Enhancing Mixture of Experts with Independent and Collaborative Learning for Long-Tail Visual Recognition

Yanhao Chen[1,†], **Zhongquan Jian**[2,3,†], **Nianxin Ke**[1], **Shuhao Hu**[1], **Junjie Jiao**[1], **Qingqi Hong**[1,3,*] and **Qingqiang Wu**[1,2,3,4,*]
[1]School of Film, Xiamen University, Xiamen, China
[2]School of Informatics, Xiamen University, Xiamen, China
[3]Institute of Artificial Intelligence, Xiamen University, Xiamen, China
[4]Key Laboratory of Digital Protection and Intelligent Processing of Intangible Cultural Heritage of Fujian and Taiwan, Ministry of Culture and Tourism, Xiamen University, Xiamen, China
{cyhao, jianzq, kenianxin, horacehsh, jiaojj1204}@stu.xmu.edu.cn, {hongqq, wuqq}@xmu.edu.cn

Abstract

Deep neural networks (DNNs) face substantial challenges in Long-Tail Visual Recognition (LTVR) due to the inherent class imbalances in real-world data distributions. The Mixture of Experts (MoE) framework has emerged as a promising approach to addressing these issues. However, in MoE systems, experts are typically trained to optimize a collective objective, often neglecting the individual optimality of each expert. This individual optimality usually contributes to the overall performance, as the goals of different experts are not mutually exclusive. We propose the Independent and Collaborative Learning (ICL) framework to optimize each expert independently while ensuring global optimality. First, Diverse Optimization Learning (DOL) is introduced to enhance expert diversity and individual performance. Then, we conceptualize experts as parallel circuit branches and introduce Competition and Collaboration Learning (CoL). Competition Learning amplifies the gradients of better-performing experts to preserve individual optimality, and Collaboration Learning encourages collaboration through mutual distillation to enhance optimal knowledge sharing. ICL achieves state-of-the-art accuracy in experiments on CIFAR-100/10-LT, ImageNet-LT, and iNaturalist 2018, respectively. Our code is available at https://github.com/PolarisLight/ICL.

1 Introduction

Deep neural networks (DNNs) have achieved significant advancements, largely driven by the availability of large-scale datasets, elegant model architectures, and efficient optimization algorithms. These developments have established DNNs as a foundation for key computer vision tasks, including object detection, semantic segmentation, and image classification. The robust performance of DNNs often relies on training with balanced datasets such as ImageNet [Liu *et al.*, 2019], MS COCO [Lin *et al.*, 2014], and Places [Zhou *et al.*, 2018]. However, data often exhibits an unexpected long-tailed distribution in real-world scenarios, with significant variation in the number of samples per class. Under such conditions, DNNs are prone to overfitting to head classes, resulting in poor performance on tail classes, which frequently hold greater practical importance. To address these challenges, Long-Tail Visual Recognition (LTVR) has emerged as a pivotal research area in computer vision, which seeks to enhance the model's performance on tail classes while preserving accuracy on head classes, thus improving their generalization and practical applicability.

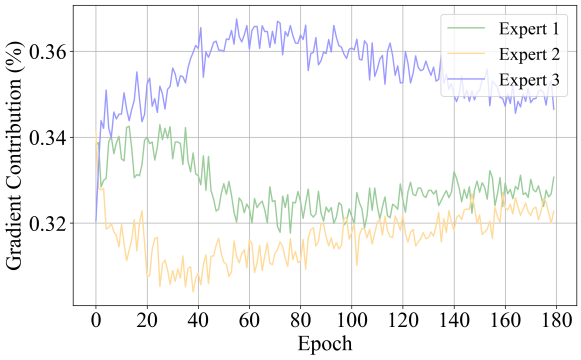


Figure 1: Gradient contributions of individual experts in existing MoE models (*e.g.*, with three experts).

To overcome the challenge of imbalanced data distribution, initial studies have attempted to correct this imbalance by employing strategies such as resampling [Wang *et al.*, 2020] and reweighting [Ren *et al.*, 2020; Park *et al.*, 2021]. The fundamental objective of these methods is to allocate greater emphasis on the tail categories, to enhance their performance by augmenting the weight or number of samples in these categories. However, this approach has been observed to harm the performance of the head category to a certain extent. Existing research indicates that over-sampling the tail category, particularly when the tail category possesses a lim-

*Corresponding Authors

ited number of samples, can result in significant overfitting issues. To mitigate these limitations, recent efforts in LTVR have focused on utilizing the Mixture of Experts (MOE). This approach's fundamental premise involves promoting diversity among experts and integrating their predictions to ensure the generation of more confident and reliable final decisions. Early MoE approaches tended to have each expert model focus on a different subset of data [Cai *et al.*, 2021; Cui *et al.*, 2023; Li *et al.*, 2022a], avoiding individual experts from facing categories with too much variation in the amount of data. In contrast, recent research [Jin *et al.*, 2023; Tan *et al.*, 2024] has adopted a shift towards a unified training approach, where all experts are exposed to the same data, to mitigate the uncertainty in model predictions.

However, existing MoE approaches emphasize the need for the experts' overall optimal performance to ensure robust model performance, ignoring the fact that each expert's goals are not mutually exclusive. We conducted a quick experiment to reveal different expert contributions during multi-expert model training. The mainstream ResNet-32 network with shared shallow and 3 experts' independent deep feature extractors was used to train 180 epochs using cross-entropy, and the average gradient share of the last convolution layer of each branch was used as the contribution of that branch, as shown in Figure 1. The final individual accuracies of the experts were 45.23%, 44.68%, and 45.36%, with the overall system accuracy reaching 47.56%. This indicates a positive correlation between gradient contribution and learning outcomes. During training, Expert 3 consistently dominated and contributed the most. However, this imbalance caused the stronger expert's optimization to stagnate prematurely, ultimately limiting the overall system performance, which depends on the joint optimization of all components. On this basis, we ask the following questions: 1) *How to drive individual experts to achieve optimization?* and 2) *How to ensure that the overall effect of the MoE system is improved under the premise of individual optimization of the experts?*

For this purpose, we propose an Independent and Collaborative Learning (ICL) framework that incorporates Diverse Optimization Learning (DOL) and Competition and Collaboration Learning (CoL). In response to the first question, DOL employs Adaptive Diversity (AD) to enhance the diversity of the experts. This ensures that each expert becomes proficient at capturing unique and complementary features, thereby strengthening the overall diversity. At the same time, to optimize the respective domains of the experts, DOL adopts Confusion Contrastive Learning (CCL), which treats the most confusing categories as negative samples and uses contrastive learning to enhance the discriminative capability of each expert. For the second question, CoL treats experts as parallel branches of a circuit, each offering a unique perspective on the input data to aid learning. It introduces parallel loss to minimize the impact of branches with larger losses, allowing those with smaller losses to influence overall optimization, thus preserving individual optimality. Mutual distillation is introduced to enhance collaboration and prevent dominant experts from falling into local optima. Minimizing the Kullback-Leibler (KL) divergence among experts' predictions fosters consistency and knowledge sharing. Our contributions can be summarized as follows:

- We propose Diverse Optimization Learning (DOL), which encourages experts to personalize optimization in unique domains from different perspectives through representation learning.
- We introduce Competition and Collaboration Learning (CoL), which allows experts to dynamically adjust their contributions to training among themselves to maintain individual strengths, while encouraging knowledge sharing to drive overall optimization.
- We demonstrate the effectiveness of our method on CIFAR-100/10-LT, ImageNet-LT, and iNaturalist 2018 datasets, showcasing its superior performance in LTVR task.

2 Related Work

2.1 Long-Tail Visual Recognition

LTVR seeks to improve tail class accuracy without harming head class performance. Early methods addressed class imbalance using re-sampling and re-weighting strategies, re-sampling methods [Kang *et al.*, 2020; Zang *et al.*, 2021] balance the distribution by oversampling tail classes or undersampling head classes, while re-weighting methods [Ren *et al.*, 2020; Park *et al.*, 2021] adjust class-specific loss weights to focus on fewer classes during training. In addition, data augmentation techniques [Zhang *et al.*, 2018; Verma *et al.*, 2019] transfer information from head to tail classes or generate synthetic tail class samples. Logit adjustment methods [Li *et al.*, 2022c; Li *et al.*, 2023] post-process logits to enhance inter-class variation. However, these methods often improve tail class accuracy at the cost of head class performance.

Recently, MoE-based models have gained attention in LTVR due to their ability to integrate multiple experts, each focusing on different subsets of data or features. For instance, BBN [Zhou *et al.*, 2020] uses a two-branch network to handle long-tailed and balanced distributions. ACE [Cai *et al.*, 2021] and ResLT [Cui *et al.*, 2023] enhance diversity and performance by specializing experts for different parts of the long-tailed distribution. RIDE [Wang *et al.*, 2021] and TLC [Li *et al.*, 2022a] reduce model variance by combining predictions from independently learned experts. SHIKE [Jin *et al.*, 2023] increases diversity by transferring shallow knowledge into deep feature extractors, while NCL [Li *et al.*, 2022b; Tan *et al.*, 2024] improves expert effectiveness by enforcing intra-expert consistency through output standardization. Despite these advancements, challenges remain in optimizing both individual and collective expert performance.

2.2 Knowledge distillation

Knowledge Distillation (KD) [Hinton *et al.*, 2015] is a technique for optimizing model performance by transferring knowledge from a large model to a small model. Self-Distillation (SD) [Zhang *et al.*, 2021], a variant of KD, allows the model itself to transfer knowledge through multi-level feature representation, which improves the optimization efficiency and performance of the model. In recent years,

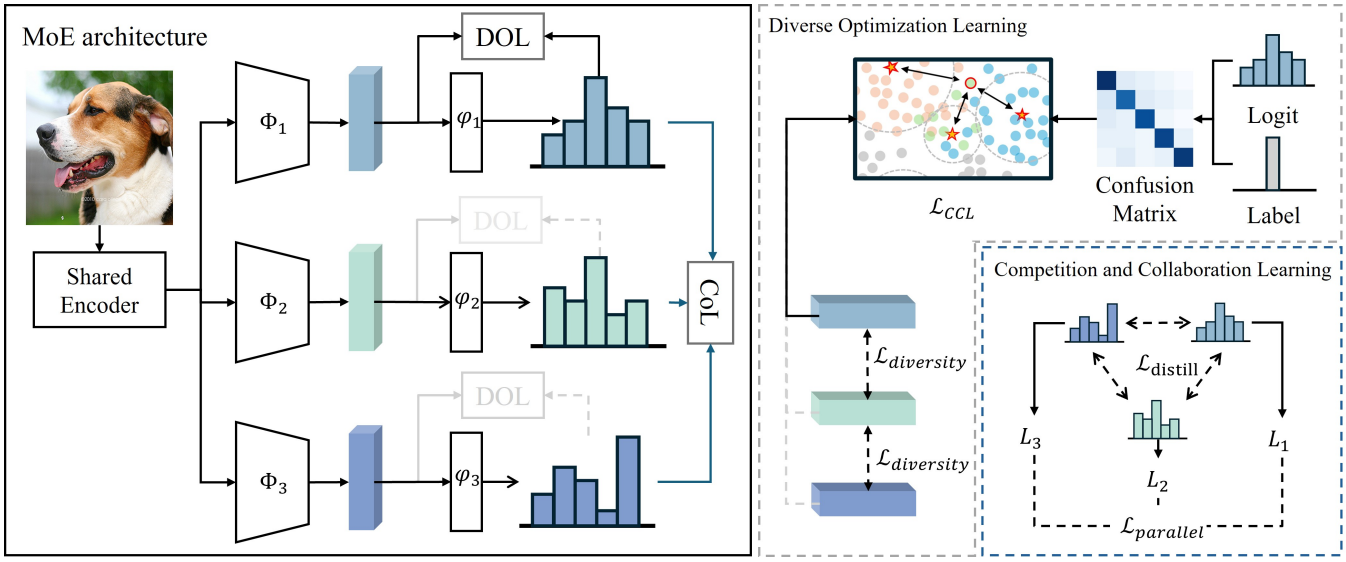


Figure 2: The proposed ICL contains two modules: DOL and CoL. Each expert in the MoE performs diversity optimization and confusion contrastive learning in the feature space and then applies competition and collaboration learning at the logits level for better model optimization from the individual to the overall.

distillation among experts has been introduced into MoE systems [Jin *et al.*, 2023; Tan *et al.*, 2024] to facilitate knowledge sharing and collaborative learning among experts by minimizing the KL dispersion among different experts. These methods improve both the performance of individual experts and the overall performance of the entire model.

2.3 Contrastive Learning

Contrastive Learning (CL) [Chen *et al.*, 2020], an effective representation learning method, improves the discriminative properties of features by maximizing the similarity of positive sample pairs and minimizing the similarity of negative sample pairs. CL is widely used in LTVR [Cui *et al.*, 2021; Hou *et al.*, 2023] to enhance the discriminative properties of features and improve the performance of the model. Prototypical Contrastive Learning (PCL) [Snell *et al.*, 2017] further enhances the aggregation of samples of the same class and the separation of samples of different classes by introducing category prototypes. This method has been shown to further improve the discriminative properties of features and the model's performance by enhancing the aggregation of samples of the same class and the separation of samples of different classes [Zhu *et al.*, 2022].

3 Methodology

Figure 2 depicts the architecture of ICL, which consists of two main modules: 1) Diverse Optimization Learning (DOL), and 2) Competition and Collaboration Learning (CoL). DOL enables experts to achieve diverse individual optimizations through representation learning, while CoL preserves their strengths and guides the MoE system toward global optimal.

3.1 Preliminaries

We denote the training dataset as $\{(x_i, y_i)\}_{i=1}^{N}$, where each data point x_i has a corresponding label $y_i \in \{1, \ldots, C\}$. The total number of samples in the training set is $N = \sum_{c=1}^{C} n_c$, where n_c represents the number of samples belonging to class c. In the long-tailed scenarios, for clarity, we set the number of samples to decrease as the class indices increase, *i.e.*, $n_1 > n_2 > \cdots > n_C$.

Consistent with prior mainstream research, our MoE model adopts the structure where experts share a shallow feature extractor f and each expert possesses an individual deep feature extractor Ψ^m along with a classifier φ^m, where m denotes the m-th expert out of a total of M experts. Given an input x, the exclusive feature of expert m is obtained:

$$\mathbf{h}^m = \Phi^m(f(x)) \tag{1}$$

where $\mathbf{h}^m$ is the encoded feature representation derived by the m-th expert, and the logits are calculated by its classifier:

$$\mathbf{z}^m = \varphi^m(\mathbf{h}^m) \tag{2}$$

Typically, the class associated with the maximum logits value is chosen as the predicted class for each expert, and the final prediction is obtained by averaging the outputs of all experts.

3.2 Diverse Optimization Learning

We propose Diverse Optimization Learning (DOL) to enhance expert diversity and achieve individual optimization in the MoE framework. Unlike existing methods that rely on predefined strategies to increase expert diversity, We propose an **Adaptive Diversity** (AD) strategy to capture features from different classes dynamically. The AD loss is to encourage experts to learn non-overlapping features:

$$\mathcal{L}_{\text{AD}} = \sum_{m=1}^{M} \sum_{m'=1}^{M} \mathbb{I}_{m \neq m'} \frac{\mathbf{h}^m \cdot \mathbf{h}^{m'}}{\|\mathbf{h}^m\| \|\mathbf{h}^{m'}\|} \tag{3}$$

where $\mathbf{h}^m$ and $\mathbf{h}^{m'}$ denote feature representations from any two experts. $\mathbb{I}$ represents the indicator function, which is 1 when $m \neq m'$ and 0 otherwise.

Building upon this, we further propose **Confusion Contrastive Learning** (CCL) to optimize the feature representations of each expert. Specifically, Each expert m maintains a confusion matrix $\mathcal{C}^m \in \mathbb{R}^{C\times C}$, where each element $\mathcal{C}^m_{c,c'}$ represents the frequency with which class c is misclassified as class c' for expert m. In each batch, $\mathcal{C}^m_{c,c'}$ is updated using a momentum-based method, which ensures a smooth and stable estimate of class misclassification frequencies over time.

$$\mathcal{C}^m_{c,c'} = \alpha\mathcal{C}^m_{c,c'} + (1-\alpha)\textit{Count}(c \to c') \tag{4}$$

where α is the momentum coefficient, and $\textit{Count}(c \to c')$ denotes the function that counts the number of samples from class c misclassified as class c'. Meanwhile, a zero-initialized class prototypes $\mathbf{p}^m_c$ for each class c, which is updated as follows:

$$\mathbf{p}^m_c = \beta\mathbf{p}^m_c + (1-\beta)\mathbf{h}^m_c \tag{5}$$

where β is the momentum coefficient and $\mathbf{h}^m_c$ is the mean feature vector of class c in the current batch for expert m. Then, for each sample i belonging to the class c, the CCL loss for expert m is defined as:

$$\mathcal{L}_{\text{CCL}} = -\log \frac{\exp\left(\frac{\mathbf{h}^m_i \cdot \mathbf{p}^m_c}{\tau}\right)}{\exp\left(\frac{\mathbf{h}^m_i \cdot \mathbf{p}^m_c}{\tau}\right) + \sum_{c'\in\mathcal{N}(c)} \exp\left(\frac{\mathbf{h}^m_i \cdot \mathbf{p}^m_{c'}}{\tau}\right)} \tag{6}$$

where τ is the temperature parameter, and $\mathcal{N}(c)$ denotes the top k most confused classes for class c. This loss function encourages the sample features to move closer to their respective class prototypes while distancing them from the prototypes of the most confused negative classes, thereby enhancing feature discrimination. We denote the DOL loss as:

$$\mathcal{L}_{\text{DOL}} = \mathbb{I}_{t<t_{AD}}\mathcal{L}_{\text{AD}} + \mathbb{I}_{t>t_{CL}} \sum_{m=1}^{M} \mathcal{L}_{\text{CCL}} \tag{7}$$

where t is the training epoch. We have found experimentally that these two losses yield better performance when applied at specific stages of training, rather than throughout the entire training process. Therefore, we introduce two variables, t_{AD} and t_{CL}. By integrating AD and CCL, DOL not only promotes expert diversity but also optimizes each expert's feature representation within its domain, thereby enhancing generalization and performance, particularly for LTVR tasks.

3.3 Competition and Collaboration Learning

In multi-branch neural networks, each branch contributes to the overall learning process by providing diverse perspectives on the input data. Simply summing the losses from all branches may lead to an imbalance, as branches with larger losses dominate the optimization. To address this issue, we propose **Competition and Collaboration Learning** (CoL), aiming at balancing the gradient contributions among experts through an optimized loss function to achieve the goal of maintaining individual dominance while reinforcing the overall effect. **Competition Learning** is implemented by utilizing a parallel loss to allow superior branches to have a greater influence on the overall optimization. This is inspired by the parallel resistances in electrical circuits, analogous to how the effective resistance in parallel circuits is dominated by the smallest resistance. Specifically, for the cross-entropy losses of M expert $L = [L_1, L_2, \ldots, L_M]$, the parallel loss is defined as:

$$L_{\text{parallel}} = \left(\sum_{i=1}^{M} \frac{1}{L_i + \epsilon}\right)^{-1} \tag{8}$$

where $\epsilon > 0$ is a small constant added to ensure numerical stability when L_m is small or zero. Its gradient of L_i of expert i is:

$$\frac{\partial L_{\text{parallel}}}{\partial L_i} = \frac{1}{\left(\sum_{j=1}^{M} \frac{1}{L_j+\epsilon}\right)^2 (L_i+\epsilon)^2} \tag{9}$$

$\mathcal{L}_{\text{parallel}}$ ensures that the gradient contribution of each expert is inversely proportional to the square of its loss, meaning that experts with better performance contribute more. This mechanism helps to ensure that high-performing experts maintain their influence as the percentage of lost value declines, thereby guiding overall optimization more effectively.

Collaboration Learning is achieved by introducing mutual distillation [Hinton *et al.*, 2015], which involves minimizing KL divergence between the logits of different experts, encouraging them to align their predictions and facilitate knowledge sharing, thereby enhancing the collaborative effect among experts.

$$\mathcal{L}_{\text{distill}} = \frac{2}{M(M-1)} \sum_{m<m'} \text{KL}(\mathbf{z}^m \parallel \mathbf{z}^{m'}) \tag{10}$$

where $\mathbf{z}^m$ and $\mathbf{z}^{m'}$ are the predicted probability distributions of experts m and m', respectively. Hence, $\mathcal{L}_{\text{CoL}}$ is the combination of the parallel and mutual distillation loss:

$$\mathcal{L}_{\text{CoL}} = \mathcal{L}_{\text{parallel}} + \mathcal{L}_{\text{distill}} \tag{11}$$

CoL loss allows the model to harness the strengths of high-performing branches in driving the optimization process while ensuring consistent and complementary learning across all experts through knowledge sharing.

3.4 Overall Training Objective

Our training framework consists of two phases: feature extractor training and classifier fine-tuning. During the feature extractor training phase, the total loss comprises CoL loss and DOL loss:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{CoL}} + \mathcal{L}_{\text{DOL}} \tag{12}$$

In the classifier fine-tuning phase, we freeze the shared feature extractor Φ and the expert-specific deep feature extractors Ψ, retraining the classification layers φ. Hence, the optimization is to minimize the Logit Adjustment (LA) [Menon *et al.*, 2021] loss:

$$\mathcal{L}_{LA} = -\sum_{m=1}^{M} \log\left(\frac{n_y \exp\left(\mathbf{z}^m\right)}{\sum_{c=1}^{C} n_c \exp\left(\mathbf{z}^m_c\right)}\right) \tag{13}$$

Since the feature extractor is frozen, the DOL loss is not applicable. Additionally, CoL is not utilized at this stage because the classifiers are focused on learning the optimal classification from their respective branches, rather than engaging in collaboration.

Category	Method	CIFAR-100-LT			CIFAR-10-LT		
		100	50	10	100	50	10
Baseline	Cross Entropy	38.3	43.9	55.7	70.4	74.8	86.4
	Focal loss [Lin *et al.*, 2017]	38.7	46.2	–	74.6	79.3	–
Representation Learning	TSC [Li *et al.*, 2022d]	43.8	47.4	59.5	79.7	82.9	88.7
	BCL [Zhu *et al.*, 2022]	51.9	56.6	64.9	84.3	87.2	91.1
	SBCL [Hou *et al.*, 2023]	44.9	48.7	57.9	–	–	–
Re-balance	WD [Alshammari *et al.*, 2022]	52.4	57.4	67.9	–	–	–
	GCL [Li *et al.*, 2022c]	48.7	53.6	–	82.7	85.5	–
	KPS [Li *et al.*, 2023]	45.0	49.2	–	81.2	84.6	–
Data Augmentation	RISDA [Chen *et al.*, 2022]	50.2	53.8	62.4	79.9	84.2	89.4
	H2T [Li *et al.*, 2024]	51.4	55.5	–	–	–	–
	DBN-Mix [Baik *et al.*, 2024]	51.0	54.9	65.0	83.5	86.8	90.9
MoE	RIDE (3E) [Wang *et al.*, 2021]	49.1	–	–	–	–	–
	ACE (3E) [Cai *et al.*, 2021]	49.6	51.9	–	81.4	84.9	–
	TLC (4E) [Li *et al.*, 2022a]	49.8	–	–	80.4	–	–
	ResLT (3E) [Cui *et al.*, 2023]	49.7	54.5	63.7	–	–	–
	NCL (3E) [Li *et al.*, 2022b]	54.2	58.2	–	85.5	87.3	–
	SHIKE (3E) [Jin *et al.*, 2023]	<u>56.3</u>	<u>59.8</u>	–	–	–	–
	NCL++ (2E) [Tan *et al.*, 2024]	<u>56.3</u>	<u>59.8</u>	–	<u>87.2</u>	<u>88.8</u>	–
	Ours (2E)	<u>56.3</u>	59.7	<u>69.0</u>	86.4	88.5	<u>91.3</u>
	Ours (3E)	**57.6**	**61.3**	**69.3**	**87.9**	**89.7**	**91.9**

Table 1: Comparison results on CIFAR-100-LT and CIFAR-10-LT with imbalance factors of 100, 50, and 10.

4 Experiments

4.1 Datasets

In this study, we used CIFAR-100/10-LT [Cui *et al.*, 2019], ImageNet-LT [Liu *et al.*, 2019] and iNaturalist 2018 [Horn *et al.*, 2018], to assess the performance of the model in addressing class imbalance problems. **CIFAR-100/10-LT** is derived from the original CIFAR-100 and CIFAR-10 datasets, constructed with imbalance factors of 100, 50, and 10. These factors indicate that the largest class contains 100, 50, and 10 times more samples than the smallest class, respectively. **ImageNet-LT** (Img-LT) is based on the original ImageNet dataset and follows a Pareto distribution with an imbalance factor of 256 across 1,000 classes. The training set includes 115.8K samples, while the test set consists of 50K samples. **iNaturalist 2018** (iNat-LT) is a large-scale real-world dataset with 437.5K samples distributed in 8,142 categories, showing highly imbalanced distribution characteristics suitable for simulating real-world data imbalance scenarios.

4.2 Implementation Details

The CIFAR-100-LT and CIFAR-10-LT dataset uses ResNet-32 [He *et al.*, 2016] as the base network, followed by the application of AutoAugment [Cubuk *et al.*, 2019] and Cutout [Devries and Taylor, 2017] techniques. In contrast, the Img-LT and iNat-LT datasets employ the ResNet-50 model and utilize RandAugment [Cubuk *et al.*, 2020] for data augmentation following NCL [Li *et al.*, 2022b].

The learning rates for these four datasets were 0.05, 0.01, 0.2, and 0.025, and all models were trained for 180 epochs. The classifiers were then retrained based on $\mathcal{L}_{LA}$ with an additional 20 epochs, with the feature extractor frozen. All experiments utilized the SGD optimizer with a momentum of 0.9, a cosine learning rate scheduler that decays to zero, and a weight decay factor of 5e-4. During the classifier training phase, we restarted the cosine learning rate scheduler. Momentum coefficients α and β are set to 0.1 empirically. The specific settings of t_{AD} and t_{CL} will be discussed in Section 4.4.

4.3 Main Results

Analysis on CIFAR-LT

The comparison results on CIFAR-100/10-LT with imbalance factors of 100, 50, and 10 are presented in Table 1. All methods are trained using ResNet-32 [He *et al.*, 2016] to ensure a fair comparison. We show the results when the number of experts is 2 and 3. Overall, ICL achieves the highest accuracy under all imbalance ratios on the CIFAR100-LT and CIFAR10-LT datasets, outperforming existing methods. Specifically, ICL sets new benchmarks of 57.6%, 61.3%, and 69.3% for the imbalance rates of CIFAR100-LT. This result significantly outperforms all classes of methods. Notably, at imbalance factors of 100 and 50, the SHIKE and NCL++ boost compared to the currently most effective MoE methods are 1.3% and 1.5%, respectively. Similarly, on CIFAR10-LT, ICL maintains excellent performance with 87.9%, 89.7%, and 91.9%. This suggests that ICL's ability to address class imbalance in CIFAR-10 is particularly effective at maintaining high accuracy under more extreme class distribution imbalance.

Method	Img-LT	iNat-LT
Single Model		
Cross Entropy	41.6	66.9
WD [Alshammari *et al.*, 2022]	53.3	70.0
BCL [Zhu *et al.*, 2022]	56.0	–
SBCL [Hou *et al.*, 2023]	57.1	70.8
GCL [Li *et al.*, 2022c]	54.9	–
GLMC [Du *et al.*, 2023]	56.3	–
H2T [Li *et al.*, 2024]	56.9	72.0
DBN-Mix [Baik *et al.*, 2024]	56.6	74.7
MoE-based Method		
RIDE [Wang *et al.*, 2021]	55.4	71.7
ACE [Cai *et al.*, 2021]	55.1	72.9
TLC [Li *et al.*, 2022a]	55.1	–
ResLT [Cui *et al.*, 2023]	55.1	72.9
NCL [Li *et al.*, 2022b]	59.5	74.9
SHIKE [Jin *et al.*, 2023]	59.7	75.4
NCL++ [Tan *et al.*, 2024]	59.6	75.2
Ours(2 experts)	59.5	75.3
Ours(3 experts)	**60.2**	**75.9**

Table 2: Comparison on Img-LT and iNat-LT.

Method	iNat-LT		
	Many	Med.	Few
Single Model			
Cross Entropy	**76.1**	69.0	62.4
TSC [Li *et al.*, 2022d]	70.6	67.8	67.8
SBCL [Hou *et al.*, 2023]	73.3	71.9	68.6
WB [Alshammari *et al.*, 2022]	71.0	70.3	69.4
DBN-Mix [Baik *et al.*, 2024]	73.0	75.6	74.7
MoE-based Method			
RIDE [Wang *et al.*, 2021]	68.3	72.6	71.8
ResLT [Cui *et al.*, 2023]	73.0	72.6	73.1
NCL [Li *et al.*, 2022b]	72.7	75.6	74.5
NCL++ [Tan *et al.*, 2024]	72.2	75.3	75.7
Ours(2 experts)	73.1	75.5	75.6
Ours(3 experts)	74.9	**75.9**	**76.1**

Table 3: Performance on iNat-LT in many, medium, and few classes.

MoE	CoL	DOL	ACC
			50.34
✓			54.42
✓	✓		56.50
✓		✓	56.56
✓	✓	✓	57.59

Table 4: Influences of key modules.

Analysis on Img-LT and iNat-LT

We report the overall Top-1 accuracy of Img-LT and iNat-LT in Table 2. ICL achieves a performance of 60.2 and 75.9, respectively, outperforming all competing methods, especially the state-of-the-art MoE-based methods SHIKE and NCL++. We further report the accuracy of the three divisions on the iNat dataset, i.e., the many-shot class (>100 training samples), the medium-shot class (20 $\sim$ 100 training samples), and the few-shot class (<20 training samples), in Table 3. ICL performed well in all category divisions, achieving the best results in both the medium-shot (75.9%) and few-shot (76.1%) categories, and ranked second in the many-shot (74.9%) categories. This result demonstrates ICL's superior performance on large-scale real-world long-tail datasets.

4.4 Ablation Study and Further Analysis

This section presents extensive ablation studies to analyze the effectiveness of each component in our proposed ICL framework. All experiments are conducted on the CIFAR-100-LT datasets with an imbalance factor of 100 if not stated.

Influence of Key Modules in ICL

We report the results of the ICL module ablation experiments in Table 4. Since the CoL is an extension of cross-entropy, we use each expert's sum of cross-entropy when the CoL is not used to ensure the basic learning objective. It can be seen that all of our proposed modules improve the overall performance of the model. When applying the MoE architecture, the accuracy is improved by 4.12% compared to the 50.34% of the single-expert model, validating the effectiveness of expert integration for long-tail learning. Based on the MoE architecture, CoL and DOL are improved by 2.08% and 2.14%, respectively. Ultimately, after combining CoL and DOL, the model achieves a maximum accuracy of 57.59%, a 3.17% improvement over the baseline MoE model. The steady performance improvement demonstrates the effectiveness of the proposed ICL.

Analysis of CoL

To show the effect of competition and collaboration learning on inter-expert contributions, we show gradient contribution experiments for models trained using CoL in Figure 3(a), where the baseline results are displayed with transparent lines for easy comparison. Compared to the baseline, where expert 3 always dominates the training process, the experts trained using CoL have a clear competitive relationship. Specifically, expert 2 and expert 3 have comparable contributions and alternately dominate the training process. Eventually, the three experts of CE achieved individual accuracy rates of 45.23%, 44.68%, and 45.36%, with a systematic accuracy rate of 47.56%.In comparison, the three experts of CoL achieved individual accuracy rates of 46.62%, 48.02%, and 47.33%, with a systematic accuracy rate of 48.28%. Its accuracy rate matches the gradient contribution ranking, which not only indicates that the gradient contribution experiment can correctly reflect the individual optimality of the experts in training, but also proves that CoL can maintain the individual optimality of the experts through the competition, and promote the model to perform a better overall effect. Furthermore, Figure 3(b) illustrates the KL divergence between the predicted probabilities of each expert's outputs for CoL

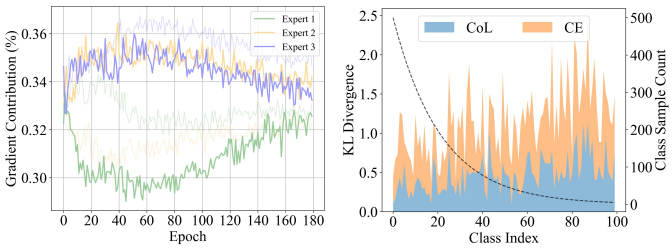


Figure 3: Visualization of competition and collaboration between experts. (a) Gradient contribution of each expert during training. (b) Sum of KL divergence among the output logits of experts.

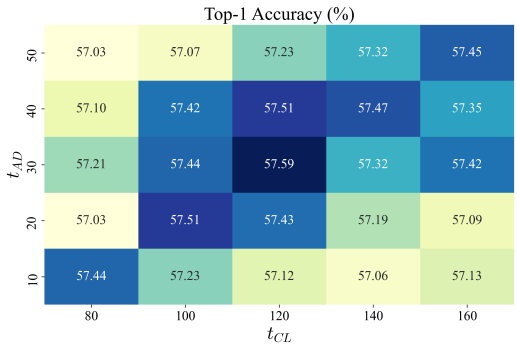


Figure 4: Parametric analysis of t_{AD} and t_{CL}.

and baseline across various categories. The KL divergence signifies the degree of uncertainty among experts regarding the inputs. CoL demonstrates a substantially smaller KL divergence than CE in all categories, particularly in the tail categories with limited samples. This finding suggests that CoL's mutual distillation effectively promotes expert knowledge sharing and reduces system uncertainty.

Analysis of DOL

In our early experiments, we found that DOL was able to achieve a 56.52% accuracy rate if its two losses were used consistently throughout the training process. This result does not utilize the potential of DOL. We hypothesized that enhancing the diversity of the model at the onset of the training phase would yield a more diverse feature extraction driven by CoL. Similarly, since the prototype construction of DOL relies on a more rational representation space, CCL losses should be introduced only after several training epochs. We therefore performed a combination of these two parameters t_{AD} and t_{CL}, to obtain the results shown in Figure 4. The model achieved the best results at 57.59% at t_{AD} and t_{CL} of 30 and 120, respectively, which is an improvement of 1.07% over the baseline 56.56%, proving our analysis. We further observed that the elements on the subdiagonal of the matrix are typically the largest within their respective rows. This is because, due to the effect of CoL, experts may tend to approximate, and thus the later the intervention of CCL, the stronger the tendency for expert homogeneity. Therefore, a longer diversity loss duration is needed to maintain expert diversity and support CCL in promoting domain-specific expertise.

CCL aims to improve the representation space, enhancing individual experts' performance. A visual comparison of the baseline CE and CCL methods is made using t-SNE [Maaten and Hinton, 2008]. As shown in Figure 5, the t-SNE plot for CCL exhibits better clustering, with more compact feature distribution and clearer class separation. This indicates that CCL outperforms CE in creating discriminative feature representations. Additionally, the effect of AD loss is demonstrated by visualizing the activation maps of three MoE experts using Grad-CAM [Selvaraju *et al.*, 2019]. Figure 6 highlights, as an example of a picture in Img-LT, the differences in expert attention across image regions, showing that AD loss enhances expert diversity, rather than just the positional relationships in the representation space.

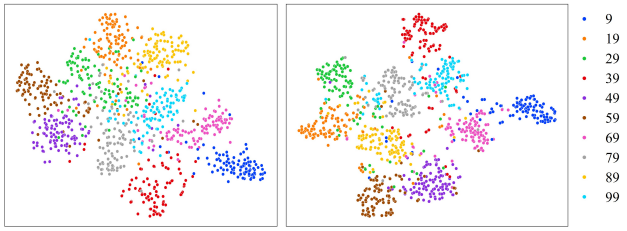
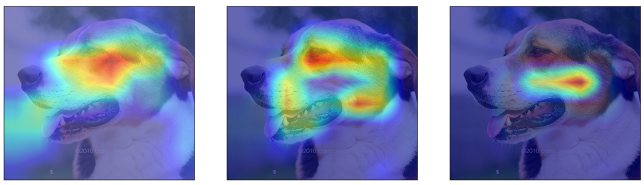


Figure 5: Visualization of t-SNE with (a) CE and (b) CCL on CIFAR100-LT with an imbalance factor of 100. We have chosen ten classes equally spaced for better view.



(a) Expert 1 (b) Expert 2 (c) Expert 3

Figure 6: Grad-CAM visualization.

5 Conclusion

In this work, we introduced the Independent and Collaborative Learning (ICL) framework, which enables the independent optimization of each expert while preserving their strengths to achieve globally optimal performance collaboratively. By incorporating Diverse Optimization Learning (DOL), we enhanced the diversity among experts and strengthened their representation space, thereby reducing uncertainty. Additionally, we propose the Competition and Collaboration Learning (CoL) approach, which conceptualizes experts as branches of parallel circuits. This framework enables the dynamic interaction between competition and collaboration, where experts simultaneously strive for individual optimization while sharing knowledge to enhance collective performance. Our approach demonstrates a promising path toward balancing autonomy and collaboration in multi-agent systems, paving the way for more effective and efficient learning paradigms.

Acknowledgements

This work is supported by the Solfeggio ear training intelligent robot and cloud platform research and development project for music education (No.2024CXY0102) and the 3D visualization digital twin integrated control system (No.2023CXY0111), and in part by the National Natural Science Foundation of China (No.62471418).

Contribution Statement

The first two authors are equal contribution.

References

[Alshammari *et al.*, 2022] Shaden Alshammari, Yu-Xiong Wang, Deva Ramanan, and Shu Kong. Long-tailed recognition via weight balancing. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6887–6897, 2022.

[Baik *et al.*, 2024] Jae Soon Baik, In Young Yoon, and Jun Won Choi. Dbn-mix: Training dual branch network using bilateral mixup augmentation for long-tailed visual recognition. *Pattern Recognition*, 147:110107, 2024.

[Cai *et al.*, 2021] Jiarui Cai, Yizhou Wang, and Jenq-Neng Hwang. Ace: Ally complementary experts for solving long-tailed recognition in one-shot. In *IEEE/CVF International Conference on Computer Vision*, 2021.

[Chen *et al.*, 2020] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 1597–1607, 2020.

[Chen *et al.*, 2022] Xiaohua Chen, Yucan Zhou, Dayan Wu, Wanqian Zhang, Yu Zhou, Bo Li, and Weiping Wang. Imagine by reasoning: A reasoning-based implicit semantic data augmentation for long-tailed classification. In *36th AAAI Conference on Artificial Intelligence, 34th Conference on Innovative Applications of Artificial Intelligence, The 12th Symposium on Educational Advances in Artificial Intelligence*, pages 356–364, 2022.

[Cubuk *et al.*, 2019] Ekin D. Cubuk, Barret Zoph, Dandelion Mane, Vijay Vasudevan, and Quoc V. Le. Autoaugment: Learning augmentation strategies from data. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019.

[Cubuk *et al.*, 2020] Ekin D. Cubuk, Barret Zoph, Jonathon Shlens, and Quoc V. Le. Randaugment: Practical automated data augmentation with a reduced search space. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2020.

[Cui *et al.*, 2019] Yin Cui, Menglin Jia, Tsung-Yi Lin, Yang Song, and Serge J. Belongie. Class-balanced loss based on effective number of samples. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 9268–9277, 2019.

[Cui *et al.*, 2021] Jiequan Cui, Zhisheng Zhong, Shu Liu, Bei Yu, and Jiaya Jia. Parametric contrastive learning. In *IEEE/CVF International Conference on Computer Vision*, pages 695–704, 2021.

[Cui *et al.*, 2023] Jiequan Cui, Shu Liu, Zhuotao Tian, Zhisheng Zhong, and Jiaya Jia. Reslt: Residual learning for long-tailed recognition. *IEEE Trans. Pattern Anal. Mach. Intell.*, 45(3):3695–3706, 2023.

[Devries and Taylor, 2017] Terrance Devries and Graham W. Taylor. Improved regularization of convolutional neural networks with cutout. *CoRR*, abs/1708.04552, 2017.

[Du *et al.*, 2023] Fei Du, Peng Yang, Qi Jia, Fengtao Nan, Xiaoting Chen, and Yun Yang. Global and local mixture consistency cumulative learning for long-tailed visual recognitions. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15814–15823, 2023.

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

[Hinton *et al.*, 2015] Geoffrey E. Hinton, Oriol Vinyals, and Jeffrey Dean. Distilling the knowledge in a neural network. *CoRR*, abs/1503.02531, 2015.

[Horn *et al.*, 2018] Grant Van Horn, Oisin Mac Aodha, Yang Song, Yin Cui, Chen Sun, Alexander Shepard, Hartwig Adam, Pietro Perona, and Serge J. Belongie. The inaturalist species classification and detection dataset. In *2018 IEEE Conference on Computer Vision and Pattern Recognition*, pages 8769–8778, 2018.

[Hou *et al.*, 2023] Chengkai Hou, Jieyu Zhang, Haonan Wang, and Tianyi Zhou. Subclass-balancing contrastive learning for long-tailed recognition. In *IEEE/CVF International Conference on Computer Vision*, pages 5372–5384, 2023.

[Jin *et al.*, 2023] Yan Jin, Mengke Li, Yang Lu, Yiu-ming Cheung, and Hanzi Wang. Long-tailed visual recognition via self-heterogeneous integration with knowledge excavation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23695–23704, 2023.

[Kang *et al.*, 2020] Bingyi Kang, Saining Xie, Marcus Rohrbach, Zhicheng Yan, Albert Gordo, Jiashi Feng, and Yannis Kalantidis. Decoupling representation and classifier for long-tailed recognition. In *8th International Conference on Learning Representations*, 2020.

[Li *et al.*, 2022a] Bolian Li, Zongbo Han, Haining Li, Huazhu Fu, and Changqing Zhang. Trustworthy long-tailed classification. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.

[Li *et al.*, 2022b] Jun Li, Zichang Tan, Jun Wan, Zhen Lei, and Guodong Guo. Nested Collaborative Learning for Long-Tailed Visual Recognition . In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6939–6948, 2022.

[Li *et al.*, 2022c] Mengke Li, Yiu-Ming Cheung, and Yang Lu. Long-tailed visual recognition via gaussian clouded

logit adjustment. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6919–6928, 2022.

[Li *et al.*, 2022d] Tianhong Li, Peng Cao, Yuan Yuan, Lijie Fan, Yuzhe Yang, Rogerio Feris, Piotr Indyk, and Dina Katabi. Targeted supervised contrastive learning for long-tailed recognition. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6908–6918, 2022.

[Li *et al.*, 2023] Mengke Li, Yiu-Ming Cheung, and Zhikai Hu. Key point sensitive loss for long-tailed visual recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(4):4812–4825, 2023.

[Li *et al.*, 2024] Mengke Li, Zhikai Hu, Yang Lu, Weichao Lan, Yiu-ming Cheung, and Hui Huang. Feature fusion from head to tail for long-tailed visual recognition. In *38th AAAI Conference on Artificial Intelligence, 36th Conference on Innovative Applications of Artificial Intelligence, 14th Symposium on Educational Advances in Artificial Intelligence*, pages 13581–13589, 2024.

[Lin *et al.*, 2014] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: common objects in context. In *13th European Conference Computer Vision*, volume 8693, pages 740–755, 2014.

[Lin *et al.*, 2017] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollar. Focal loss for dense object detection. In *2017 IEEE International Conference on Computer Vision*, 2017.

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

[Maaten and Hinton, 2008] Laurensvander Maaten and GeoffreyE. Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research,Journal of Machine Learning Research*, 2008.

[Menon *et al.*, 2021] Aditya Krishna Menon, Sadeep Jayasumana, Ankit Singh Rawat, Himanshu Jain, Andreas Veit, and Sanjiv Kumar. Long-tail learning via logit adjustment. In *9th International Conference on Learning Representations*, 2021.

[Park *et al.*, 2021] Seulki Park, Jongin Lim, Younghan Jeon, and Jin Young Choi. Influence-balanced loss for imbalanced visual classification. In *IEEE/CVF International Conference on Computer Vision*, pages 715–724, 2021.

[Ren *et al.*, 2020] Jiawei Ren, Cunjun Yu, Shunan Sheng, Xiao Ma, Haiyu Zhao, Shuai Yi, and Hongsheng Li. Balanced meta-softmax for long-tailed visual recognition. In *Advances in Neural Information Processing Systems*, 2020.

[Selvaraju *et al.*, 2019] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. *International Journal of Computer Vision*, 128(2):336–359, 2019.

[Snell *et al.*, 2017] Jake Snell, Kevin Swersky, and Richard S. Zemel. Prototypical networks for few-shot learning. In *Advances in Neural Information Processing Systems*, pages 4077–4087, 2017.

[Tan *et al.*, 2024] Zichang Tan, Jun Li, Jinhao Du, Jun Wan, Zhen Lei, and Guodong Guo. Ncl++: Nested collaborative learning for long-tailed visual recognition. *Pattern Recognition*, 147:110064, 2024.

[Verma *et al.*, 2019] Vikas Verma, Alex Lamb, Christopher Beckham, Amir Najafi, Ioannis Mitliagkas, David Lopez-Paz, and Yoshua Bengio. Manifold mixup: Better representations by interpolating hidden states. In *Proceedings of the 36th International Conference on Machine Learning*, pages 6438–6447, 2019.

[Wang *et al.*, 2020] Tao Wang, Yu Li, Bingyi Kang, Junnan Li, Jun Hao Liew, Sheng Tang, Steven C. H. Hoi, and Jiashi Feng. The devil is in classification: A simple framework for long-tail instance segmentation. In *16th European Conference Computer Vision*, volume 12359, pages 728–744, 2020.

[Wang *et al.*, 2021] Xudong Wang, Long Lian, Zhongqi Miao, Ziwei Liu, and Stella X. Yu. Long-tailed recognition by routing diverse distribution-aware experts. In *9th International Conference on Learning Representations*, 2021.

[Zang *et al.*, 2021] Yuhang Zang, Chen Huang, and Chen Change Loy. Fasa: Feature augmentation and sampling adaptation for long-tailed instance segmentation. In *IEEE/CVF International Conference on Computer Vision*, pages 3437–3446, 2021.

[Zhang *et al.*, 2018] Hongyi Zhang, Moustapha Cissé, Yann N. Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. In *6th International Conference on Learning Representations*, 2018.

[Zhang *et al.*, 2021] Linfeng Zhang, Chenglong Bao, and Kaisheng Ma. Self-distillation: Towards efficient and compact neural networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, page 1–1, 2021.

[Zhou *et al.*, 2018] Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10 million image database for scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, page 1452–1464, 2018.

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

[Zhu *et al.*, 2022] Jianggang Zhu, Zheng Wang, Jingjing Chen, Yi-Ping Phoebe Chen, and Yu-Gang Jiang. Balanced contrastive learning for long-tailed visual recognition. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6898–6907, 2022.