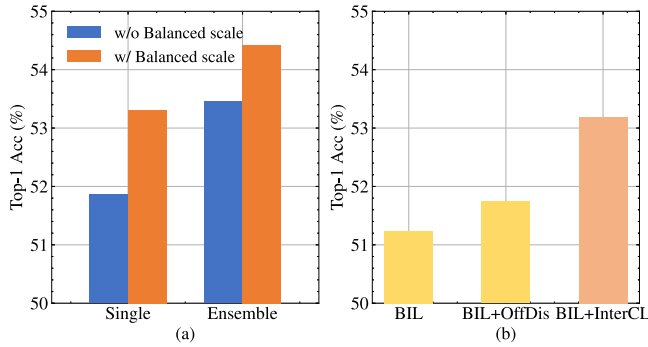

Fig. 12. (a) Comparisons of whether using balanced scale in InterCL. (b) Comparisons of using offline distillation or our InterCL. Analysis is conducted on CIFAR100-LT with an IF of 100.

Table 9
The impact of training data with different ratios. Experiments are conducted on ImageNet-LT with ResNet-10.

Data ratio	Baseline Acc. (%)	NCL++ Acc. (%)
20%	40.64	44.24 (+3.60)
40%	44.04	47.93 (+3.89)
60%	44.66	48.05 (+3.39)
80%	44.91	48.70 (+3.79)
100%	45.70	49.00 (+3.30)

Table 10
Comparisons between our NCL++ and other contrastive learning methods. Experiments are conducted on CIFAR100-LT with an IF of 100.

Method	Acc. (%)
Hybrid-PSC [80]	45.0
PaCo [73]	52.0
NCL [74]	53.3
NCL++ (ours)	54.8

For example, our NCL++ improves the performance by 3.60% when only using 20% data, and improves the performance by 3.30% when using all data.

Comparisons to contrastive learning methods. There are some similarities between our method and contrastive learning methods. For contrastive learning, it aims to narrow the distance between positive pairs (augment copies from the same image) by contrastive learning losses. For our NCL++, it employs the KL loss to minimize the distance between positive sample pairs. Specifically, the definition of the positive sample pairs are two folds: one is the different augment copies from the same image in our IntraCL, and the other is the different predictions of the same image from different networks in InterCL. However, there are also some differences among them. For example, contrastive learning is an unsupervised method while our method is a supervised method. It is precisely because of supervised learning where each sample has a certain label, we further propose the nested feature learning to increase the network's ability to distinguish the sample from the negative hard categories. We have compared the performance of our method and other contrastive learning based methods (including Hybrid-PSC [80] and PaCo [73]) as shown in Table 10. For the supervised contrastive learning methods, like [90], the previous work [73] has conducted lots of experiments to show that it is not directly applicable to long-tailed recognition. Therefore, here we do not take supervised contrastive learning methods for comparisons. As we can see, our method outperforms them by a considerable margin. Regarding the aspect of hard sample mining, Hybrid-PSC enhances training on the tail by reweighting, while PaCo enhances training on the tail by resampling. These approaches mine hard samples based on categories. While in our method, we proposed HCM to focus on each individual sample, extracting hard categories specific to that sample. Therefore, our method is still different from the hard example mining, and has not been proposed before. We further take a simple experiment to compare the performance of our IntraCL and a classic contrastive learning MoCo. Actually, when replacing the IntraCL with MoCo's contrastive learning module, our NCL++ degenerates to the previous version NCL [74]. As shown in Table 10, our IntraCL could achieve better performance, which shows the advantages of the proposed method.

5. Conclusions

In this work, we have proposed a Nested Collaborative Learning (NCL++) to enhance the discriminative ability of the model in long-tailed learning by collaborative learning. Five components including BIL, IntraCL, InterCL, HCM and NFL were proposed in our NCL++. The goal of BIL is to enhance the discriminative ability of a single network. IntraCL and InterCL conduct the collaborative learning to reduce the learning uncertainty. For HCM and NFL, they aim to select hard negative categories and then formulate the learning in a nested

way to improve the meticulous distinguishing ability of the model. Extensive experiments have verified the superiorities of our method over the state-of-the-art.

Limitations and broader impacts. One limitation is that more GPU memory and computing power are needed when training our NCL++ with multiple experts and augmented copies, which will bring higher training costs. But fortunately, one expert is also enough to achieve promising performance in inference, which still shows some advantages compared to the ensemble-based multi-expert approaches. Moreover, the proposed method improves the accuracy and fairness of the classifier, which promotes the visual model to be further put into practical use. To some extent, it helps to collect large datasets without forcing class balancing preprocessing, which improves efficiency and effectiveness of work. The negative impacts can yet occur in some misuse scenarios, e.g., identifying minorities for malicious purposes. Therefore, the appropriateness of the purpose of using long-tailed classification technology is supposed to be ensured with attention.

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.

Data availability

Data will be made available on request.

Acknowledgments

References

[1] Kaiming He, Xiangyu Zhang, Shaoqing Ren, Jian Sun, Deep residual learning for image recognition, in: CVPR, 2016.

[2] Jie Hu, Li Shen, Gang Sun, Squeeze-and-excitation networks, in: CVPR, 2018.

[3] Shaoqing Ren, Kaiming He, Ross Girshick, Jian Sun, Faster r-cnn: Towards real-time object detection with region proposal networks, in: NeurIPS, 2015.

[4] Shifeng Zhang, Longyin Wen, Xiao Bian, Zhen Lei, Stan Z Li, Single-shot refinement neural network for object detection, in: CVPR, 2018.

[5] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, Jiaya Jia, Pyramid scene parsing network, in: CVPR, 2017.

[6] Jun Fu, Jing Liu, Haijie Tian, Yong Li, Yongjun Bao, Zhiwei Fang, Hanqing Lu, Dual attention network for scene segmentation, in: CVPR, 2019.

[7] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, Li Fei-Fei, Imagenet: A large-scale hierarchical image database, in: CVPR, 2009.

[8] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, C Lawrence Zitnick, Microsoft coco: Common objects in context, in: ECCV, 2014.

[9] Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, Antonio Torralba, Places: A 10 million image database for scene recognition, IEEE TPAMI (2017).

[10] Yu-Xiong Wang, Deva Ramanan, Martial Hebert, Learning to model the tail, in: NeurIPS, 2017.

[11] Ziwei Liu, Zhongqi Miao, Xiaohang Zhan, Jiayun Wang, Boqing Gong, Stella X Yu, Large-scale long-tailed recognition in an open world, in: CVPR, 2019.

[12] Saptarshi Sinha, Hiroki Ohashi, Katsuyuki Nakamura, Class-difficulty based methods for long-tailed visual recognition, Int. J. Comput. Vis. 130 (10) (2022) 2517–2531.

[13] Lu Yang, He Jiang, Qing Song, Jun Guo, A survey on long-tailed visual recognition, Int. J. Comput. Vis. 130 (7) (2022) 1837–1872.

[14] Weiqiu Wang, Zhicheng Zhao, Pingyu Wang, Fei Su, Hongying Meng, Attentive feature augmentation for long-tailed visual recognition, IEEE Trans. Circuits Syst. Video Technol. 32 (9) (2022) 5803–5816.

[15] Ming-Liang Zhang, Xu-Yao Zhang, Chuang Wang, Cheng-Lin Liu, Towards prior gap and representation gap for long-tailed recognition, Pattern Recognit. 133 (2023) 109012.

[16] Xinqiao Zhao, Jimin Xiao, Siyue Yu, Hui Li, Bingfeng Zhang, Weight-guided class complementing for long-tailed image recognition, Pattern Recognit. 138 (2023) 109374.

[17] Jingru Tan, Bo Li, Xin Lu, Yongqiang Yao, Fengwei Yu, Tong He, Wanli Ouyang, The equalization losses: Gradient-driven training for long-tailed object recognition, IEEE Trans. Pattern Anal. Mach. Intell. (2023).

[18] Yucan Zhou, Qinghua Hu, Yu Wang, Deep super-class learning for long-tail distributed image classification, Pattern Recognit. 80 (2018) 118–128.

[19] Haibo He, Edwardo A. Garcia, Learning from imbalanced data, IEEE TKDE (2009).

[20] Mateusz Buda, Atsuto Maki, Maciej A. Mazurowski, A systematic study of the class imbalance problem in convolutional neural networks, Neural Netw. (2018).

[21] Yin Cui, Menglin Jia, Tsung-Yi Lin, Yang Song, Serge Belongie, Class-balanced loss based on effective number of samples, in: CVPR, 2019.

[22] Chen Huang, Yining Li, Chen Change Loy, Xiaoou Tang, Learning deep representation for imbalanced classification, in: CVPR, 2016.

[23] Mengye Ren, Wenyuan Zeng, Bin Yang, Raquel Urtasun, Learning to reweight examples for robust deep learning, in: ICML, 2018.

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

[25] Kaidi Cao, Colin Wei, Adrien Gaidon, Nikos Arechiga, Tengyu Ma, Learning imbalanced datasets with label-distribution-aware margin loss, in: NeurIPS, 2019.

[26] Liuyu Xiang, Guiguang Ding, Jungong Han, Learning from multiple experts: Self-paced knowledge distillation for long-tailed classification, in: ECCV, 2020.

[27] Xudong Wang, Long Lian, Zhongqi Miao, Ziwei Liu, Stella X Yu, Long-tailed recognition by routing diverse distribution-aware experts, in: ICLR, 2021.

[28] Yu Li, Tao Wang, Bingyi Kang, Sheng Tang, Chunfeng Wang, Jintao Li, Jiashi Feng, Overcoming classifier imbalance for long-tail object detection with balanced group softmax, in: CVPR, 2020.

[29] Jiarui Cai, Yizhou Wang, Jenq-Neng Hwang, ACE: Ally complementary experts for solving long-tailed recognition in one-shot, in: ICCV, 2021.

[30] Yifan Zhang, Bryan Hooi, Lanqing Hong, Jiashi Feng, Test-agnostic long-tailed recognition by test-time aggregating diverse experts with self-supervision, 2021, arXiv preprint arXiv:2107.09249.

[31] Qiushan Guo, Xinjiang Wang, Yichao Wu, Zhipeng Yu, Ding Liang, Xiaolin Hu, Ping Luo, Online knowledge distillation via collaborative learning, in: CVPR, 2020.

[32] Xu Lan, Xiatian Zhu, Shaogang Gong, Knowledge distillation by on-the-fly native ensemble, 2018, arXiv preprint arXiv:1806.04606.

[33] Ying Zhang, Tao Xiang, Timothy M. Hospedales, Huchuan Lu, Deep mutual learning, in: CVPR, 2018.

[34] Boyan Zhou, Quan Cui, Xiu-Shen Wei, Zhao-Min Chen, Bbn: Bilateral-branch network with cumulative learning for long-tailed visual recognition, in: CVPR, 2020.

[35] Dong Cao, Xiangyu Zhu, Xingyu Huang, Jianzhu Guo, Zhen Lei, Domain balancing: Face recognition on long-tailed domains, in: CVPR, 2020.

[36] Jiawei Ren, Cunjun Yu, Shunan Sheng, Xiao Ma, Haiyu Zhao, Shuai Yi, Hongsheng Li, Balanced meta-softmax for long-tailed visual recognition, 2020, arXiv preprint arXiv:2007.10740.

[37] Yongshun Zhang, Xiu-Shen Wei, Boyan Zhou, Jianxin Wu, Bag of tricks for long-tailed visual recognition with deep convolutional neural networks, in: AAAI, 2021.

[38] Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, W Philip Kegelmeyer, SMOTE: Synthetic minority over-sampling technique, J. Artif. Intell. Res. 16 (2002) 321–357.

[39] Tao Wang, Yu Li, Bingyi Kang, Junnan Li, Junhao Liew, Sheng Tang, Steven Hoi, Jiashi Feng, The devil is in classification: A simple framework for long-tail instance segmentation, in: ECCV, Springer, 2020, pp. 728–744.

[40] Zizhao Zhang, Tomas Pfister, Learning fast sample re-weighting without reward data, in: ICCV, 2021, pp. 725–734.

[41] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, Piotr Dollár, Focal loss for dense object detection, in: ICCV, 2017.

[42] Jiaqi Wang, Wenwei Zhang, Yuhang Zang, Yuhang Cao, Jiangmiao Pang, Tao Gong, Kai Chen, Ziwei Liu, Chen Change Loy, Dahua Lin, Seesaw loss for long-tailed instance segmentation, in: CVPR, 2021.

[43] Jingru Tan, Changbao Wang, Buyu Li, Quanquan Li, Wanli Ouyang, Changqing Yin, Junjie Yan, Equalization loss for long-tailed object recognition, in: CVPR, 2020.

[44] Peilin Zhao, Yifan Zhang, Min Wu, Steven CH Hoi, Mingkui Tan, Junzhou Huang, Adaptive cost-sensitive online classification, IEEE Trans. Knowl. Data Eng. 31 (2) (2018) 214–228.

[45] Han-Jia Ye, Hong-You Chen, De-Chuan Zhan, Wei-Lun Chao, Identifying and compensating for feature deviation in imbalanced deep learning, 2020, arXiv preprint arXiv:2001.01385.

[46] Youngkyu Hong, Seungju Han, Kwanghee Choi, Seokjun Seo, Beomsu Kim, Buru Chang, Disentangling label distribution for long-tailed visual recognition, in: CVPR, 2021.

[47] Aditya Krishna Menon, Sadeep Jayasumana, Ankit Singh Rawat, Himanshu Jain, Andreas Veit, Sanjiv Kumar, Long-tail learning via logit adjustment, 2020, arXiv preprint arXiv:2007.07314.

[48] Songyang Zhang, Zeming Li, Shipeng Yan, Xuming He, Jian Sun, Distribution alignment: A unified framework for long-tail visual recognition, in: CVPR, 2021.

[49] Tianhao Li, Limin Wang, Gangshan Wu, Self supervision to distillation for long-tailed visual recognition, in: ICCV, 2021.

[50] Yan Zhao, Weicong Chen, Xu Tan, Kai Huang, Jihong Zhu, Adaptive logit adjustment loss for long-tailed visual recognition, in: AAAI, 2022.

[51] Bolian Li, Zongbo Han, Haining Li, Huazhu Fu, Changqing Zhang, Trustworthy long-tailed classification, in: CVPR, 2021.

[52] Geoffrey Hinton, Oriol Vinyals, Jeff Dean, Distilling the knowledge in a neural network, 2015, arXiv preprint arXiv:1503.02531.

[53] Tommaso Furlanello, Zachary Lipton, Michael Tschannen, Laurent Itti, Anima Anandkumar, Born again neural networks, in: ICML, 2018.

[54] Adriana Romero, Nicolas Ballas, Samira Ebrahimi Kahou, Antoine Chassang, Carlo Gatta, Yoshua Bengio, Fitnets: Hints for thin deep nets, 2014, arXiv preprint arXiv:1412.6550.

[55] Nikolaos Passalis, Anastasios Tefas, Learning deep representations with probabilistic knowledge transfer, in: ECCV, 2018.

[56] Tianhong Li, Jianguo Li, Zhuang Liu, Changshui Zhang, Few sample knowledge distillation for efficient network compression, in: CVPR, 2020, pp. 14639–14647.

[57] Sergey Zagoruyko, Nikos Komodakis, Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer, 2016, arXiv preprint arXiv:1612.03928.

[58] Defang Chen, Jian-Ping Mei, Can Wang, Yan Feng, Chun Chen, Online knowledge distillation with diverse peers, in: AAAI, 2020.

[59] Nikita Dvornik, Cordelia Schmid, Julien Mairal, Diversity with cooperation: Ensemble methods for few-shot classification, in: ICCV, 2019.

[60] Prashant Bhat, Elahe Arani, Bahram Zonooz, Distill on the go: Online knowledge distillation in self-supervised learning, in: CVPRW, 2021, pp. 2678–2687.

[61] Zhilu Zhang, Mert Sabuncu, Self-distillation as instance-specific label smoothing, in: NeurIPS, Vol. 33, 2020, pp. 2184–2195.

[62] Linfeng Zhang, Jiebo Song, Anni Gao, Jingwei Chen, Chenglong Bao, Kaisheng Ma, Be your own teacher: Improve the performance of convolutional neural networks via self distillation, in: ICCV, 2019, pp. 3713–3722.

[63] Yuenan Hou, Zheng Ma, Chunxiao Liu, Chen Change Loy, Learning lightweight lane detection cnns by self attention distillation, in: ICCV, 2019, pp. 1013–1021.

[64] Mary Phuong, Christoph H. Lampert, Distillation-based training for multi-exit architectures, in: ICCV, 2019, pp. 1355–1364.

[65] Li Yuan, Francis EH Tay, Guilin Li, Tao Wang, Jiashi Feng, Revisit knowledge distillation: A teacher-free framework, 2019.

[66] Seyed Iman Mirzadeh, Mehrdad Farajtabar, Ang Li, Nir Levine, Akihiro Matsukawa, Hassan Ghasemzadeh, Improved knowledge distillation via teacher assistant, in: AAAI. Vol. 34, no. 04, 2020, pp. 5191–5198.

[67] Devesh Walawalkar, Zhiqiang Shen, Marios Savvides, Online ensemble model compression using knowledge distillation, in: ECCV, Springer, 2020, pp. 18–35.

[68] Ekin D. Cubuk, Barret Zoph, Jonathon Shlens, Quoc V. Le, Randaugment: Practical automated data augmentation with a reduced search space, in: CVPRW, 2020.

[69] Alexander Hermans, Lucas Beyer, Bastian Leibe, In defense of the triplet loss for person re-identification, 2017, arXiv preprint arXiv:1703.07737.

[70] Grant Van Horn, Oisin Mac Aodha, Yang Song, Yin Cui, Chen Sun, Alex Shepard, Hartwig Adam, Pietro Perona, Serge Belongie, The inaturalist species classification and detection dataset, in: CVPR, 2018.

[71] Alex Krizhevsky, Geoffrey Hinton, et al., Learning Multiple Layers of Features from Tiny Images, Citeseer, 2009.

[72] Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, Kaiming He, Aggregated residual transformations for deep neural networks, in: CVPR, 2017.

[73] Jiequan Cui, Zhisheng Zhong, Shu Liu, Bei Yu, Jiaya Jia, Parametric contrastive learning, in: ICCV, 2021.

[74] Jun Li, Zichang Tan, Jun Wan, Zhen Lei, Guodong Guo, Nested collaborative learning for long-tailed visual recognition, in: CVPR, 2022.

[75] Zichang Tan, Ajian Liu, Jun Wan, Hao Liu, Zhen Lei, Guodong Guo, Stan Z Li, Cross-batch hard example mining with pseudo large batch for ID vs. Spot face recognition, IEEE Trans. Image Process. 31 (2022) 3224–3235.

[76] Muhammad Abdullah Jamal, Matthew Brown, Ming-Hsuan Yang, Liqiang Wang, Boqing Gong, Rethinking class-balanced methods for long-tailed visual recognition from a domain adaptation perspective, in: CVPR, 2020.

[77] Zhengzhuo Xu, Zenghao Chai, Chun Yuan, Towards calibrated model for long-tailed visual recognition from prior perspective, in: NeurIPS. Vol. 34, 2021.

[78] Jaehyung Kim, Jongheon Jeong, Jinwoo Shin, M2m: Imbalanced classification via major-to-minor translation, in: CVPR, 2020.

[79] Zhisheng Zhong, Jiequan Cui, Shu Liu, Jiaya Jia, Improving calibration for long-tailed recognition, in: CVPR, 2021.

[80] Peng Wang, Kai Han, Xiu-Shen Wei, Lei Zhang, Lei Wang, Contrastive learning based hybrid networks for long-tailed image classification, in: CVPR, 2021.

[81] Yin-Yin He, Jianxin Wu, Xiu-Shen Wei, Distilling virtual examples for long-tailed recognition, in: ICCV, 2021.

[82] Tianhao Li, Limin Wang, Gangshan Wu, Self supervision to distillation for long-tailed visual recognition, in: ICCV, 2021.

[83] Beier Zhu, Yulei Niu, Xian-Sheng Hua, Hanwang Zhang, Cross-domain empirical risk minimization for unbiased long-tailed classification, in: AAAI, 2022.

[84] Xiaohua Chen, Yucan Zhou, Dayan Wu, Wanqian Zhang, Yu Zhou, Bo Li, Weiping Wang, Imagine by reasoning: A reasoning-based implicit semantic data augmentation for long-tailed classification, in: AAAI, 2022.

[85] Zhi Hou, Baosheng Yu, Dacheng Tao, BatchFormer: Learning to explore sample relationships for robust representation learning, in: CVPR, 2022.

[86] Jialun Liu, Jingwei Zhang, Wenhui Li, Chi Zhang, Yifan Sun, Memory-based jitter: Improving visual recognition on long-tailed data with diversity in memory, in: AAAI, 2022.

[87] Shaden Alshammari, Yu-Xiong Wang, Deva Ramanan, Shu Kong, Long-tailed recognition via weight balancing, in: CVPR, 2022.

[88] Hongyi Zhang, Moustapha Cisse, Yann N. Dauphin, David Lopez-Paz, Mixup: Beyond empirical risk minimization, in: ICLR, 2018.

[89] Laurens Van der Maaten, Geoffrey Hinton, Visualizing data using t-SNE, J. Mach. Learn. Res. 9 (11) (2008).

[90] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, Dilip Krishnan, Supervised contrastive learning, in: Advances in Neural Information Processing Systems. Vol. 33, 2020, pp. 18661–18673.

Zichang Tan received the B.E. degree from the department of Automation, Huazhong University of Science and Technology (HUST) in 2016, and the Ph.D. degree from the Institute of Automation, Chinese Academy of Sciences in 2021. He was named as an outstanding graduate of the college when he graduated, and was a winner of the 2020 CAS President Award. Since July 2021, he has joined Baidu Inc. as a researcher. His main research interests include deep learning, computer vision and biometrics in particular.

Jun Li received the B.E. degree from the Harbin Institute of Technology (HIT), Harbin, China in 2021. He was named as an outstanding graduate of the college when he graduated. He is currently studying for a Master's degree at the Institute of Automation, Chinese Academy of Science (CASIA). His main research interests include deep learning, computer vision and long-tailed learning.

Jinhao Du received the B.E. degree from Central South University (CSU), Changsha, China in 2021. She was named as an outstanding graduate of Hunan Province when she graduated. She is currently studying for a Master's degree at Peking University. Her main research interests include deep learning, computer vision and long-tailed learning.

Jun Wan received his B.S. degree from the China University of Geosciences, Beijing, China, in 2008, and the Ph.D. degree from the Institute of Information Science, Beijing Jiaotong University, Beijing, China, in 2015. Since January 2015, he has been a Faculty Member with the National Laboratory of Pattern Recognition (NLPR), Institute of Automation, Chinese Academy of Science (CASIA), China, where he currently serves as an Associate Professor. He is an associate editor of the IET Biometrics (2020–2022), an Area Chair for ICME 2021 and 2020, the Senior Program Committee for AAAI 2021, and has served as co-editor of special issues in IEEE TPAMI, IEEE TBIOM. His main research interests include computer vision and machine learning.

Zhen Lei received the BS degree in automation from the University of Science and Technology of China (USTC), in 2005, and the Ph.D. degree from the Institute of Automation, Chinese Academy of Sciences, in 2010, where he is currently a professor. He has published more than 200 papers in international journals and conferences. His research interests are in computer vision, pattern recognition, image processing, and face recognition in particular. He served as an area chair of the International Joint Conference on Biometrics in 2014, the IAPR/IEEE International Conference on Biometric in 2015, 2016, 2018, and the IEEE International Conference on Automatic Face and Gesture Recognition in 2015. He is the winner of 2019 IAPR YOUNG BIOMETRICS INVESTIGATOR AWARD. He is a IAPR Fellow.

Guodong Guo received the B.E. degree in Automation from Tsinghua University, Beijing, China, the Ph.D. degree in Computer Science from University of Wisconsin, Madison, WI, USA. He is the Head of the Institute of Deep Learning (IDL), Baidu Research, and also affiliated with the Dept. of Computer Science and Electrical Engineering, West Virginia University (WVU), USA. In the past, he studied, visited or worked in several places, including Institute of Automation, Chinese Academy of Sciences; INRIA, Sophia Antipolis, France; Ritsumeikan University, Kyoto, Japan; and Microsoft Research, Beijing, China; He authored a book, "Face, Expression, and Iris Recognition Using Learning-based Approaches" (2008), co-edited two books, "Support Vector Machines Applications" (2014) and "Mobile Biometrics" (2017), and co-authored a book, "Multi-Modal Face Presentation Attack Detection" (2020). He published over 180 technical papers, and he is the creator of the visual BMI (body mass index)

estimator. His research interests include computer vision, biometrics, machine learning, and multimedia. He is an AE of several journals, including IEEE Trans. on Affective Computing. He received the North Carolina State Award for Excellence in Innovation in 2008, New Researcher of the Year (2010–2011), and Outstanding Researcher (2017–2018, 2013–2014) at CEMR, WVU. He was selected the "People's Hero of the Week" by BSJB under Minority Media and Telecommunications Council (MMTC) in 2013. Two of his papers were selected as "The Best of FG'13" and "The Best of FG'15", respectively.