VICAL: Vicinal Consistency Alignment for Long-Tailed Visual Recognition

Jianggang Zhu[1,2], Zheng Wang[3], Bin Zhu[4], Yi-Ping Phoebe Chen[5], and Jingjing Chen*[1,2]

[1] Institute of Trustworthy Embodied AI, Fudan University, Shanghai, China
[2] Shanghai Key Laboratory of Multimodal Embodied AI, Shanghai, China
[3] Zhejiang University of Technology, Hangzhou, China
[4] Singapore Management University, Singapore, Singapore
[5] La Trobe University, Melbourne, Australia

Abstract. Multi-expert models have become the dominant paradigm for long-tailed learning, largely attributed to their presumed ability to benefit from expert diversity. However, we revisit this central assumption and reveal that diversity induced by logit adjustment or explicit regularizers does not guarantee better ensemble accuracy. Our work suggests that multi-expert models benefit more from variance reduction than diversity maximization. We introduce **VICAL**, a **VI**cinal **C**onsistency **AL**ignment framework that improves long-tailed recognition not by enforcing expert diversity, but by reducing prediction variance. Specifically, our approach comprises two key components: Self-Consistency Learning and Deep Ensemble Distillation. Self-Consistency Learning discourages reliance on unstable high-frequency information, smoothing the local loss landscape and mitigating overfitting, especially for tail classes. Deep Ensemble Distillation promotes cross-expert low-frequency semantic agreement using a low-resolution view, thereby sidestepping optimization conflicts with established knowledge. Extensive experiments on CIFAR-LT, ImageNet-LT, and iNaturalist 2018 show that VICAL consistently outperforms state-of-the-art methods, validating the effectiveness of our consistency-driven design.

Keywords: Long-Tailed Visual Recognition · Multi-Expert Models · Consistency Learning

1 Introduction

Computer vision has witnessed remarkable progress over the last decade, largely driven by the construction of large-scale and high-quality datasets such as ImageNet [26] and MS COCO [18]. However, many of these benchmarks implicitly assume a balanced label distribution, which does not reflect real-world data distributions. In practice, data typically follow a long-tailed distribution, where a

* Corresponding Author. Email: {jgzhu20, chenjingjing}@fudan.edu.cn

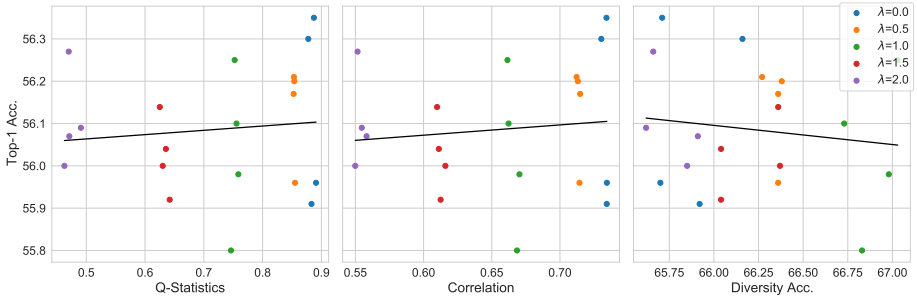


Fig. 1: We use three different diversity metrics, *i.e.*, Q-statistics, correlation, and diversity accuracy, to measure the relationship between model diversity and ensemble accuracy on CIFAR-100-LT. Each point denotes an independently trained model, and λ controls the differences in logit adjustment strength across experts. A larger λ indicates more diverse experts by controlling the logit adjustment intensity for each expert. Varying λ changes expert agreement but produces no clear improvement in ensemble accuracy. (The Pearson correlation coefficients are 0.11, 0.11, and -0.13 for Q-statistics, correlation, and diversity accuracy.)

few head classes dominate the data volume while the majority of tail classes suffer from extreme data scarcity. Deep neural networks trained on such data exhibit a pronounced bias toward the head classes [15, 44], limiting applicability in domains such as medical diagnostics [13] and autonomous driving [24].

To address the imbalance, numerous approaches have been proposed [3, 7, 15, 21, 44, 45]. Among them, logit adjustment [21] is an effective approach to adjust the classification boundary and achieve Fisher consistency. More recently, multi-expert models [1, 31, 39, 41] have received significant attention due to their superior performance. This success is largely attributed to their ability to simultaneously mitigate both model variance and bias [31], thereby reducing prediction uncertainty [33]. Some works [1, 39, 41] also hypothesize that this success derives from model diversity, which maximizes expert specialization by forcing each expert to focus on different data distributions. These approaches have achieved strong empirical performance, advancing long-tailed visual recognition.

However, the fundamental premise of these diversity-driven methods, *the relationship between explicit expert diversity enforcement and ensemble prediction accuracy*, remains systematically unverified. In this paper, we re-evaluate this core assumption using multiple ensemble diversity metrics and surprisingly reveal that expert diversity induced by logit adjustment or explicit loss regularizers does not universally yield better ensemble accuracy. More formally, the expected test error for ensemble models can be approximately decomposed as [32]

$$Error = Noise + Bias + Variance - Diversity \tag{1}$$

Eq. 1 states diversity is an entangled hidden dimension in the bias-variance decomposition of ensemble learning rather than an independent variable [32]. Current expert diversity-enforcing methods [31, 41] in long-tailed learning may

overlook this crucial coupling. As shown in Fig. 1, we can obtain diverse outputs with lower Q-statistics and correlation coefficients by controlling the logit adjustment intensities for each expert, but without a notable improvement in ensemble accuracy. We provide results on diversity induced by explicit regularizers [31] in the Supplementary Material, which show that explicitly forcing diversity degrades individual expert representation quality. These observations prompt us to reconsider the underlying factors that truly determine the effectiveness of multi-expert models.

Guided by this insight, we propose the Vicinal Consistency Alignment framework, VICAL, for *variance reduction* rather than *explicit expert diversity constraints*. Our approach is grounded in a key insight: a robust model should exhibit minimal prediction variation against both localized high-frequency perturbations and global resolution shifts, especially for tail classes. Specifically, VICAL comprises two complementary components operating across decoupled frequency domains: Self-Consistency (SC) Learning and Deep Ensemble Distillation (DED). First, SC enforces intra-expert prediction invariance by constructing a localized vicinity through the interpolation of strong augmented views. By aligning predictions between the original views and these heavily corrupted interpolations, SC penalizes the exploitation of unstable high-frequency patterns, effectively smoothing the local loss landscape of tail classes. Second, DED promotes inter-expert consistency by utilizing a low-resolution view to extract resolution-invariant semantics. Because this downsampled input inherently lacks high-frequency details, DED strictly confines the cross-expert alignment to low-frequency semantic agreement. This decoupling averts optimization conflicts, preserving the robust, individual features established by SC while filtering conflicting knowledge. Together, these components significantly reduce prediction variance both within and across experts.

Our main contributions can be summarized as follows:

- Our work first empirically re-evaluates the hypothesis that expert diversity induced by logit adjustment or explicit regularizers directly correlates with ensemble accuracy in long-tailed recognition. Our study reveals that utilizing logit adjustment or explicit regularizers to maximize diversity can compromise individual expert accuracy, thereby undermining the overall ensemble.
- We propose Vicinal Consistency Alignment (VICAL), a novel consistency learning framework that moves beyond explicit diversity constraints to implicit variance reduction.
- We implement VICAL through two core and complementary components: Self-Consistency Learning, which enables individual experts to discourage reliance on unstable high-frequency patterns, and Deep Ensemble Distillation, which facilitates cross-expert alignment through low-frequency semantic agreement.
- VICAL showcases its superiority and generalizability through extensive experiments on popular long-tailed benchmarks, including CIFAR-LT, ImageNet-LT, and iNaturalist 2018, achieving significant gains over existing state-of-the-art methods.

2 Related Work

Long-Tailed Visual Recognition Conventional strategies for mitigating class imbalance primarily include re-sampling, re-weighting, and logit adjustment. Re-sampling [3, 15] techniques typically over-sample instances of tail classes or under-sample head classes, while re-weighting [3, 8] typically assigns higher loss weights to tail classes. Logit adjustment [21, 25] further refines this idea by explicitly modifying decision boundaries to accommodate class imbalance. On the other hand, a large body of contrastive learning methods [7, 9, 45] has been adopted to enhance representation quality by learning a balanced feature space.
Multi-Expert Models Recently, multi-expert models [1, 17, 28, 31, 39, 41] have become the dominant paradigm in long-tailed recognition. These methods typically train multiple experts independently or cooperatively, and aggregate predictions of all experts during testing. A common hypothesis underpinning these approaches is that encouraging expert diversity can improve ensemble performance by allowing each expert to specialize in different categories. For example, RIDE [31] enforces diversity by penalizing the similarity of expert predictions using the explicit diversity regularizer. SADE [39], MDCS [41], and BalPoE [1] use varying logit adjustment intensities across experts, compelling each expert to specialize in distinct categories. However, the validity of this foundational premise has rarely been scrutinized. Our work initiates from this critical gap.
Consistency Learning Consistency learning is a powerful regularization technique in semi-supervised learning [27, 29, 41], typically leveraging unlabeled data by enforcing output stability under input perturbations. These perturbations are commonly generated through varying data augmentation strategies, such as the strong-weak pair employed in FixMatch [27]. Beyond FixMatch, Mean Teacher [29] also uses generic consistency to propagate pseudo-labels to unlabeled data. Consistency learning has also been shown to improve model robustness and representation learning [12]. Mixup [37] creates vicinal samples by mixing instances from different classes, yielding more robust decision boundaries. GLMC [10] utilizes Mixup and CutMix [36] to achieve global and local mixture consistency for long-tailed learning. In contrast, VICAL adopts frequency-decoupled consistency, which acts as a spectral bottleneck to suppress high-frequency overfitting and enforce a robust, low-frequency consensus. This design reduces variance within and across experts in long-tailed recognition.

3 Do More Diverse Experts Yield Better Performance?

A line of recent methods [1, 2, 20, 39, 41, 43] has been proposed to improve the output diversity of multi-expert models by enabling each expert to focus on different categories. However, one premise for these practices to be effective is that diverse experts are beneficial for long-tailed learning, and this simple assumption has not been verified. In this section, we first provide preliminaries for long-tailed recognition and then evaluate whether improving the expert diversity through logit adjustment and an explicit regularizer in existing long-tailed learning approaches enhances recognition performance.

3.1 Preliminaries

The goal of long-tailed visual recognition is to train on skewed data while achieving promising performance on the balanced test set. Formally, let $\mathbb{D} = \{x_i, y_i\}_{i=1}^{N}$ be the training set where x_i denotes the i-th image sample and y_i is the corresponding label. For an ensemble model with M experts, given the input x_i, the prediction probability of j-th class in m-th expert parameterized with θ_m is

$$p_j(x_i;\theta_m) = \frac{exp(z_{ij}^m)}{\sum_{k=1}^{C} exp(z_{ik}^m)}, \tag{2}$$

where z_{ij}^m is the logit for class j of sample i, and C is the number of classes. The posterior in Eq. 2 does not account for the train-test class prior shift. After calibration by logit adjustment [21, 25], Eq. 2 can be reformulated as

$$p_j(x_i;\theta_m) = \frac{exp(z_{ij}^m + \tau_m \log(P_j))}{\sum_{k=1}^{C} exp(z_{ik}^m + \tau_m \log(P_k))}, \tag{3}$$

$$\mathcal{L}_{ce}^m = -\log(p_j(x_i;\theta_m)). \tag{4}$$

Here, P_j is the class prior of class j, the hyperparameter λ controls calibration strength, and $\mathcal{L}_{ce}^m$ denotes the basic classification loss of individual experts. A large value of τ_m (>1) will render the prediction biased towards the tail class and a small value (<1) towards the head class. Previous works [1, 20, 39, 41] assign small and large λs for head and tail experts, focusing on head and tail classes, respectively. Following the common setup [1,39,41], we set $M = 3$ in subsequent sections unless otherwise mentioned.

3.2 Statistical Analysis

In our work, we find that expert diversity induced by logit adjustment (See Fig. 1) or an explicit loss regularizer (See Supplementary Material) has no positive correlation with ensemble performance. Specifically, we train multi-expert models on CIFAR-100-LT (IF=100) under varying logit adjustment intensities: $(\tau_1, \tau_2, \tau_3) = (1-\lambda,\ 1,\ 1+\lambda)$ ($\lambda \in \{0, 0.5, 1, 1.5, 2\}$) for the long-tailed, uniform, and reversed experts, respectively. We measure diversity using Q-statistics [35] and the correlation coefficient ρ, with the diversity factor σ [41] as an auxiliary accuracy metric. As shown in Fig. 1, larger λ values yield lower Q-statistics and ρ, indicating that more diverse predictions are obtained. However, this induced diversity fails to yield ensemble accuracy gains. Furthermore, when applying explicit regularizers (See Supplementary Material), we observe a strong positive correlation between Q-statistics and ensemble accuracy. This indicates that artificially pushing experts to be diverse harms overall performance, motivating our shift in focus from pursuing diversity to reducing model variance.

4 Vicinal Consistency Alignment

In this paper, we emphasize that reducing the variance of the multi-expert model is more effective than enforcing diversity in Eq. 1. To this end, we present a complementary and tiered consistency-learning framework, VICAL, to achieve vicinal consistency alignment, with two key components: Self-Consistency (SC) Learning and Deep Ensemble Distillation (DED). Specifically, we construct a vicinity for each instance using an interpolation of two strongly augmented full-resolution views, together with a separate low-resolution view. The vicinity is strictly confined to transformations of the same input. First, SC acts as the local smoother and stabilizes the individual expert's prediction within the vicinity of each sample. Second, DED aligns the individual output of the low-resolution view with the global semantic consensus based on SC. Notably, by separating these objectives across different input resolutions, VICAL successfully reduces model variance both within and across experts while avoiding knowledge conflicts. In the following section, we will describe our proposed VICAL in greater detail.

4.1 Self-Consistency Learning

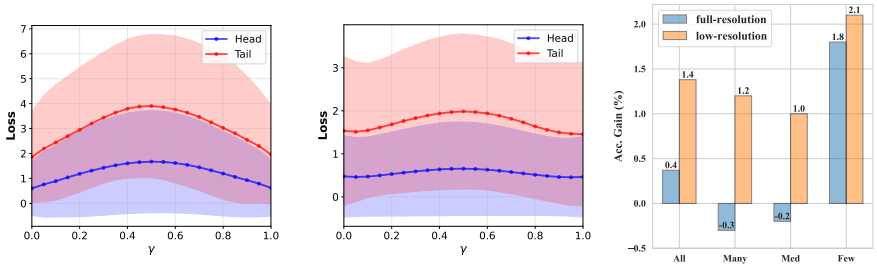
Vanilla Self-Distillation Self-distillation (SD) is an elegant and efficient approach to reducing model variance and improving prediction consistency [11, 29, 38]. Here, we consider an exponential moving average (EMA) model as the teacher to produce soft labels, which are more informative than hard labels. For simplicity, let $p^{\mathcal{S}}$ and $p^{\mathcal{T}}$ denote the predicted probability of the student and teacher model, respectively. The vanilla SD [29] of the i-th sample in the m-th expert can be expressed as:

$$\mathcal{L}_{sd} = KL(p^{\mathcal{T}}(x_i; \hat{\theta}_m)||p^{\mathcal{S}}(x_i; \theta_m)), \tag{5}$$

where $KL(\cdot||\cdot)$ represents the Kullback–Leibler divergence and $\hat{\theta}_m$ denotes the EMA model of m-th expert, updated from the online model θ_m. We omit the parameters θ_m and $\hat{\theta}_m$ in the following sections for simplicity.

Inconsistent Loss Landscape We find vanilla SD inadequate for enforcing robustness within the vicinity of each training sample. Specifically, we construct an interpolated view $\widetilde{\mathrm{v}} = \gamma \mathrm{v}_1 + (1-\gamma)\mathrm{v}_2$ within the vicinity, where v_1 and v_2 are two strong augmentations of the same image x, and $\gamma \sim Beta(\alpha, \alpha)$, with $\gamma \in [0, 1]$. Since v_1 and v_2 are semantically correlated, the loss of the convex combination $\widetilde{\mathrm{v}}$ for the tail class is expected to be as smooth as that of the head class when varying with γ. However, as shown in Fig. 2a, the vanilla SD exhibits high sensitivity to γ for the tail class, suggesting a sharp loss landscape and substantial prediction uncertainty.

Local and High-frequency Perturbations To mitigate sharp loss variations within the vicinity, we propose a self-consistency (SC) objective as the local smoother that penalizes prediction divergence across interpolated views. Specifically, we employ two strong augmented views v_1 and v_2, and introduce unstable



(a) (b) (c)

Fig. 2: Loss landscapes as a function of the interpolation factor $\gamma \in [0,1]$ between two augmented views v_1 and v_2 on CIFAR-100-LT (IF=100) with a single expert: (a) **vanilla self-distillation**, (b) **SC**. SC induces a smoother loss curve as γ varies, especially for the tail classes. (c) Accuracy gains over the SC baseline when DED uses a full-resolution or low-resolution input.

high-frequency perturbation via the interpolated view $\widetilde{v}$, which creates unnatural edges, ghosts, and textural artifacts. By feeding this heavily corrupted view $\widetilde{v}$ into the online student network, we compel the model to align its prediction with the EMA teacher of the original views. Due to data scarcity, tail classes are highly sensitive to vicinal perturbations, thereby receiving stronger regularization during consistency alignment. Consequently, SC penalizes the exploitation of unstable high-frequency patterns and achieves prediction consistency within the vicinity for each expert. By suppressing these brittle features, SC alleviates overfitting in the tail classes and yields flatter local minima (See Fig. 2b).

As illustrated in Fig. 3, the interpolated view $\widetilde{v}$ is fed into the online network to produce the prediction $p^{\mathcal{S}}(z^m_{\widetilde{v}})$ of the m-th expert as the student distribution. The views v_1 and v_2 are fed into the target network to calculate the mean logits $\overline{z}^m$ to serve as the stable teacher. Formally, the modified objective function of the m-th expert can be expressed as

$$\mathcal{L}^m_{sd} = KL(p^{\mathcal{T}}(\overline{z}^m)||p^{\mathcal{S}}(z^m_{\widetilde{v}})), \tag{6}$$

where $\overline{z}^m = \frac{1}{2}(z^m_{v_1} + z^m_{v_2})$ is the mean of EMA logits for v_1 and v_2 and serves as the stabilized target.

4.2 Deep Ensemble Distillation

Global and Low-frequency Alignment Ensemble distillation offers a promising approach for variance reduction [31]. While SC improves robustness to high-frequency perturbations introduced by interpolation, unconditionally forcing the student to align with the global ensemble consensus poses a severe risk of optimization conflict with established knowledge of SC. Concretely, if an extra full-resolution view were employed for distillation, ensemble averaging inherently blurs individual features established by SC, leading to degraded performance

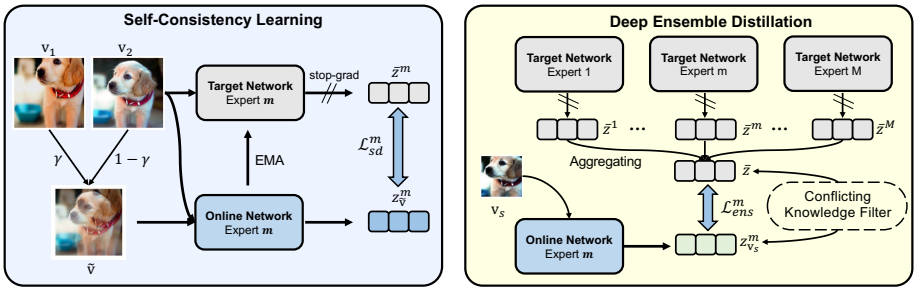


Fig. 3: The framework of our proposed VICAL, consisting of Self-Consistency (SC) Learning and Deep Ensemble Distillation (DED). SC utilizes the interpolated view $\widetilde{\mathrm{v}}$ to discourage reliance on unstable high-frequency patterns, significantly improving robustness of tail classes. DED strictly confines the cross-expert alignment to low-frequency semantic agreement via a low-resolution view v_s, avoiding optimization conflicts with the SC objective. DED further filters conflicting transfers while retaining complementary expert knowledge through the Conflicting Knowledge Filter.

compared to using SC alone (See Fig. 2c). To sidestep this conflict, our Deep Ensemble Distillation (DED) module restricts the distillation objective strictly to a low-resolution view v_s. Because the downsampled v_s lacks high-frequency information, this resolution asymmetry confines the ensemble alignment strictly to low-frequency semantics. Consequently, the DED module offers complementary signals based on SC, and our design decouples the SC and DED objectives across different frequency domains.

As shown in Fig. 3, the low-resolution view v_s is fed into the online network and produces logits $z^m_{\mathrm{v}_s}$ of the m-th expert. The outputs of all target models are averaged to obtain $\overline{z}$, which provides the global semantic consensus. The logits $z^m_{\mathrm{v}_s}$ are forced to align with the mean $\overline{z}$ of full-resolution views, which enhances low-frequency consistency and improves overall robustness.

Conflicting Knowledge Filter To further ensure that each expert gains complementary knowledge, we introduce the Conflicting Knowledge Filter (CKF) that sidesteps incorrect knowledge transfer, preventing individual knowledge from degenerating into uncertain global consensus. The final ensemble distillation formulation can be expressed as

$$\mathcal{L}^m_{ens} = \frac{1}{|\mathbb{D}^C_c|} \sum_{\mathrm{v}_s \in \mathbb{D}^C_c} KL(p^{\mathcal{T}}(\overline{z}) || p^{\mathcal{S}}(z^m_{\mathrm{v}_s})), \tag{7}$$

$$\mathbb{D}_c = \{x | \mathrm{argmax}(p^{\mathcal{T}}(\overline{z})) \neq y \ \textit{and} \ \mathrm{argmax}(p^{\mathcal{S}}(z^m_{\mathrm{v}_s})) = y\} \tag{8}$$

where $\overline{z}$ denotes the aggregated logits across experts, and $\mathbb{D}^C_c$ denotes the complement of $\mathbb{D}_c$, *i.e.*, the non-conflict knowledge set, excluding samples that the student correctly predicts but the teacher fails. The CKF maintains individual experts' specialized knowledge while enforcing semantic-level agreement.

Method	CIFAR-100-LT				CIFAR-10-LT				ImageNet-LT
	200	100	50	10	200	100	50	10	256
LDAM-DRW [3]	38.5	42.0	46.6	58.7	74.7	77.0	81.0	88.2	45.8
BCL‡ [45]	-	51.9	56.6	64.9	-	84.3	87.2	91.1	57.1
PaCo‡ [7]	47.8	52.0	56.0	64.2	82.3	85.4	88.0	91.5	57.2
ProCo‡ [9]	-	52.8	57.1	65.5	-	85.9	88.2	91.9	58.0
RIDE [31]	44.6	48.0	51.7	61.8	77.8	81.2	83.7	86.3	56.8
SADE [39]	44.7	49.8	53.9	63.6	78.0	82.9	85.8	90.0	58.8
NCL‡ [17]	49.5	54.2	58.2	63.8	82.2	85.5	87.3	91.1	60.5
BalPoE‡ [1]	-	55.9	60.1	68.1	-	86.8	88.5	91.9	61.6
MDCS‡ [41]	-	56.1	60.1	-	-	87.2	88.3	-	60.2
ECL‡ [33]	51.4	56.3	59.9	67.3	83.6	86.5	88.9	91.8	61.7
PRL [43]	-	52.8	57.3	65.6	-	-	-	-	60.7
NCL++‡ [28]	-	56.3	59.8	-	-	87.2	88.8	-	60.9
ICL‡ [4]	-	57.6	61.3	69.3	-	87.9	89.7	91.9	60.2
Ours‡	**55.3**	**59.7**	**63.9**	**71.6**	**86.7**	**90.0**	**91.9**	**94.5**	**62.9**

Table 1: Top-1 accuracy on CIFAR-LT using ResNet-32 as backbone and on ImageNet-LT using ResNeXt-50 as backbone. We report the results of 400 and 200 epochs for CIFAR-LT and ImageNet-LT, respectively. ‡ denotes models trained with strong augmentation [7]. The best result is shown in bold, and the second-best result is underlined.

4.3 Training and Inference

Training The overall training objective comprises three components: cross-entropy loss $\mathcal{L}_{ce}^m$ of original views v_1 and v_2, self-consistency distillation $\mathcal{L}_{sd}^m$ of $\widetilde{\mathrm{v}}$, and ensemble distillation loss of $\mathcal{L}_{ens}^m$ of v_s. Finally, the loss is as follows:

$$\mathcal{L} = \sum_{m=1}^{M} \left(\mathcal{L}_{ce}^m + \eta \mathcal{L}_{ens}^m + \beta \mathcal{L}_{sd}^m \right), \tag{9}$$

where η and β are two hyperparameters to control the distillation strength, respectively. Note that each expert is trained with a uniform-aware objective to produce class-balanced predictions.

Inference The online network's exposure to the interpolated samples degrades the precision of its Batch Normalization statistics. Consequently, we only keep the target network during the inference phase to produce high-quality predictions, and the online network is discarded.

5 Experiments

To validate the efficacy of our proposed VICAL, we conduct comprehensive experiments on long-tailed CIFAR-10 [3], long-tailed CIFAR-100 [3], ImageNet-LT [19], and iNaturalist 2018 [30]. Details of these datasets and extended experiments are available in the Supplementary Material.

Method	Many	Med	Few	All
100 epochs				
RIDE [31]	70.9	72.4	73.1	72.6
ACE [2]	-	-	-	72.9
BalPoE [1]	73.2	75.5	74.7	75.0
BalPoE‡ [1]	-	-	-	73.5
MDCS‡ [41]	71.8	73.1	72.4	72.5
Ours‡	**74.3**	**77.3**	**76.2**	**76.6**
200 epochs				
TS-MOF [42]	72.8	73.6	70.5	72.3
ResLT [6]	73.0	72.6	73.1	72.9
ML [22]	-	-	-	74.9
NCL++‡ [28]	72.2	75.3	75.7	75.2
SHIKE‡ [14]	-	-	-	75.4
ICL‡ [4]	74.9	75.9	76.1	75.9
Ours‡	**75.5**	**78.1**	**77.3**	**77.5**

Table 2: Top-1 accuracy on iNaturalist 2018 using ResNet-50 as backbone. We report the results of 100 epochs and 200 epochs, respectively. ‡ denotes models trained with RandAugment [5].

ME	SC Eq. 6	DED Eq. 7	CKF Eq. 8	Top-1 Acc.
✓				55.6
✓	✓			57.8
✓	✓	✓		58.6
✓	✓	✓	✓	59.1

Table 3: Ablation study on CIFAR-100-LT training for 250 epochs. ME: multi-expert model. SC: Self-Consistency Learning. DED: Deep Ensemble Distillation. CKF: Conflicting Knowledge Filter.

Methods	Top-1 Acc.
full-resolution	58.1
+ high-pass	58.1
+ low-pass	58.6
low-resolution	59.1

Table 4: Ablation study of applying spatial frequency filters (high-pass vs. low-pass) to the full-resolution input images for DED on CIFAR-100-LT.

5.1 Experimental Settings

Evaluation Protocol We follow the standard evaluation protocol [3, 7, 44, 45] and validate the top-1 accuracy for all categories on the corresponding balanced test set. Following common practice [15], we split the classes into three groups, *i.e.*, many-shot (>100 images), medium-shot (20∼100 images), and few-shot (<20 images), and report the respective accuracy of each group.

Implementation Details We implement our approach with PyTorch [23] and employ the SGD optimizer with a momentum of 0.9 by default. The models are trained for 400, 200, and 100/200 epochs, and the batch size is set to 64, 256, and 512 for CIFAR-LT, ImageNet-LT, and iNaturalist 2018, respectively. For network architecture and data augmentation strategy, we mainly follow [1, 41]. The initial learning rate is 0.2 for iNaturalist 2018 and 0.1 for the other datasets. The learning rate follows a cosine annealing schedule for ImageNet-LT and iNaturalist 2018, and is multiplied by 0.1 at epochs 320 and 360 for CIFAR-LT. We set α to 1.0, η to 0.8, and β to 1.0 for all experiments. The momentum coefficient of the target network is set to 0.99 by default. The resolution of the low-resolution view is set to 16×16 for CIFAR-LT and 96×96 for the large-scale datasets. More details can be found in the Supplementary Material.

Methods	Multi-Expert	Intra-Expert Stabilization	Inter-Expert Alignment	Objective Decoupling	Distillation Gate
PCL [16]	×	Predictive Consistency	×	×	×
GLMC [10]	×	Inter-Class Mixup	×	×	×
LTRL [40]	×	Reflective Learning	×	×	×
RIDE [31]	✓	×	Explicit Divergence	×	×
BalPoE [1]	✓	Inter-class Mixup	Explicit Divergence	×	×
MDCS [41]	✓	Vanilla Self-Distillation	Explicit Divergence	×	Confident Sampling
NCL++ [28]	✓	Vanilla Self-Distillation	Full-Resolution Mimicry	×	× (Unconditional)
ICL [4]	✓	×	Full-Resolution Mimicry	×	× (Unconditional)
VICAL (Ours)	✓	**Vicinal Interpolation**	**Low-Resolution Semantic**	**Resolution Asymmetry**	**CKF**

Table 5: Mechanism comparisons of VICAL against existing consistency-driven and multi-expert methods. Unlike prior approaches that rely on explicit divergence or coupled full-resolution distillation, VICAL decouples intra-expert local smoothing from inter-expert global semantic agreement, which avoids optimization conflicts.

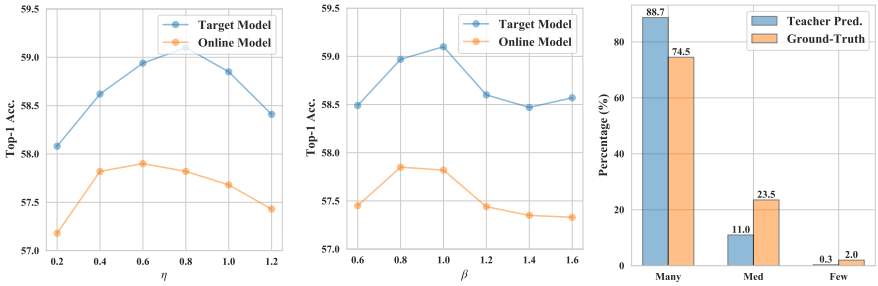


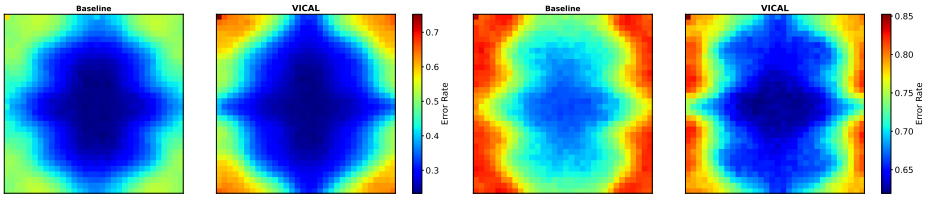
Fig. 4: Ablation study of the hyperparameters (a) η and (b) β on CIFAR-100-LT (IF=100). We report the best results achieved by the online and target models. (c) By the terminal training phase, the histogram of the conflicting knowledge set $\mathbb{D}_c$.

5.2 Main Results

Comparison with state-of-the-art We compare VICAL with previous state-of-the-art methods such as NCL++ [28], MDCS [41] and BalPoE [1]. BalPoE serves as a representative baseline that integrates Mixup during training. Tab. 1 lists the results on CIFAR-LT and ImageNet-LT, and Tab. 2 shows the comparison on iNaturalist 2018. Our VICAL framework consistently achieves state-of-the-art performance across all datasets. Tab. 5 lists the core mechanism differences of VICAL against existing consistency-driven and multi-expert approaches. Compared to NCL++ [28] and MDCS [41], which employ full-resolution views for distillation, VICAL achieves substantial improvement, validating the efficacy of our vicinal consistency alignment.

Component Analysis Tab. 3 presents the ablation study of all components of our VICAL on CIFAR-100-LT. The baseline achieves 55.6% top-1 accuracy. Each module provides a substantial improvement, indicating that the components are complementary: SC aligns predictions within each expert, while DED achieves low-frequency semantic agreement across experts.

Effect of Hyperparameters Fig. 4a and Fig. 4b present the ablation studies of hyperparameters η and β. The target model achieves the best accuracy at



(a) **(b)**

Fig. 5: Model sensitivity to additive noise aligned with different Fourier basis vectors on the CIFAR-100-LT validation set: (a) **Head** and (b) **Tail**. The center represents **low-frequency**, while the edges represent **high-frequency**. For head classes, VICAL is slightly more sensitive to extreme high-frequency noise. In contrast, VICAL substantially improves tail class robustness in the low-frequency region.

$p^{\mathcal{T}}$	$p^{\mathcal{S}}$	Ratio(%)	Many	Med	Few	All
-	-	-	74.3	59.1	36.6	57.7
All	All	100	74.8	59.8	38.4	58.6
Correct	-	73	74.6	59.8	37.9	58.4
Wrong	Wrong	25	73.6	59.3	38.4	58.0
Wrong	Correct	2	67.7	55.1	34.2	53.2
Ours	Ours	98	75.5	60.1	38.7	59.1

Table 6: We compare the results of several Conflicting Knowledge Filters on the CIFAR-100-LT (IF=100). The baseline is VICAL w/o DED. The Ratio denotes the percentage of samples used for DED throughout the training.

$\eta = 0.8$ while the online network peaks at $\eta = 0.6$. Similarly, the optimal value of β for the target model is 1. The superior performance of the target model is attributed to its clean inputs. In contrast, exposure to corrupted samples introduces noise into the online network and degrades its Batch Normalization statistics. Consequently, we use the target network during inference.

5.3 Further Analysis

Frequency-based Analysis Fig. 2c shows the results when using views at different resolutions for DED. Tab. 4 further shows the results of applying different filters, which confirms that discarding unstable high-frequency patterns effectively mitigates cross-expert knowledge conflicts. To validate our frequency-decoupled mechanism, Fig. 5 visualizes the model's error rate on the CIFAR-100-LT validation set against Fourier basis noise, following [34]. VICAL significantly expands the low-frequency robustness region for tail classes, while increasing sensitivity to extreme high-frequency noise for head classes. This asymmetric shift demonstrates that VICAL restricts the exploitation of unstable high-frequency shortcuts, compelling the model to establish robust dependencies on low-frequency structures.

Method	All			Many			Med			Few		
	Acc.	Bias	Var	Acc.	Bias	Var	Acc.	Bias	Var	Acc.	Bias	Var
CE	31.6	0.60	0.47	57.3	0.28	0.35	28.2	0.61	0.51	6.3	0.94	0.57
RIDE [31]	40.5	0.50	0.42	60.5	0.28	0.30	38.7	0.50	0.44	20.1	0.74	0.52
MDCS [41]	46.1	-	0.36	-	-	0.24	-	-	0.38	-	-	0.46
MDCS†	48.2	0.43	0.37	68.9	0.22	0.24	55.1	0.34	0.34	30.4	0.63	0.48
λ=2	49.9	0.42	0.35	**74.6**	**0.16**	**0.19**	57.0	0.33	0.31	29.3	0.65	0.48
λ=0	49.9	0.42	0.35	74.4	0.17	0.20	56.8	0.32	0.32	29.5	0.65	0.48
+ Mixup [37]	49.7	0.41	0.36	69.4	0.22	0.23	56.1	0.34	0.32	32.7	**0.59**	0.47
+ AugMix [12]	49.3	0.42	0.36	71.7	0.19	0.22	55.7	0.33	0.32	30.7	0.62	0.48
Ours	**52.3**	**0.39**	**0.33**	73.6	0.18	**0.19**	**59.3**	**0.31**	**0.29**	**33.9**	**0.59**	**0.45**

Table 7: Comparisons of the mean accuracy, bias, and variance of baselines and our proposed method on CIFAR-100-LT (IF=100), following the previous setup [31, 41]. † represents our reproduced results using the official code on the specified datasets for a fair comparison. $\lambda = 0$ refers to our proposed method without SC and DED.

Identifying Conflicting Knowledge We provide Tab. 6 to identify which teacher-student prediction pairs create conflicts. Specifically, we divide the samples into several groups based on whether the teachers' and students' predictions are correct. Although samples where the teacher is incorrect and the student is correct account for only 2%, they can cause a significant performance drop and must be removed. In Fig. 4c, we provide the histogram of the conflicting knowledge set $\mathbb{D}_c$. During the terminal training phase, for misclassified samples from difficult head classes, the ensemble consensus tends to predict them as another head class. CKF safeguards valuable individual expert knowledge from being compromised by uncertain consensus.

Heatmaps of Entropy Profiles To validate the efficacy of SC in improving consistency and reducing prediction uncertainty, we present heatmaps to show how the entropy of interpolated samples on the test dataset varies with respect to γ in Fig. 6. Specifically, we train three ResNet-32 networks independently using Balanced Softmax [25] (BS) loss, BS jointly with vanilla SD, and BS jointly with our SC module on CIFAR-100-LT. The entropy of tail samples on the test set is typically higher than that of head samples. SC significantly reduces the uncertainty and induces a smoother entropy profile, especially for tail classes.

Model Bias and Variance We aim to enhance model robustness and reduce variance through the proposed VICAL. To validate the effectiveness, we follow the prior setup [31, 41] and train 20 independent models on subsets of CIFAR-100-LT. Tab. 7 shows the comparison results. Increasing λ enhances output diversity as shown in Fig. 1, yet yields no corresponding improvement in model bias or variance. Mixup [37] and AugMix [12] enhance model robustness by interpolating inputs across different classes and combining multiple random perturbations into a single input, respectively. Although Mixup and AugMix effectively mitigate bias in tail classes, this advantage comes at the cost of significant per-

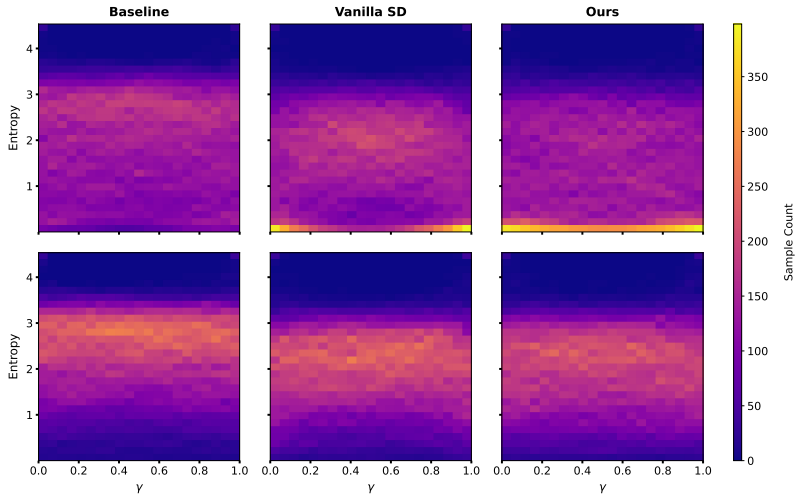


Fig. 6: Heatmaps of the entropy profiles as a function of the interpolation factor $\gamma \in [0,1]$ between two full-resolution views v_1 and v_2 on the CIFAR-100-LT test set with a single expert. Head samples (**Top**) are more susceptible to γ, whereas tail samples (**Bottom**) have higher entropy overall. SC reduces the entropy of interpolated samples, particularly for tail classes.

formance degradation in the head. In contrast, VICAL not only reduces variance and bias on tail classes but also preserves the head's performance, demonstrating its superior regularization effect on long-tailed learning.

6 Conclusion

In this paper, we critically re-evaluate the prevailing hypothesis in long-tailed learning that assumes a positive correlation between model diversity and ensemble accuracy. We unveil that diversity induced by logit adjustment or explicit regularizers does not universally guarantee better ensemble accuracy. Motivated by these findings, we introduce Vicinal Consistency Alignment (VICAL), a novel framework that shifts the focus from explicit diversity constraints to implicit variance reduction. By penalizing reliance on unstable high-frequency patterns and enforcing low-frequency semantic agreement, VICAL improves the stability within each expert and sidesteps optimization conflicts through views at asymmetric resolutions. Extensive experiments show that VICAL consistently achieves state-of-the-art performance across benchmarks. More importantly, it effectively reduces both variance and bias, particularly for tail classes. Our work demonstrates that vicinal consistency alignment offers a powerful alternative for improving generalization in long-tailed recognition.

Acknowledgements

This project was supported by the National Natural Science Foundation of China (NSFC) Projects under Grants No. 62522206 and No. 62521004.

References

1. Aimar, E.S., Jonnarth, A., Felsberg, M., Kuhlmann, M.: Balanced product of calibrated experts for long-tailed recognition. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 19967–19977 (2023)
2. Cai, J., Wang, Y., Hwang, J.N.: Ace: Ally complementary experts for solving long-tailed recognition in one-shot. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 112–121 (2021)
3. Cao, K., Wei, C., Gaidon, A., Arechiga, N., Ma, T.: Learning imbalanced datasets with label-distribution-aware margin loss. Advances in neural information processing systems **32** (2019)
4. Chen, Y., Jian, Z., Ke, N., Hu, S., Jiao, J., Hong, Q., Wu, Q.: Enhancing mixture of experts with independent and collaborative learning for long-tail visual recognition. In: Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence. pp. 828–836 (2025)
5. Cubuk, E.D., Zoph, B., Shlens, J., Le, Q.V.: Randaugment: Practical automated data augmentation with a reduced search space. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops. pp. 702–703 (2020)
6. Cui, J., Liu, S., Tian, Z., Zhong, Z., Jia, J.: Reslt: Residual learning for long-tailed recognition. IEEE transactions on pattern analysis and machine intelligence **45**(3), 3695–3706 (2022)
7. Cui, J., Zhong, Z., Liu, S., Yu, B., Jia, J.: Parametric contrastive learning. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 715–724 (2021)
8. Cui, Y., Jia, M., Lin, T.Y., Song, Y., Belongie, S.: Class-balanced loss based on effective number of samples. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 9268–9277 (2019)
9. Du, C., Wang, Y., Song, S., Huang, G.: Probabilistic contrastive learning for long-tailed visual recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence **46**(9), 5890–5904 (2024)
10. Du, F., Yang, P., Jia, Q., Nan, F., Chen, X., Yang, Y.: Global and local mixture consistency cumulative learning for long-tailed visual recognitions. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 15814–15823 (2023)
11. Furlanello, T., Lipton, Z., Tschannen, M., Itti, L., Anandkumar, A.: Born again neural networks. In: International conference on machine learning. pp. 1607–1616. PMLR (2018)
12. Hendrycks, D., Mu, N., Cubuk, E.D., Zoph, B., Gilmer, J., Lakshminarayanan, B.: Augmix: A simple data processing method to improve robustness and uncertainty. arXiv preprint arXiv:1912.02781 (2019)
13. Holste, G., Zhou, Y., Wang, S., Jaiswal, A., Lin, M., Zhuge, S., Yang, Y., Kim, D., Nguyen-Mau, T.H., Tran, M.T., et al.: Towards long-tailed, multi-label disease classification from chest x-ray: Overview of the cxr-lt challenge. Medical Image Analysis **97**, 103224 (2024)

14. Jin, Y., Li, M., Lu, Y., Cheung, Y.m., Wang, H.: Long-tailed visual recognition via self-heterogeneous integration with knowledge excavation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 23695–23704 (2023)
15. Kang, B., Xie, S., Rohrbach, M., Yan, Z., Gordo, A., Feng, J., Kalantidis, Y.: Decoupling representation and classifier for long-tailed recognition. In: Eighth International Conference on Learning Representations (ICLR) (2020)
16. Kang, N., Chang, H., MA, B., Bai, S., Shan, S., Chen, X.: Predictive consistency learning for long-tailed recognition. In: 34th British Machine Vision Conference 2023, BMVC 2023, Aberdeen, UK, November 20-24, 2023. BMVA (2023)
17. Li, J., Tan, Z., Wan, J., Lei, Z., Guo, G.: Nested collaborative learning for long-tailed visual recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6949–6958 (2022)
18. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: European conference on computer vision. pp. 740–755. Springer (2014)
19. Liu, Z., Miao, Z., Zhan, X., Wang, J., Gong, B., Yu, S.X.: Large-scale long-tailed recognition in an open world. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 2537–2546 (2019)
20. Ma, C., Elezi, I., Deng, J., Dong, W., Xu, C.: Three heads are better than one: Complementary experts for long-tailed semi-supervised learning. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 38, pp. 14229–14237 (2024)
21. Menon, A.K., Jayasumana, S., Rawat, A.S., Jain, H., Veit, A., Kumar, S.: Long-tail learning via logit adjustment. arXiv preprint arXiv:2007.07314 (2020)
22. Park, C., Yim, J., Jun, E.: Mutual learning for long-tailed recognition. In: Proceedings of the IEEE/CVF winter conference on applications of computer vision. pp. 2675–2684 (2023)
23. Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al.: Pytorch: An imperative style, high-performance deep learning library. Advances in neural information processing systems **32** (2019)
24. Peri, N., Dave, A., Ramanan, D., Kong, S.: Towards long-tailed 3d detection. In: Conference on Robot Learning. pp. 1904–1915. PMLR (2023)
25. Ren, J., Yu, C., Ma, X., Zhao, H., Yi, S., et al.: Balanced meta-softmax for long-tailed visual recognition. Advances in neural information processing systems **33**, 4175–4186 (2020)
26. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A.C., Fei-Fei, L.: ImageNet Large Scale Visual Recognition Challenge. International Journal of Computer Vision (IJCV) **115**(3), 211–252 (2015)
27. Sohn, K., Berthelot, D., Carlini, N., Zhang, Z., Zhang, H., Raffel, C.A., Cubuk, E.D., Kurakin, A., Li, C.L.: Fixmatch: Simplifying semi-supervised learning with consistency and confidence. Advances in neural information processing systems **33**, 596–608 (2020)
28. Tan, Z., Li, J., Du, J., Wan, J., Lei, Z., Guo, G.: Ncl++: Nested collaborative learning for long-tailed visual recognition. Pattern Recognition **147**, 110064 (2024)
29. Tarvainen, A., Valpola, H.: Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. Advances in neural information processing systems **30** (2017)

30. Van Horn, G., Mac Aodha, O., Song, Y., Cui, Y., Sun, C., Shepard, A., Adam, H., Perona, P., Belongie, S.: The inaturalist species classification and detection dataset. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 8769–8778 (2018)
31. Wang, X., Lian, L., Miao, Z., Liu, Z., Yu, S.X.: Long-tailed recognition by routing diverse distribution-aware experts. arXiv preprint arXiv:2010.01809 (2020)
32. Wood, D., Mu, T., Webb, A.M., Reeve, H.W., Luján, M., Brown, G.: A unified theory of diversity in ensemble learning. Journal of machine learning research **24**(359), 1–49 (2023)
33. Xu, Z., Chai, Z., Xu, C., Yuan, C., Yang, H.: Towards effective collaborative learning in long-tailed recognition. IEEE Transactions on Multimedia **26**, 3754–3764 (2023)
34. Yin, D., Gontijo Lopes, R., Shlens, J., Cubuk, E.D., Gilmer, J.: A fourier perspective on model robustness in computer vision. Advances in Neural Information Processing Systems **32** (2019)
35. Yule, G.U.: Vii. on the association of attributes in statistics: with illustrations from the material of the childhood society, &c. Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character **194**(252-261), 257–319 (1900)
36. Yun, S., Han, D., Oh, S.J., Chun, S., Choe, J., Yoo, Y.: Cutmix: Regularization strategy to train strong classifiers with localizable features. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 6023–6032 (2019)
37. Zhang, H., Cisse, M., Dauphin, Y.N., Lopez-Paz, D.: mixup: Beyond empirical risk minimization. arXiv preprint arXiv:1710.09412 (2017)
38. Zhang, L., Song, J., Gao, A., Chen, J., Bao, C., Ma, K.: Be your own teacher: Improve the performance of convolutional neural networks via self distillation. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 3713–3722 (2019)
39. Zhang, Y., Hooi, B., Hong, L., Feng, J.: Self-supervised aggregation of diverse experts for test-agnostic long-tailed recognition. Advances in Neural Information Processing Systems **35**, 34077–34090 (2022)
40. Zhao, Q., Dai, Y., Lin, S., Hu, W., Zhang, F., Liu, J.: Ltrl: Boosting long-tail recognition via reflective learning. In: European Conference on Computer Vision. pp. 1–18. Springer (2024)
41. Zhao, Q., Jiang, C., Hu, W., Zhang, F., Liu, J.: Mdcs: More diverse experts with consistency self-distillation for long-tailed recognition. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 11597–11608 (2023)
42. Zhao, Z., Gong, Z., Wang, P., Wen, H., Guo, C., Xue, B., Lin, X., Wang, Z., Zhang, Q., Wang, Y.: Ts-mof: Two-stage multi-objective fine-tuning for long-tailed recognition. In: Advances in Neural Information Processing Systems. vol. 38, pp. 5224–5252 (2025)
43. Zhao, Z., Wen, H., Wang, Z., Wang, P., Wang, F., Lai, S., Zhang, Q., Wang, Y.: Breaking long-tailed learning bottlenecks: A controllable paradigm with hypernetwork-generated diverse experts. Advances in Neural Information Processing Systems **37**, 7493–7520 (2024)
44. Zhou, B., Cui, Q., Wei, X.S., Chen, Z.M.: Bbn: Bilateral-branch network with cumulative learning for long-tailed visual recognition. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 9719–9728 (2020)

45. Zhu, J., Wang, Z., Chen, J., Chen, Y.P.P., Jiang, Y.G.: Balanced contrastive learning for long-tailed visual recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6908–6917 (2022)