

# Towards Effective Collaborative Learning in Long-Tailed Recognition

Zhengzhuo Xu\*, Zenghao Chai\*, Chengyin Xu,  
Chun Yuan<sup>†</sup>, *Senior Member, IEEE*, Haiqin Yang<sup>†</sup>, *Senior Member, IEEE*

**Abstract**—Real-world data usually suffers from severe class imbalance and long-tailed distributions, where minority classes are significantly underrepresented compared to the majority ones. Recent research prefers to utilize multi-expert architectures to mitigate the model uncertainty on the minority, where collaborative learning is employed to aggregate the knowledge of experts, i.e., online distillation. In this paper, we observe that the knowledge transfer between experts is imbalanced in terms of class distribution, which results in limited performance improvement of the minority classes. To address it, we propose a re-weighted distillation loss by comparing two classifiers’ predictions, which are supervised by online distillation and label annotations, respectively. We also emphasize that feature-level distillation will significantly improve model performance and increase feature robustness. Finally, we propose an Effective Collaborative Learning (ECL) framework that integrates a contrastive proxy task branch to further improve feature quality. Quantitative and qualitative experiments on four standard datasets demonstrate that ECL achieves state-of-the-art performance and the detailed ablation studies manifest the effectiveness of each component in ECL.

**Index Terms**—Image Classification, Long Tail Recognition, Collaborative Learning, Knowledge Distillation.

## I. INTRODUCTION

RECENT advancements in computer vision, e.g., visual recognition [1], video analysis [2] and person re-ID [3], [4], heavily rely on the large-scale, high-quality, and balanced datasets, such as ImageNet [5], COCO [6] and Place [7], which require laborious collections and careful annotations. Regrettably, collecting rare instances entails gathering more dominant samples because real-world data naturally exhibits imbalanced distributions w.r.t. its categories. Hence, datasets typically follow a long-tailed distribution, with only a few labels having a majority of the samples, while most labels are associated with limited instances. In Long Tail Recognition (LTR), the minority classes (**tail**) are always overwhelmed by the majority classes (**head**), resulting in low performance for the tail. As a result, the models trained on the long-tailed dataset show great uncertainty, where the outputs for few-shot classes vary remarkably, despite the same training settings.

Most existing work addresses the LTR issue by improving the feature representations of tail classes or re-balancing the

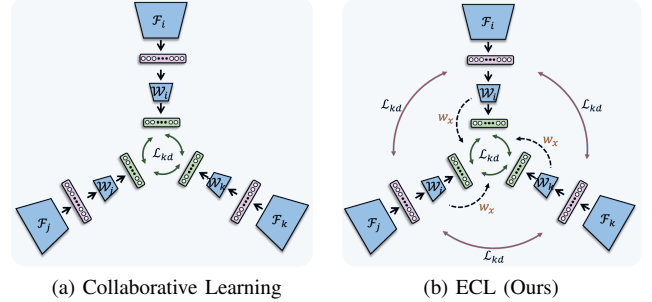


Fig. 1: The illustration of collaborative learning in the multi-expert framework.  $\mathcal{F}$ : feature encoder.  $\mathcal{W}$ : classification head. Different from previous work, we re-balance the distillation and conduct online distillation on both feature and logit levels.

contribution of different classes. However, some intuitive approaches like over-sampling the tail [8] or under-sampling the head [9] result in severe robustness problems especially in tail classes. Although some well-designed approaches enrich tail samples in more elegant ways, such as through feature combinations [10], [11], [12] or pseudo sample generation [13], [9], [14], the problem of model preference towards head classes remains unresolved. To calibrate the label distribution gap between the train and test dataset, the Balanced Cross-entropy (BC) loss is proposed based on *Bayesian Theory*, which compensates the model bias by label frequency on standard *softmax* Cross-Entropy (CE) loss [15], [16], [17], [11]. Based on the effective BC loss, Multi-Expert (ME) [18], [19], [20] framework is proposed to further address model uncertainty on the tail classes. For example, NCL [20] trains several expert networks in parallel and aggregates each expert’s knowledge in a nested collaborative manner, i.e., online Knowledge Distillation (KD) on the logit level (see Fig. 1a), where we refer to each network as an expert.

However, our experimental observations indicate that the transfer knowledge (distillation logit value) is not balanced w.r.t. class in vanilla collaborative learning (see Sec. IV). The tail samples are always under-represented during the distillation process, which damages the balanced knowledge transfer. Such imbalance leads the online distillation to boost the head performance while suppressing the transfer of tail knowledge. Consequently, the tail remains unimproved compared to the single expert baseline. Recent research [21] suggests that the KD-trained classifier is more confident for the over-represented samples than the label-trained one because the

<sup>†</sup>Corresponding authors: C. Yuan is with the Tsinghua Shenzhen International Graduate School, Tsinghua University, Shenzhen 518055, China. (email: yuanc@sz.tsinghua.edu.cn). H. Yang is with International Digital Economy Academy, Shenzhen 518045, China. (email: hqyang@iee.org).

\*Equal contribution authors, listing order is random. Z. Xu, Z. Chai and C. Xu are with the Tsinghua Shenzhen International Graduate School, Tsinghua University, Shenzhen 518055, China. (e-mail: {xzz20, xucy20}@mails.tsinghua.edu.cn, zenghaochai@gmail.com).

distillation tends to learn more generalized *context* knowledge compared to label supervision, which mainly provides content-invariant knowledge. Inspired by it, we propose a novel re-weighted distillation loss by comparing the predictions of two different classifiers. Moreover, we propose to perform additional collaborative distillation at the feature level, which significantly boosts model performance and feature robustness. We further incorporate a contrastive proxy task with a parallel branch to improve feature representations. As a result, we propose a novel Effective Collaborative Learning (ECL) framework to improve vanilla NCL, which distinguishes from previous ME frameworks in two aspects:

**Single expert training.** We propose the Balanced Knowledge Transfer (BKT) module to conduct balanced knowledge distillation. Following the feature encoder, we add an extra reference classifier parallel to the original classifier. The reference classifier is only supervised by the BC loss and is not involved in the expert collaboration, allowing it to only focus on the content-invariant knowledge. We compare the predictions of two classifiers to estimate whether the input samples are over-confident or not and re-weight the KD loss to assign the under-represented samples with larger weights (Fig. 1b). For each expert, we introduce a siamese branch to conduct Contrastive Proxy Task (CPT) and update parameters in a momentum-based moving average scheme [22]. The CPT is designed to increase the feature similarity of an image’s two views to facilitate model discriminative ability. Note that we will discard the additional reference classifier and siamese branch during the inference phase to keep the consistent architecture with previous ME approaches.

**Expert knowledge aggregation.** In the proposed ECL, each expert is collaboratively learned with others. Note that the knowledge is transferred not only on the logit level but also on the feature level (see Fig. 1b), which facilitates stable representation learning. Our Feature Level Distillation (FLD) is a simple yet effective improvement that encourages all experts to extract well-represented features. We also present in-depth analysis to investigate how FLD influences the model performance qualitatively and quantitatively (see Sec. VI).

With the above observations, insights, and techniques, we build our final ECL (Fig. 1b&3), which contains three key components, namely the balanced knowledge transfer module, feature level distillation, and contrastive proxy task. Extensive experiments in four benchmarks justify the superiority of ECL. In summary, our contributions are as follows:

- 1) We pinpoint the imbalance of transfer knowledge in previous collaborative learning methods and propose a balanced knowledge distillation loss to tackle it.
- 2) We propose to conduct knowledge distillation on both feature and logit levels, which significantly enhances model performance and robustness.
- 3) We propose the ECL framework to collaboratively train multiple experts to overcome the head preference and tail uncertainty in long-tailed recognition.
- 4) We present extensive experiments and demonstrate ECL achieves state-of-the-art performance on CIFAR10/100-LT, ImageNet-LT, and iNaturalist 2018 datasets.

This paper is organized as follows: Sec. II provides a brief overview of related work. In Sec. III, we introduce the relevant concepts and baselines. We discuss our motivation based on experimental observations in Sec. IV and provide a detailed design in Sec. V. Sec. VI demonstrates the effectiveness of ECL through extensive experiments and ablation studies. Finally, Sec. VII concludes our work.

1. We pinpoint the imbalance of transfer knowledge in previous collaborative learning methods and propose a balanced knowledge distillation loss to tackle it.
2. We propose to conduct knowledge distillation on both feature and logit levels, which significantly enhances model performance and robustness.
3. We propose the ECL framework to collaboratively train multiple experts to overcome the head preference and tail uncertainty in long-tailed recognition.
4. We present extensive experiments and demonstrate ECL achieves state-of-the-art performance on CIFAR10/100-LT, ImageNet-LT, and iNaturalist 2018 datasets.

## II. RELATED WORK

**Feature-wise Rebalance Learning.** To avoid damaging model generalization severely from simply over/under-sampling the tail/head classes [23], [24], [9], recent advances resort combination of the head to enrich the feature of tail samples [13], [9], [25] or increase the tail frequency implicitly [10], [26], [11], [14]. The two-stage methods [27], [24], [28] decouple feature learning from downstream tasks (e.g., classification) to reduce the bias on the classifier. Several methods [29], [22], [30] also leverage self-supervised learning to eliminate the influence of imbalanced distribution. SSP [31] and HybridSC [32] demonstrated that self-supervised or semi-supervised training can boost performance through larger train epochs and GPU memory. Recent state-of-the-art [33], [32], [34], [35] introduces fixed or learnable proxy to overcome performance degradation due to the absence of label supervision.

**Reweight-wise Learning.** To mitigate the inherent statistical bias in LTR, researchers have designed meticulous loss to learn larger *margins* among different classes [27], [16], [17], [15], [11], [36], [37], [38] or assign various *weights* for different classes based on the label frequencies [39], [40], [41], [28], [42]. In particular, the simple yet effective BC loss [16], [15], [17], [11] has been widely adopted by state-of-the-art [20], [33], [35], [43]. Unfortunately, BC loss is not always compatible with the above feature-wise methods for the inconsistency of the statistical label frequency.

**Multi-expert Learning.** To tackle the tail uncertainty [44], [19], the multi-expert framework is increasingly valued, which typically contains two components, i.e., *single expert training* and *experts knowledge aggregation* [8], [45], [18], [19], [43]. BBN [8] trains two experts with instance sampling and inverse sampling, respectively, and aggregates their knowledge in a cumulative weighting manner. LFME [18] trains multiple experts with different instance groups and weights the logits from different experts as the final output. RIDE [19] enlarges the KL divergence to train experts and cascades all the experts via decision gates for inference. TADE [43] trains experts by BC loss with different assumed statistical prior and weights

each expert’s output, which is obtained via post-hoc contrastive training. Another feasible expert aggregation manner is knowledge distillation [46]. DiVe [47] shows the effectiveness of distillation in the LTR. SSD [48] trains expert backbone by self-distillation learning and classifier through balanced sampling. CBD [44] trains different teachers by various data augmentations and random seeds. Then, it trains students with balance sampling and knowledge from the above teachers. NCL [20] trains experts in a nested manner and adopts online inter-distillation with each other to reduce the tail uncertainty. However, these methods mainly conduct logit-level distillation while ignoring the imbalance of transfer knowledge.

### III. PRELIMINARIES

#### A. Task Definition.

Given an  $N$ -sample dataset  $\mathcal{D} = \{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^N$  from  $C$  classes, where  $\mathbf{x}_i \in \mathcal{X}$  denotes the  $i$ -th instance with its label,  $\mathbf{y}_i \in \mathcal{Y} := \{\mathbf{y}_1, \dots, \mathbf{y}_C\}$ . We assume the dataset  $\mathcal{D}$  is long-tailed distributed and denote each category as  $\mathcal{C}_i$  and its instance number as  $n_i = |\mathcal{C}_i|$ . Furthermore, we consider a base classification model  $\mathcal{M} := \{\mathcal{F}_\theta, \mathcal{W}_\phi\}$ . It contains a learnable *feature encoder*  $\mathcal{F}_\theta$  and a *classifier*  $\mathcal{W}_\phi$ , parameterized by  $\theta$ ,  $\phi$ , respectively. Given an input image  $\mathbf{x}$ , the encoder extracts the feature representation  $\mathbf{v} := \mathcal{F}_\theta(\mathbf{x}) \in \mathbb{R}^d$ . Then, the classifier (typically fully connected layers) outputs the logits  $\mathbf{z} := \mathcal{W}_\phi(\mathbf{v}) \in \mathbb{R}^C$ . We assume  $K$  experts in the collaborative learning framework with the same architecture  $\mathcal{M}$  and denote the  $k$ -th expert as  $E_k := \{\mathcal{F}_{\theta_k}, \mathcal{W}_{\phi_k}\}$ .

#### B. Balanced Cross-entropy Loss.

**Balanced Cross-entropy (BC)** loss is effective and widely adopted in LTR tasks [16], [15], [17], [11], [20], [35]. It compensates the statistical bias via logits adjustment on standard **Cross-Entropy (CE)** loss. Consider the expert  $E_k$  is supervised by CE loss with standard *softmax*:

$$\mathcal{L}_{\text{CE}} = -\log(p(\mathbf{y}_i|\mathbf{x}; \theta_k, \phi_k)) = \log \left[ 1 + \sum_{\mathbf{y}_j \neq \mathbf{y}_i} e^{\mathbf{z}_{\mathbf{y}_j} - \mathbf{z}_{\mathbf{y}_i}} \right]. \quad (1)$$

Here, we denote the label distribution prior of train/test data as  $p_s(\mathbf{y})/p_t(\mathbf{y})$  respectively. Based on the Bayesian theory, *the posterior is proportional to prior times likelihood*, where the likelihood  $p_s(\mathbf{x}|\mathbf{y})$  maximization is equal to the model parameters (i.e.,  $\theta, \phi$ ) learning. Typically, the posterior  $p_t(\mathbf{y}|\mathbf{x})$  is equivalent to likelihood  $p_s(\mathbf{x}|\mathbf{y})$  between train and test set when  $p_s(\mathbf{y}) \equiv p_t(\mathbf{y})$ . However, if we take the statistical distribution of label  $\mathbf{y}$  as its prior  $p(\mathbf{y})$ , we can derive the following bias from the mismatch of  $p_s(\mathbf{y})$  and  $p_t(\mathbf{y})$ :

$$\begin{aligned} p_t(\mathbf{y}|\mathbf{x}) &= \frac{p_s(\mathbf{x}|\mathbf{y}) \cdot p_s(\mathbf{x})}{p_s(\mathbf{y})} \cdot \frac{p_t(\mathbf{y})}{p_t(\mathbf{x})} \propto \frac{p_s(\mathbf{x}|\mathbf{y}) \cdot p_t(\mathbf{y})}{p_s(\mathbf{y})} \\ &= \frac{\frac{p_t(\mathbf{y})}{p_s(\mathbf{y})} \cdot e^{\mathbf{z}_{\mathbf{y}_i}}}{\sum_j \frac{p_t(\mathbf{y}_j)}{p_s(\mathbf{y}_j)} \cdot e^{\mathbf{z}_{\mathbf{y}_j}}} = \frac{e^{\mathbf{z}_{\mathbf{y}_i} - (\log(p_s(\mathbf{y}_i)) - \log(p_t(\mathbf{y}_i)))}}{\sum_j e^{\mathbf{z}_{\mathbf{y}_j} - (\log(p_s(\mathbf{y}_j)) - \log(p_t(\mathbf{y}_j)))}}, \end{aligned} \quad (2)$$

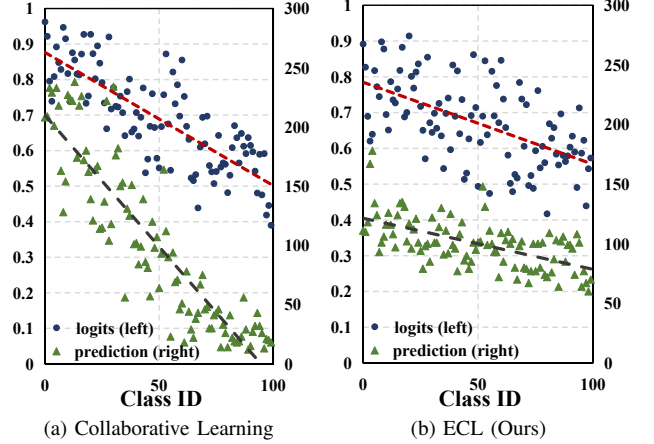


Fig. 2: Average distillation *logits value* and *prediction distribution* w.r.t. classes. We conduct an evaluation of vanilla NCL and proposed ECL on CIFAR100-LT ( $\gamma = 100$ ). The class index is ranked according to the training instance number.

where  $p_s(\mathbf{x})$  and  $p_t(\mathbf{x})$  are regular terms (i.e., normal distribution). Here, we get statistical bias of class  $\mathbf{y}_i$  as  $\log p_s(\mathbf{y}_i) - \log p_t(\mathbf{y}_i)$ . Combining Eq. 1 and Eq. 2, we compensate for it in CE loss with a hyper-parameter  $\tau$  as follows:

$$\mathcal{L}_{\text{BC}} = -\log \left[ \frac{e^{\mathbf{z}_{\mathbf{y}_i} + \tau \cdot (\log(p_s(\mathbf{y}_i)) - \log(p_t(\mathbf{y}_i)))}}{\sum_j e^{\mathbf{z}_{\mathbf{y}_j} + \tau \cdot (\log(p_s(\mathbf{y}_j)) - \log(p_t(\mathbf{y}_j)))}} \right]. \quad (3)$$

#### C. Nested Collaborative Learning

To reduce the great uncertainty in long-tailed learning, Li *et al.* [20] propose Nested Collaborative Learning (NCL) to learn multiple experts parallelly and aggregate the expert knowledge via nested online distillation on the logit-level (see Fig. 1a). The NCL performs online inter-distillation on both partial and full views, while incorporating an instance discrimination task as well. All experts adopt the same BC loss and hyper-parameter settings. It contributes to complementary expert learning and achieves state-of-the-art performance whether by using a single expert or an ensemble. Our ECL is motivated by the experimental observations on it, which will be elaborated in the following section.

## IV. MOTIVATION

Our motivation stems from the following inspiring observations **1**: *The transfer knowledge during online distillation is imbalanced w.r.t. classes in vanilla nested collaborative learning.* **2**: *The optimal hyper-parameter of Eq. 3 for each expert is not consistent, which hinders performance improvement.*

For multi-expert collaborative learning approaches, the expert knowledge aggregation typically conduct at the logit level. For observation **1**, we demonstrate that previous distillation at logit level (Fig. 1) is ineffective. In Fig. 2a, we visualize the logit-level transfer knowledge (i.e., distilled logits value) and model prediction numbers w.r.t. class of vanilla collaborative learning and ours. The tail classes present lower knowledge



logit level, we implement the re-weighted KD loss between each expert pair and the total loss will be:

$$\mathcal{L}_{\text{kd}}^{\text{logit}} = \frac{\sum_k \sum_{q \neq k} \sum_{\mathbf{x}_i} \hat{w}_{\mathbf{x}_i} \cdot \tau^2 \cdot \text{KL}(\varsigma(\frac{\mathbf{z}_i^{k,c}}{\tau}) || \varsigma(\frac{\mathbf{z}_i^{q,c}}{\tau}))}{N \cdot K \cdot (K - 1)}, \quad (7)$$

where  $\varsigma$  indicates *softmax*,  $\tau$  is the temperature factor and  $\text{KL}(p||q) = \sum_i p_i \cdot \log(p_i/q_i)$ .  $\mathbf{z}_i^{k,c}$  is the logits given by the *cls* head  $\mathcal{W}_\phi^c$  of expert  $k$  and  $\hat{w}_{\mathbf{x}_i}$  is given by Eq. 6.

Different from previous methods, we pinpoint that the distillation on the feature level will capture more robust knowledge in the LTR tasks, which can be formulated as follows:

$$\mathcal{L}_{\text{kd}}^{\text{feature}} = \frac{\sum_k \sum_{q \neq k} \sum_{\mathbf{x}_i} \tau^2 \cdot \text{KL}(\varsigma(\frac{\mathbf{v}_i^k}{\tau}) || \varsigma(\frac{\mathbf{v}_i^q}{\tau}))}{K(K - 1)}. \quad (8)$$

Experimentally, the online distillation on feature level shows significant effectiveness compared to logit level. we will present in-depth investigations on the reason for its performance and generalization in Sec. VI.

### C. Contrastive Proxy Task

To learn more generalized features, we follow [22] to adopt a contrastive proxy task in MoCo v2 manner. As Fig. 3 shows, an extra MoCo encoder is employed to perform instance discrimination, in which parameters are updated in a momentum-based moving average scheme to provide negative samples. For the feature  $\mathbf{v}_i^k$  given by expert  $k$  MoCo head, we denote the normalized embedding of its copy image with different augmentations as  $\tilde{\mathbf{v}}_i^k$ . For more negative pairs, a dynamic queue  $\mathcal{Q}^k$  is employed to record historical feature representations to save GPU memory. The info-NCE loss is adopted to increase the feature similarity of the same image while reducing the feature similarity of different images pairs, which is computed as:

$$\mathcal{L}_{\text{con}} = - \sum_k \log \frac{\exp(\mathbf{v}_i^{k,T} \tilde{\mathbf{v}}_i^k / \tau)}{\sum_{\tilde{\mathbf{v}}_j^k \in \{\mathcal{Q}^k \cup \mathbf{v}_i^{k,T}\}} \exp(\mathbf{v}_i^{k,T} \tilde{\mathbf{v}}_j^k / \tau)}. \quad (9)$$

### D. Model Training

Based on the above designs, we propose our final ECL in the multi-expert architecture, as Fig. 1&3 shows. To train the whole model, we leverage the classification loss  $\mathcal{L}_{\text{sup}}$ , distillation loss  $\mathcal{L}_{\text{kd}}$ , and contrastive loss  $\mathcal{L}_{\text{con}}$  for supervision. Formally,  $\mathcal{L}_{\text{sup}}$  compute the all experts BC loss between the predicted logits and the ground-truth labels for *ref* & *cls* head:

$$\mathcal{L}_{\text{sup}} = \frac{1}{K} \sum_k (\mathcal{L}_{\text{BC}}^{\text{ref}} + \mathcal{L}_{\text{BC}}^{\text{cls}}) \quad (10)$$

$\mathcal{L}_{\text{kd}}$  estimate the KL divergence on logit & feature level, while  $\mathcal{L}_{\text{con}}$  is used for the contrastive proxy task. Finally, the overall loss is formulated as:

$$\mathcal{L}_{\text{all}} = \mathcal{L}_{\text{sup}} + \alpha(\mathcal{L}_{\text{kd}}^{\text{logit}} + \mathcal{L}_{\text{kd}}^{\text{feature}}) + \beta \mathcal{L}_{\text{con}}, \quad (11)$$

where  $\alpha$  and  $\beta$  are the hyperparameters to balance the contribution of collaborative and contrastive learning.

### E. Model Inference

Note that the MoCo branch and *ref* head  $\mathcal{W}_\phi^r$  are only designed for effective model training. Therefore, in the inference phase, we only preserve the feature encoder  $\mathcal{F}_\theta$  and *cls* head  $\mathcal{W}_\phi$  to keep consistent model size with previous work. In addition, we can achieve higher performance by averaging the output logits from all experts as an ensemble model. In this case, our model size will be the same as the previous NCL.

## VI. EXPERIMENT

### A. Datasets

**CIFAR-10/100-LT.** CIFAR-10/100 [54] have 10/100 classes with 60,000 images in  $32 \times 32$  resolution. We follow [39], [27] to sample the train set of each class with exponential functions to create the long-tailed versions while remaining the validation set uniformly distributed. The imbalance factor  $\gamma$  indicates the skewness of the dataset, which is the ratio between the most and the least frequent classes. We employ  $\gamma = [10, 50, 100, 200]$  for comprehensive comparisons.

**ImageNet-LT** is the subset of the large-scale balanced ImageNet-1k [5], widely used in classification and localization tasks. The train data in ImageNet-LT are sampled through Pareto distribution with power value  $\alpha = 6$ . It contains 115.8K images from 1,000 classes. The most/least class number is 1,280/5 respectively ( $\gamma = 256$ ). we utilize the balanced validation set constructed by [39] for fair comparisons.

**iNaturalist 2018** [55] is the large-scale real-world LTR dataset. With over 437.5K images and 8,142 classes ( $\gamma = 500$ ), it suffers from severe label long-tailed distribution and fine-grained challenges. We follow [27] to utilize the official splits of training and validation sets in our experiments.

### B. Evaluation Metrics

**Top-1 Acc.** In LTR, the model is trained in an imbalanced dataset while evaluated in a balanced test set. Therefore, we adopt the common evaluation protocol Top-1 Acc. to estimate the model performance of each category.

**Group Acc.** In LTR, we focus more on the tail performance. Therefore, we follow [43] to group the test data into Many-shot ( $> 100$ ), Medium-shot ( $20 \sim 100$ ), and Few-shot ( $< 20$ ) according to the corresponding sample number w.r.t. different classes in the train set, and evaluate the accuracy of these groups separately.

**Loss-Acc Landscapes.** To investigate the feature-level distillation, we visualize the loss/accuracy landscapes of different models [49]. More specifically, we perturb the model weights by varying degrees through a series of Gaussian noises. The noise level is normalized to the  $l_2$ -norm of each filter to represent the effects of different weight amplitudes.

**Class-wise Average Feature Distance.** On the balanced test dataset, a well-trained encoder should map the input images into a distinguishable feature space. To evaluate the feature similarity, For class  $y_i$ , we calculate the class-wise average  $l_2$

TABLE I: Top-1 accuracy (%) on CIFAR-10/100-LT with ResNet32 backbone.  $\gamma$ : imbalance factor. Results are sorted according to method category. RW: re-weight wise methods. FW: feature improvement wise methods. ME: multi-expert frameworks. Underline: the best performance in each group. **Bold**: the best performance overall. We report the performance from original papers and reproduce results for unavailable settings according to their official repos.

Dataset	Type			CIFAR100				CIFAR10			
	RW	FW	ME	10	50	100	200	10	50	100	200
$\gamma$											
CE [39]	-	-	-	55.7	44.0	38.3	34.6	86.4	75.0	70.4	66.2
Focal Loss [50]	✓			55.8	44.3	38.4	35.6	86.6	76.7	70.4	68.9
$\tau$ Norm [24]	✓			59.1	48.2	43.6	39.3	87.8	82.8	75.1	70.3
Causal Norm [51]	✓			59.6	50.3	44.1	-	<u>88.5</u>	83.6	<u>80.6</u>	-
LADE [17]	✓			61.6	50.1	45.6	<u>40.7</u>	88.3	82.1	79.1	<u>73.9</u>
DRO [52]	✓			<u>63.4</u>	<u>57.6</u>	<u>47.3</u>	-	-	-	-	-
TDE + IDR [53]	✓			-	50.3	44.9	-	-	<u>84.5</u>	79.6	-
M2m [13]		✓		58.2	-	42.9	-	87.9	-	78.3	-
CAM [9]		✓		-	51.7	47.8	-	-	<u>83.6</u>	<u>80.0</u>	-
DiVE (2 Experts) [47]		✓	✓	<u>62.0</u>	51.1	45.4	-	-	-	-	-
CMO+RIDE (4 Experts) [14]		✓	✓	<u>60.2</u>	<u>53.0</u>	<u>50.0</u>	-	-	-	-	-
TSC [34]		✓		59.0	47.4	43.8	-	<u>88.7</u>	82.9	79.7	-
LDAM+DRW [27]	✓	✓		58.7	46.6	42.0	38.5	88.2	81.3	77.0	74.7
MiSLAS [28]	✓	✓		63.2	52.3	47.0	-	90.0	85.7	82.1	-
Prior-LT [11]	✓	✓		61.3	51.1	45.5	42.1	89.7	84.3	82.8	78.5
PaCo [33]	✓	✓		64.2	56.0	<u>52.0</u>	<u>47.8</u>	<u>91.5</u>	<u>88.0</u>	<u>85.4</u>	<u>82.3</u>
BCL [35]	✓	✓		<u>64.9</u>	<u>56.6</u>	<u>51.9</u>	-	91.1	<u>87.2</u>	<u>84.3</u>	-
GCL [38]	✓	✓		-	53.6	48.7	44.9	-	85.5	82.7	79.0
LFME (3 Experts) [18]			✓	57.8	47.2	42.3	39.0	87.1	81.5	75.3	72.9
BBN (2 Experts) [8]		✓	✓	59.1	47.0	42.6	-	88.3	82.2	79.8	-
RIDE (4 Experts) [19]			✓	61.8	51.7	48.0	44.6	86.3	83.7	81.2	77.8
Hybrid-SC (2 Experts) [32]		✓	✓	-	51.9	46.7	-	-	85.4	81.4	-
SADE (3 Experts) [43]	✓		✓	63.6	53.9	49.8	44.7	90.0	85.8	82.9	78.0
SSD (2 Experts) [48]			✓	62.3	50.5	46.0	-	-	-	-	-
ACE (4 Experts) [45]			✓	-	51.9	49.6	-	-	84.9	81.4	-
NCL (3 Experts) [20]	✓		✓	<u>63.8</u>	<u>58.2</u>	<u>54.2</u>	<u>49.5</u>	<u>91.1</u>	<u>87.3</u>	<u>85.5</u>	<u>82.2</u>
ECL (3 Experts)	✓	✓	✓	<b>67.3</b>	<b>59.9</b>	<b>56.3</b>	<b>51.4</b>	<b>91.8</b>	<b>88.9</b>	<b>86.5</b>	<b>83.6</b>

distance between the outputs from two experts for all features ( $\mathcal{A}_{y_i}$ ). Here, we calculate the distance between expert  $m$  and  $n$  as follows:

$$D_i^{m,n} = \frac{1}{|\mathcal{A}_{y_i}|} \cdot \sum_{\mathbf{v}_t \in \mathcal{A}_{y_i}} \|\mathbf{v}_t^m - \mathbf{v}_t^n\|_2. \quad (12)$$

The smaller  $D_i$  indicates that the experts learn stable feature representations w.r.t. class  $y_i$ , making it easier to finetune model heads on the downstream tasks.

**Expected Calibration Error.** Calibration indicates the model prediction reflects the actual likelihood of accuracy [59]. Let  $\hat{p}_i$  be the confidence of the image  $x_i$ , and divide dataset  $\mathcal{D}$  into several bin  $\mathcal{B}$  with size  $m$  according to the value of  $\hat{p}_i$ . Then, the reliability diagrams are proposed to visualize the model calibration by measuring the distance to the ideal  $\sum_{i \in \mathcal{B}_m} \mathbb{1}(\hat{y} = y_i) \equiv \sum_{i \in \mathcal{B}_m} \mathbb{1}(\hat{y} = y_i)$  for all  $m \in \{1, \dots, M\}$ . The Expected Calibration Error (ECE) is proposed to quantitatively measure classifiers' calibration:

$$ECE = \frac{1}{|\mathcal{D}|} \sum_{m=1}^M \sum_{i \in \mathcal{B}_m} |\mathbb{1}(\hat{y} = y_i) - \hat{p}_i|. \quad (13)$$

### C. Implementation Details

For CIFAR-LT, we follow LTR-WD [42] to set weight decay  $5e - 3$  for ResNet-32 and use stochastic gradient

descent with momentum 0.9. All models are trained for 200 epochs with the learning rate 0.01 and mini-batch 64. The learning scheduler is Cosine Annealing [60] with an ending rate of 0. Further, Cutout [61] and AutoAug [62] are used to compensate for origin data augmentation strategies [63]. We adopt the MoCo augmentation [22] for better image views in the contrastive branch. For large-scale datasets, we follow LTR-WD [42] to set weight decay  $5e - 4/1e - 4$  for ImageNet-LT/iNaturalist 2018 and train 180/90 epochs, respectively. We replace AutoAug with RandAug [64] while keeping other settings consistent with CIFAR-LT. Finally, we adopt horizontal flips as the post-hoc augmentation for better performance.

Following previous work [22], [33], [20], we set temperature factor  $\tau = 1$  and keep all MoCo hyper-parameters consistent with NCL. For the hyper-parameters setting of ECL, we set  $K = 3$  experts,  $\alpha = 0.6$  and  $\beta = 1.0$  by default. Results are averaged from 5 (CIFAR-LT) or 3 (large-scale datasets) random seeds.

### D. Competing Methods

**Baselines.** The vanilla baseline (CE) conducts plain training with standard cross-entropy loss [39]. The common networks are ResNet-32 (CIFAR-10/100-LT), ResNet-50 [63] (ImageNet-LT, iNaturalist 2018) and ResNeXt-50 [65]

TABLE II: Top-1 accuracy (%) on ImageNet-LT & iNaturalist 2018. Results are sorted by publication time. R-50: ResNet-50. RX-50: ResNeXt-50. Our ECL consistently outperforms state-of-the-art by a large margin.

Method	ImageNet-LT		iNaturalist2018
	R-50	RX-50	R-50
CE [39]	38.9	44.4	60.9
OLTR [56]	40.4	-	63.9
CB [39]	40.9	-	63.5
LDAM+DRW [27]	45.8	-	68.0
BBN [8]	48.3	49.3	66.3
NCM [24]	44.3	47.3	63.1
c-RT [24]	47.3	49.6	65.2
$\tau$ -Norm [24]	46.7	49.4	65.6
LWS [24]	47.7	49.7	65.9
BS [15]	53.0	-	66.4
RIDE (4 Expert) [19]	55.4	56.8	72.6
DisAlign [57]	52.9	53.4	70.6
DiVE [47]	53.1	-	71.7
SSD (2 Expert) [48]	-	56.0	71.5
ACE (4 Expert) [45]	54.7	56.6	72.9
PaCo [33]	56.1	57.2	72.2
TSC [34]	52.4	-	69.7
RIDE+CMO (4 Expert) [14]	56.2	-	72.8
BCL [35]	56.0	57.1	71.8
CKT [58]	-	54.2	-
GCL [38]	53.7	54.9	72.0
NCL (3 Expert) [20]	59.5	60.5	74.9
ECL (3 Expert)	<b>60.6</b>	<b>61.7</b>	<b>75.8</b>

TABLE III: Ablation study of ECL. We report ResNet32 on CIFAR100-LT ( $\gamma = 100$ ) and ResNet50 on ImageNet-LT. BKT: balanced knowledge transfer module. FLD: feature level distillation. CPT: contrastive proxy task.

BKT	FLD	CPT	CIFAR100-LT	$\Delta$	ImageNet-LT	$\Delta$
-	-	-	52.2	-	55.1	-
✓	-	-	53.7	+ 1.5	56.4	+ 1.3
-	✓	-	54.7	+ 2.5	58.9	+ 3.8
-	-	✓	54.2	+ 2.0	57.6	+ 2.5
✓	-	✓	54.9	+ 2.7	57.9	+ 2.8
-	✓	✓	55.8	+ 3.6	59.9	+ 4.8
✓	✓	✓	56.3	+ 4.1	60.6	+ 5.5

(ImageNet-LT). In addition, to align with previous works that contain some additional proposal-independent tricks implicitly, we adopt the same settings with NCL for all our reproduced results for fair comparisons.

**Feature-wise methods** modify the feature sampling or learning manners to cope with long-tailed datasets. M2m [13] generates pseudo samples for training and optimizing. CAM [9] and CMO [14] enrich the training samples via feature combination. DiVE [47] adopts knowledge distillation and takes the teacher feature as an additional training sample for the student model. Recent state-of-the-art [34], [33], [35] adopts contrastive frameworks to improve representation learning.

**Re-weight methods** focus on label weighting [39], [50], [27], [28] or logits adjusting [16], [17], [15], [11], [52], [53], [38] based on standard cross entropy loss. In addition, some methods [24], [51], [42] are also effective by directly adjusting the classifier’s weight.

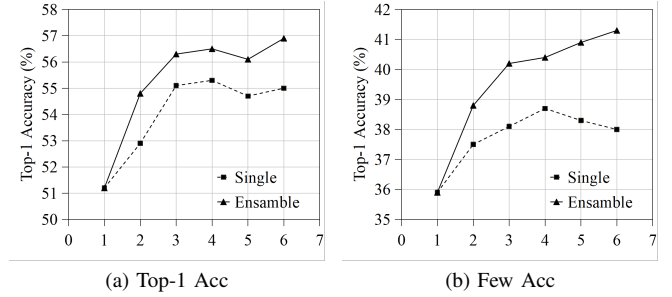


Fig. 4: Comparison of different expert number  $K$  on CIFAR100-LT ( $\gamma = 100$ ). The ensemble performance is computed based on the averaging logits of all experts. We report Top-1/Few-shot Acc and show that a larger expert number  $K$  brings higher model performance. We set  $K = 3$  to leverage the performance and training memory consumption.

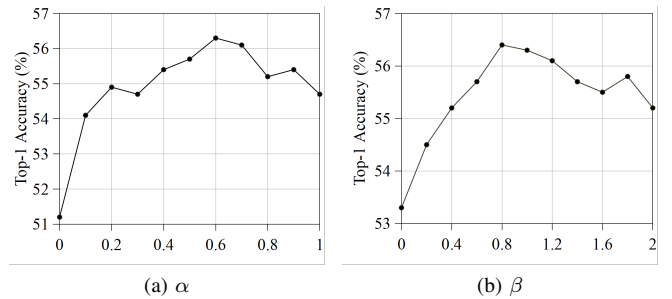


Fig. 5: Hyper-parameter analysis of  $\alpha$  and  $\beta$  on CIFAR100-LT ( $\gamma = 100$ ). We fix  $\beta = 1$  in subfigure (a) and  $\alpha = 0.6$  in subfigure (b).  $\alpha = 0$  means no collaborative learning in our ECL, which results in poor Top-1 Acc performance.

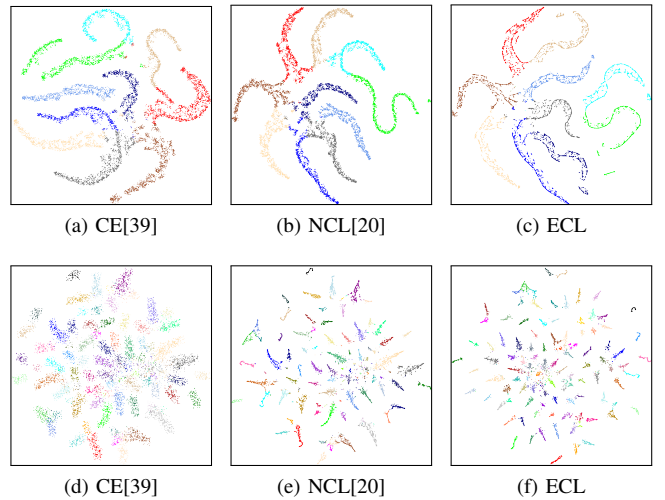


Fig. 6: Visualized t-SNE results of ResNet32 on CIFAR10-LT (a-c) and CIFAR100-LT (d-e). The scatters of the same color indicate the same categories. Our ECL shows better intra-class and inter-class distance to disentangle different categories.

**Multi-expert methods** have shown powerful generalization in LTR and can be classified into two categories. 1) Each expert learn *different* aspects of knowledge w.r.t. specific classes and then aggregates together [18], [8], [19], [45], [32]. 2) Each expert learns the *same* knowledge w.r.t. class to reduces the

uncertainty on minority classes [43], [48], [20]. Note that our ECL belongs to the latter.

### E. Comparison with state-of-the-art

We conduct comprehensive comparisons on CIFAR-LT (see Tab. I) and large-scale datasets (see Tab. II). For comparing methods, we report the performance in their original papers and reproduce the missing settings through their official code repositories. For contrastive approaches [33], [20], we keep the training epochs *consistent* with ours for fair comparisons. We group previous methods into 3 categories as discussed in Sec. II. ECL adopts RW (BC loss), FW (feature distillation), and ME (multi-expert architecture). Note that we report the ensemble results for all ME methods.

As illustrated in Tab. I-II, ECL outperforms previous approaches remarkably on all CIFAR-LT settings, ImageNet-LT, and iNaturalist 2018. Compared to state-of-the-art performance, ECL improves the NCL by 2.1% (CIFAR100-LT,  $\gamma = 100$ ), 1.1% (ImageNet-LT), and 0.9% (iNaturalist 2018) respectively. Compared to two-stage methods like MiS-LAS [28] and GCL [38], our ECL outperforms them in an end-to-end manner. Although we train the model in the multi-expert framework, we can adopt a single expert for evaluation without extra computation and memory consumption. We will discuss the single expert performance in Sec. VI-F.

### F. Further Analysis

**Ablation study of ECL.** We elaborately design three main modules to compound ECL, namely the Balanced Knowledge Transfer module (BKT), Feature Level Distillation (FLD), and Contrastive Proxy Task (CPT). We conduct extensive ablation experiments on CIFAR100-LT ( $\gamma = 100$ ) to demonstrate the contribution of each component. As Tab. III shows, our proposals are complementary to the performance, and FLD contributes the primary parts. As discussed in Sec. IV (observation 2), FLD promotes more robust feature learning without the toxicity from prior label bias. Like NCL [20] and PaCo [33], the CPT consistently improves model performance without inference burden. In addition, the BKT module consistently improves logit level distillation, allowing machine domain knowledge to be transferred with unbiased weight.

**Effect of expert number  $K$ .** We conducted experiments to explore the influence of expert number  $K$ . As Fig. 4 shows, the model performance improves consistently with larger  $K$ . When  $K = 1$ , the model is equal to the baseline with CPT without collaborative learning. When  $K = 2$ , the performance improves significantly, which firmly manifests the effectiveness of BKT and FLD. However, when  $K \geq 3$ , the single expert is difficult to get further improvement, especially on few-shot accuracy. Hence, we set  $K = 3$  to trade off the computational overhead and model performance.

**Hyper-parameters analysis.** In the final loss (Eq. 11), we trade off the collaborative learning with  $\alpha$  and contrastive learning with  $\beta$ . Fig. 5 is designed to search for the optimal value on CIFAR100-LT ( $\gamma = 100$ ). In Fig. 5a, we set  $\beta = 1$

TABLE IV: Performance comparison with NCL in detail.

Dataset		CIFAR100-LT		ImageNet-LT	
Metric		Acc $\uparrow$	ECE $\downarrow$	Acc $\uparrow$	ECE $\downarrow$
NCL [20]	Single	53.6	5.11	57.7	3.80
ECL		55.1 (+1.5)	2.33 (-2.78)	59.3 (+1.6)	1.96 (-1.84)
NCL [20]	Ensamble	54.4	4.62	59.5	2.92
ECL		56.3 (+1.9)	1.82 (-2.80)	60.6 (+1.1)	1.33 (-1.59)

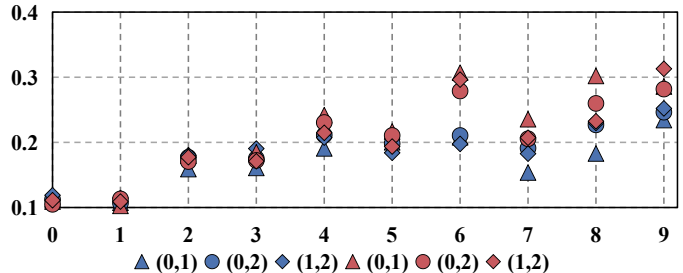


Fig. 7: The class-wise feature  $l_2$  distance between each expert pair on CIFAR10-LT. The red/blue points indicate NCL and ECL. ECL effectively reduces the differences between different experts on the feature of the same images.

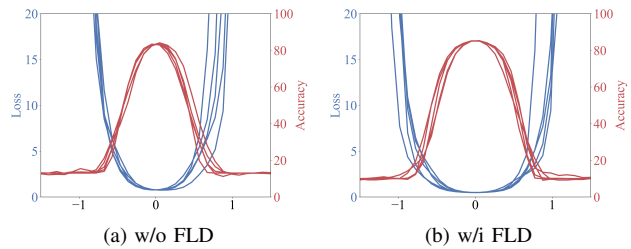


Fig. 8: The loss/accuracy landscapes of ECL without (a) or with (b) feature level distillation. All plots contain 5 landscapes with 5 randomly generated directions.

by default. When  $\alpha = 0$ , the model is degraded to the baseline with CPT. Top-1 Acc. increases rapidly when we add distillation loss ( $\alpha > 0$ ). The best trade-off between distillation and classification loss achieves at  $\alpha = 0.6$ . In Fig. 5b, we set  $\alpha = 0.6$  by default. The best performance is achieved when  $\beta \sim 1$ , which shows a balance between classification and instance discrimination.

**Feature representation quality.** In Tab. III, we notice that the feature level distillation is crucial in ECL. To delve into its mechanism, we conduct visualization experiments on CIFAR10-LT and CIFAR100-LT ( $\gamma = 100$ ) in Fig. 6. Specifically, we utilize t-SNE[66] to map the  $K$ -dimensional features into 2D distribution for visualization. Fig. 6a & 6d show that the baseline cannot achieve satisfactory clustering results where few-shot categories are coupled together. The poor inter-class distance prevents further performance gains of the classifier. Note that collaborative learning (NCL) remarkably alleviates this issue as shown in Fig. 6b & 6e. Our ECL further contributes to more compact intra-class distributions and enlarges the inter-class distance (Fig. 6c & 6f), which demonstrates that ECL provides higher quality features.

In addition, we visualize the feature distance among each expert and summarize the average distance w.r.t. class index, which is sorted by instances number. As Fig. 7 shows, all ex-



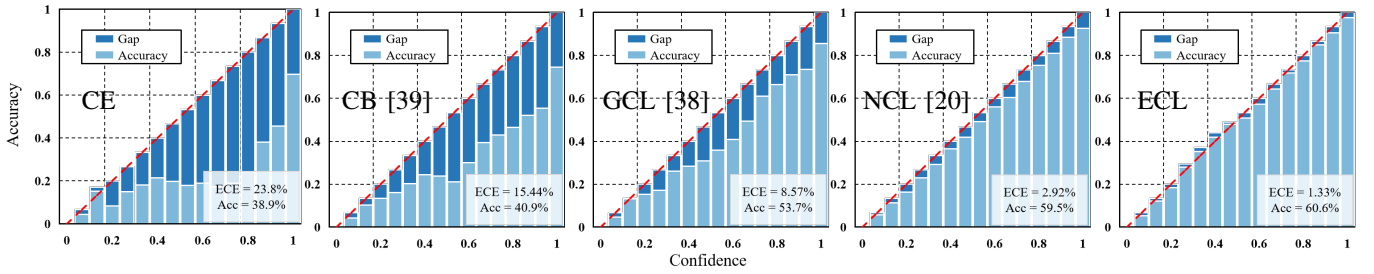


Fig. 9: Reliability diagrams on ImageNet-LT with 15 bins. We select ResNet50 models trained via plain CE, CB, GCL, NCL, and our ECL. The prediction probabilities of our ECL indicate optimal expected costs in Bayesian decision scenarios.

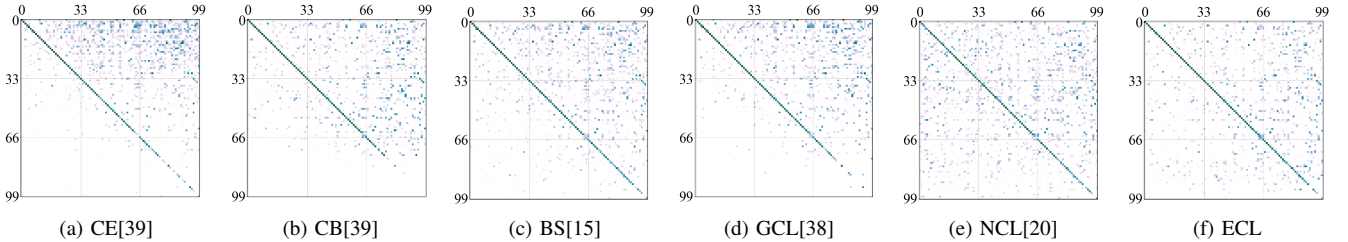


Fig. 10: Visualized log-confusion matrix on CIFAR100-LT ( $\gamma = 100$ ).  $x$ -axis: ground truth.  $y$ -axis: predicted label. The deeper color indicates larger values. ECL shows the best class accuracy and the most balanced misclassification distribution.

perts extract similar features of many-shot samples. However, the feature representations present differentiated distribution in few-shot samples. Our ECL alleviates this problem remarkably via feature-level distillation, yielding its better classification performance.

**Loss/Accuracy landscapes.** To validate the model robustness, we adopt the tool in [49] to visualize the loss/accuracy landscapes of models with/without feature level distillation. We conduct experiments on CIFAR10-LT ( $\gamma = 100$ ) based on our ECL. As described in Sec. VI-B, we perturb the model weights by a series of Gaussian noises with varying degrees. As Fig. 8 shows, it turns out that the loss/accuracy landscapes become much flatter if we adopt the feature level distillation on ECL. This observation demonstrates that the distillation operation help models to extract more robust representations to overcome the random noise perturbation.

**Model calibration.** In Fig. 9, we present the reliability diagrams with 15 bins on the ImageNet-LT. For all models in comparison, the accuracy bars are below the ideal  $y = x$  red line, which indicates that the models are all overconfident in their predictions. Compared to baseline CE, all methods alleviate the overconfidence issue and promote model calibration to some extent. Compared to NCL, ECL further reduces ECE, which demonstrates our success in regulating all classes. We present more detailed comparisons to the state-of-the-art on the best single model and ensembles in Tab. IV. Our ECL consistently outperforms the NCL in either single and ensemble views, and the single expert of ECL achieves comparable performance with the NCL ensemble.

**Do long-tail problems get alleviated?** One of the primary goals of LTR is to improve performance in few-shot categories. Hence, we plot the confusion matrices on CIFAR100-LT. For better visualizations, we adopt logarithmic operations for all matrix values. In Fig. 10, the baseline (10a) prefers to

train a trivial predictor, which simplifies images as many-shot labels to minimize the error rate. Several recent methods (10b-10e) alleviate such issues to some extent. Compared to them, our proposal (10f) shows the best accuracy (diagonal) and a more balanced misclassification distribution (non-diagonal). It firmly demonstrates our superiority in erasing the bias in LTR and our success in regularizing the few-shot classes.

## VII. CONCLUSION

This paper systematically analyzes the multi-expert framework in the long tail visual recognition, which trains several experts collaboratively to overcome the model preference for the majority and the high uncertainty on the minority. We point out that there is imbalanced knowledge transfer among experts' distillation, which leads to the inconspicuous improvement of collaborative learning on tail performance. A balanced distillation loss is proposed to improve the efficiency of collaborative learning by comparing two classifiers' predictions, which are supervised by different signals. Furthermore, we claim that distillation at the feature level will greatly improve the feature quality and model performance. To learn representations more thoroughly, we integrate a contrastive proxy task and finally propose an effective collaborative learning framework, which helps the model extract robust features and learn meticulous distinguishing ability. We conduct both quantitative and qualitative experiments on four standard datasets to verify the superiority and effectiveness of ECL. Extensive experiments and visualizations demonstrate that ECL achieves state-of-the-art performance with better feature representations.

## REFERENCES

- [1] J. Wu, L. Song, Q. Zhang, M. Yang, and J. Yuan, "Forestdet: Large-vocabulary long-tailed object detection and instance segmentation," *IEEE Transactions on Multimedia*, vol. 24, pp. 3693–3705, 2021.

- [2] X. Zhang, C. Zhu, H. Wu, Z. Liu, and Y. Xu, "An imbalance compensation framework for background subtraction," *IEEE Transactions on Multimedia*, vol. 19, no. 11, pp. 2425–2438, 2017.
- [3] P. Wang, Z. Zhao, F. Su, and H. Meng, "Ltreid: Factorizable feature generation with independent components for long-tailed person re-identification," *IEEE Transactions on Multimedia*, 2022.
- [4] M. Ding, S. Zhang, and J. Yang, "Improving pedestrian detection from a long-tailed domain perspective," in *Proceedings of the 29th ACM International Conference on Multimedia*, 2021, pp. 2918–2926.
- [5] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei, "ImageNet Large Scale Visual Recognition Challenge," *IJCV*, vol. 115, no. 3, pp. 211–252, 2015.
- [6] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *ECCV*. Springer, 2014, pp. 740–755.
- [7] B. Zhou, A. Lapedriza, A. Khosla, A. Oliva, and A. Torralba, "Places: A 10 million image database for scene recognition," *IEEE TPAMI*, 2017.
- [8] B. Zhou, Q. Cui, X.-S. Wei, and Z.-M. Chen, "Bbn: Bilateral-branch network with cumulative learning for long-tailed visual recognition," in *CVPR*, 2020, pp. 9719–9728.
- [9] Y. Zhang, X.-S. Wei, B. Zhou, and J. Wu, "Bag of tricks for long-tailed visual recognition with deep convolutional neural networks," in *AAAI*, 2021, pp. 3447–3455.
- [10] J. Gao, J. Chen, H. Fu, and Y.-G. Jiang, "Dynamic mixup for multi-label long-tailed food ingredient recognition," *IEEE Transactions on Multimedia*, 2022.
- [11] Z. Xu, Z. Chai, C. Yuan *et al.*, "Towards calibrated model for long-tailed visual recognition from prior perspective," *NeurIPS*, vol. 34, pp. 7139–7152, 2021.
- [12] P. Chu, X. Bian, S. Liu, and H. Ling, "Feature space augmentation for long-tailed data," in *ECCV*. Springer, 2020, pp. 694–710.
- [13] J. Kim, J. Jeong, J. Shin *et al.*, "M2m: Imbalanced classification via major-to-minor translation," in *CVPR*, 2020, pp. 13 896–13 905.
- [14] S. Park, Y. Hong, B. Heo, S. Yun, and J. Y. Choi, "The majority can help the minority: Context-rich minority oversampling for long-tailed classification," in *CVPR*, 2022, pp. 6887–6896.
- [15] J. Ren, C. Yu, X. Ma, H. Zhao, S. Yi *et al.*, "Balanced meta-softmax for long-tailed visual recognition," *NeurIPS*, vol. 33, pp. 4175–4186, 2020.
- [16] A. K. Menon, S. Jayasumana, A. S. Rawat, H. Jain, A. Veit, and S. Kumar, "Long-tail learning via logit adjustment," in *ICLR*, 2021.
- [17] Y. Hong, S. Han, K. Choi, S. Seo, B. Kim, and B. Chang, "Disentangling label distribution for long-tailed visual recognition," in *CVPR*. Computer Vision Foundation / IEEE, 2021, pp. 6626–6636.
- [18] L. Xiang, G. Ding, J. Han *et al.*, "Learning from multiple experts: Self-paced knowledge distillation for long-tailed classification," in *ECCV*. Springer, 2020, pp. 247–263.
- [19] X. Wang, L. Lian, Z. Miao, Z. Liu, and S. Yu, "Long-tailed recognition by routing diverse distribution-aware experts," in *ICLR*, 2021.
- [20] J. Li, Z. Tan, J. Wan, Z. Lei, and G. Guo, "Nested collaborative learning for long-tailed visual recognition," in *CVPR*, 2022, pp. 6949–6958.
- [21] Y. Niu, L. Chen, C. Zhou, and H. Zhang, "Respecting transfer gap in knowledge distillation," in *NeurIPS*, 2022.
- [22] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, "Momentum contrast for unsupervised visual representation learning," in *CVPR*, 2020, pp. 9729–9738.
- [23] H. Han, W. Wang, B. Mao *et al.*, "Borderline-smote: A new over-sampling method in imbalanced data sets learning," in *ICIC*, ser. Lecture Notes in Computer Science, vol. 3644. Springer, 2005, pp. 878–887.
- [24] B. Kang, S. Xie, M. Rohrbach, Z. Yan, A. Gordo, J. Feng, and Y. Kalantidis, "Decoupling representation and classifier for long-tailed recognition," in *ICLR*, 2020.
- [25] J. Wang, T. Lukasiewicz, X. Hu, J. Cai, and Z. Xu, "RSG: A simple but effective module for learning imbalanced datasets," in *CVPR*. Computer Vision Foundation / IEEE, 2021, pp. 3784–3793.
- [26] J. Hao, C. Wang, G. Yang, Z. Gao, J. Zhang, and H. Zhang, "Annealing genetic gan for imbalanced web data learning," *IEEE Transactions on Multimedia*, vol. 24, pp. 1164–1174, 2021.
- [27] K. Cao, C. Wei, A. Gaidon, N. Arechiga, and T. Ma, "Learning imbalanced datasets with label-distribution-aware margin loss," *NeurIPS*, vol. 32, 2019.
- [28] Z. Zhong, J. Cui, S. Liu, and J. Jia, "Improving calibration for long-tailed recognition," in *CVPR*. Computer Vision Foundation / IEEE, 2021, pp. 16 489–16 498.
- [29] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *ICML*. PMLR, 2020, pp. 1597–1607.
- [30] Z. Tao, X. Liu, Y. Xia, X. Wang, L. Yang, X. Huang, and T.-S. Chua, "Self-supervised learning for multimedia recommendation," *IEEE Transactions on Multimedia*, 2022.
- [31] Y. Yang, Z. Xu *et al.*, "Rethinking the value of labels for improving class-imbalanced learning," *NeurIPS*, vol. 33, pp. 19 290–19 301, 2020.
- [32] P. Wang, K. Han, X.-S. Wei, L. Zhang, and L. Wang, "Contrastive learning based hybrid networks for long-tailed image classification," in *CVPR*, 2021, pp. 943–952.
- [33] J. Cui, Z. Zhong, S. Liu, B. Yu, and J. Jia, "Parametric contrastive learning," in *ICCV*, 2021, pp. 715–724.
- [34] T. Li, P. Cao, Y. Yuan, L. Fan, Y. Yang, R. S. Feris, P. Indyk, and D. Katabi, "Targeted supervised contrastive learning for long-tailed recognition," in *CVPR*, 2022, pp. 6918–6928.
- [35] J. Zhu, Z. Wang, J. Chen, Y.-P. P. Chen, and Y.-G. Jiang, "Balanced contrastive learning for long-tailed visual recognition," in *CVPR*, 2022, pp. 6908–6917.
- [36] M. A. Jamal, M. Brown, M.-H. Yang, L. Wang, and B. Gong, "Rethinking class-balanced methods for long-tailed visual recognition from a domain adaptation perspective," in *CVPR*, 2020, pp. 7610–7619.
- [37] X. Li, J. Li, L. Zhu, G. Wang, and Z. Huang, "Imbalanced source-free domain adaptation," in *Proceedings of the 29th ACM International Conference on Multimedia*, 2021, pp. 3330–3339.
- [38] M. Li, Y.-m. Cheung, Y. Lu *et al.*, "Long-tailed visual recognition via gaussian clouded logit adjustment," in *CVPR*, 2022, pp. 6929–6938.
- [39] Y. Cui, M. Jia, T.-Y. Lin, Y. Song, and S. Belongie, "Class-balanced loss based on effective number of samples," in *CVPR*, 2019, pp. 9268–9277.
- [40] Z. Peng, W. Huang, Z. Guo, X. Zhang, J. Jiao, and Q. Ye, "Long-tailed distribution adaptation," in *Proceedings of the 29th ACM International Conference on Multimedia*, 2021, pp. 3275–3282.
- [41] J. Sun, W. Yang, J.-H. Xue, and Q. Liao, "An equalized margin loss for face recognition," *IEEE Transactions on Multimedia*, vol. 22, no. 11, pp. 2833–2843, 2020.
- [42] S. Alshammari, Y.-X. Wang, D. Ramanan, and S. Kong, "Long-tailed recognition via weight balancing," in *CVPR*, 2022, pp. 6897–6907.
- [43] Y. Zhang, B. Hooi, L. Hong, and J. Feng, "Self-supervised aggregation of diverse experts for test-agnostic long-tailed recognition," *NeurIPS*, vol. 35, pp. 34 077–34 090, 2022.
- [44] A. Iscen, A. Araujo, B. Gong, and C. Schmid, "Class-balanced distillation for long-tailed visual recognition," in *BMVC*. BMVA Press, 2021, p. 165.
- [45] J. Cai, Y. Wang, J.-N. Hwang *et al.*, "Ace: Ally complementary experts for solving long-tailed recognition in one-shot," in *ICCV*, 2021, pp. 112–121.
- [46] G. Hinton, O. Vinyals, J. Dean *et al.*, "Distilling the knowledge in a neural network," *arXiv preprint arXiv:1503.02531*, vol. 2, no. 7, 2015.
- [47] Y.-Y. He, J. Wu, X.-S. Wei *et al.*, "Distilling virtual examples for long-tailed recognition," in *ICCV*, 2021, pp. 235–244.
- [48] T. Li, L. Wang, and G. Wu, "Self supervision to distillation for long-tailed visual recognition," in *ICCV*, 2021, pp. 630–639.
- [49] H. Li, Z. Xu, G. Taylor, C. Studer, and T. Goldstein, "Visualizing the loss landscape of neural nets," in *NeurIPS*, 2018.
- [50] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *ICCV*, 2017, pp. 2980–2988.
- [51] K. Tang, J. Huang, and H. Zhang, "Long-tailed classification by keeping the good and removing the bad momentum causal effect," *NeurIPS*, vol. 33, pp. 1513–1524, 2020.
- [52] D. Samuel, G. Chechik *et al.*, "Distributional robustness loss for long-tail learning," in *ICCV*, 2021, pp. 9495–9504.
- [53] S. Yu, J. Guo, R. Zhang, Y. Fan, Z. Wang, and X. Cheng, "A re-balancing strategy for class-imbalanced classification based on instance difficulty," in *CVPR*, 2022, pp. 70–79.
- [54] A. Krizhevsky, G. Hinton *et al.*, "Learning multiple layers of features from tiny images," 2009.
- [55] G. Van Horn, O. Mac Aodha, Y. Song, Y. Cui, C. Sun, A. Shepard, H. Adam, P. Perona, and S. Belongie, "The inaturalist species classification and detection dataset," in *CVPR*, 2018, pp. 8769–8778.
- [56] Z. Liu, Z. Miao, X. Zhan, J. Wang, B. Gong, and S. X. Yu, "Large-scale long-tailed recognition in an open world," in *CVPR*, 2019.
- [57] S. Zhang, Z. Li, S. Yan, X. He, and J. Sun, "Distribution alignment: A unified framework for long-tail visual recognition," in *CVPR*, 2021.
- [58] S. Parisot, P. M. Esperança, S. McDonagh, T. J. Madarasz, Y. Yang, and Z. Li, "Long-tail recognition via compositional knowledge transfer," in *CVPR*, 2022, pp. 6939–6948.
- [59] A. Ashukha, A. Lyzhov, D. Molchanov, and D. Vetrov, "Pitfalls of in-domain uncertainty estimation and ensembling in deep learning," in *ICLR*, 2020.

- [60] I. Loshchilov and F. Hutter, “SGDR: Stochastic gradient descent with warm restarts,” in *ICLR*, 2017.
- [61] T. DeVries, G. W. Taylor *et al.*, “Improved regularization of convolutional neural networks with cutout,” *arXiv preprint arXiv:1708.04552*, 2017.
- [62] Y. Chen, Y. Li, T. Kong, L. Qi, R. Chu, L. Li, and J. Jia, “Scale-aware automatic augmentation for object detection,” in *CVPR*, 2021.
- [63] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *CVPR*, 2016, pp. 770–778.
- [64] E. D. Cubuk, B. Zoph, J. Shlens, and Q. V. Le, “Randaugment: Practical automated data augmentation with a reduced search space,” in *CVPR workshops*, 2020, pp. 702–703.
- [65] S. Xie, R. Girshick, P. Dollár, Z. Tu, and K. He, “Aggregated residual transformations for deep neural networks,” in *CVPR*, 2017, pp. 1492–1500.
- [66] L. Van der Maaten and G. Hinton, “Visualizing data using t-sne,” *Journal of machine learning research*, vol. 9, no. 11, 2008.