# Feature Space Distillation for Bayesian Bias Consistency in Multi-Expert Long-Tailed Recognition

**Paper ID 4291**


## Abstract

The models supervised with long-tailed data usually exhibit class-wise *bias* and tail *uncertainty*, which means the majority class predictions *overwhelm* the minority ones and the logits of identical models *vary* signally for the *same* minority images, despite the *same* training manner and model architecture. Recently, Balanced Cross-entropy Loss (BCL) has shown superiority to remarkably eliminate the above statistical *bias*. However, we observe that BCL can be flawed in integrating with re-sampling or feature-mixing strategies in Multi-Expert (ME) frameworks, which alleviates the tail uncertainty effectively. In this paper, we pinpoint such contradiction lies in the essence that these strategies will *influence the label distribution either implicitly or explicitly*, which deteriorates the consistency of statistical prior. To tackle the dilemma, we propose homomorphic Feature Space Distillation (FSD) to aggregate the knowledge from experts in ME to handle the tail *uncertainty* while keeping the prior *bias* consistent. In FSD, feature-level distillation is conducted instead of classical logits-level to obtain more generalized features and avoid inaccurate prior *bias* from teacher classifiers. We theoretically prove that FSD remains the inherent statistical *bias* and is thus compatible with BCL and ensures *Fisher consistency*. Moreover, we integrate recent conspicuous advances to present a novel baseline for Long-Tailed Recognition (LTR). With the novel FSD, our proposal achieves state-of-the-art performance and alleviates the class-wise *bias* and tail *uncertainty* simultaneously in CIFAR10/100-LT, ImageNet-LT, and iNaturalist 2018. The detailed ablation study manifests the superiority of the collaboration of FSD and BCL.


## Introduction

Recent progress in computer vision, e.g., visual recognition (Girshick et al. 2014; Huang et al. 2017; Ronneberger, Fischer, and Brox 2015), video analysis (Mao et al. 2018; Wu et al. 2020; Kwon et al. 2020; Zhang et al. 2021a) and motion capture (Huang et al. 2022; Nakano et al. 2020; Xu et al. 2022; Garau et al. 2021; Zeng et al. 2020), heavily relies on the large-scale, high-quality, and class-balanced datasets, such as ImageNet (Russakovsky et al. 2015), COCO (Lin et al. 2014) and Place (Zhou et al. 2017), which require laborious collections and carefully annotations. Unfortunately, collecting rare instances tends to accompany more dominant samples because real-world data is naturally imbalanced distributed w.r.t. its categories. When the dataset is organized according to the sample number of each category, it typically exhibits a *long-tailed* distribution, i.e., only a few labels (**head**) occupy most samples, while most labels (**tail**) are associated with limited instances.

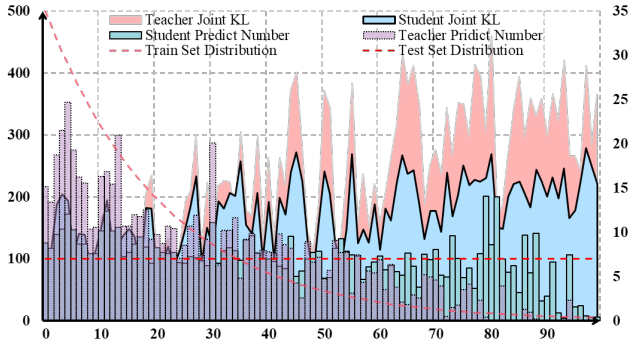



Figure 1: Main challenges of LTR. *Longtailness*: the teacher predictors tend to misclassify the input as the head while our final student expert makes a more balanced prediction. *Tail uncertainty*: The student expert shows much lower joint KL on the tail. $x$-axis: class index. $y$-axis: instance number (left) and distance (right). Joint KL: the sum KL distance among 3 corresponding experts initialized with different seeds.

We observe that the Long Tail Recognition (LTR) encounters the "longtailness" of data distribution and the "scarcity" of tail classes (c.f. Fig. 1). 1) *Longtailness* indicates that the head samples overwhelm the tail, where the models tend to *bias* to the head, i.e., the classifier simply regards an image as the head, which deteriorates the accuracy and robustness on the balanced test dataset. 2) *Scarcity* means that the models exhibit considerable *variance* on the tail because of its data paucity. The predicted probability for identical models initialized with different random seeds varies remarkably for the same tail input, even under the same training settings and architecture. The bias and uncertainty mentioned exacerbate the instability, immensely hindering the robustness of models supervised by long-tailed datasets.

Recently, Balanced Cross-entropy Loss (Ren et al. 2020) or Bayesian Compensated Loss (Menon et al. 2020; Hong et al. 2021; Xu et al. 2021) (BCL) borrows the *Bayesian Theory* to calibrate the gap in label distribution between train and test set. Concretely, it figures out the above statistical prior gap as Bayesian bias and compensates it by label frequency on standard *softmax* cross-entropy loss (CE). Multi-Expert (ME) (Xiang et al. 2020; Wang et al. 2020) frame-

work trains several experts to focus on different aspects of knowledge, such that reach a trade-off between head performance and tail uncertainty from scarcity.

However, few works address the bias and uncertainty simultaneously. In fact, BCL is incompatible or even counterproductive with some experts in ME if re-sampling (Cui et al. 2019; Zhou et al. 2020; Xiang et al. 2020) or feature-mixing strategies (Chou et al. 2020; Zhang et al. 2021c; Chu et al. 2020) are adopted. The essence is that *these techniques explicitly or implicitly influence the statistical prior*. Hence, the bias is inconsistent among different experts and it is hard to deduce the properly compensated bias for each expert.

To tackle the above challenges, we propose Feature Space Distillation (FSD) framework to aggregate each expert knowledge and keep the Bayesian bias consistent for all experts in ME. The key idea of FSD is to learn different feature-level knowledge from teacher experts to prevent the inaccurate bias learned by the classifier, and present feature-level distillation with BCL to alleviate the bias to head with lower tail variance. Specifically, the novel FSD distinguishes from previous ME in two aspects:

**For *teacher expert training*,** instead of adopting complicated sampling or feature mixing strategies that potentially influence the statistical prior, we train experts with class irrelevant augmentations for consistent Bayesian bias and robust features through simple yet effective CE. Each expert is initialized with different random seeds and shares the same network architecture. In this way, each expert will learn different tail knowledge to reduce the potential variance.

**For *expert knowledge aggregation*,** we catch the essence that the bias mainly originates from the classifier in LTR (Kang et al. 2019). Hence, we distil at the feature-level instead of classic logits-level (Hinton et al. 2015; Xiang et al. 2020; Li et al. 2022a) to reduce the tail uncertainty and avoid distorted bias (c.f. Fig. 2) from teacher experts. We conduct BCL on the student classifier to eliminate Bayesian bias from the label statistical prior and theoretically prove FSD ensures *Fish consistency* (Lin et al. 2004).

Finally, we integrate the latest conspicuous advances in long-tailed community and propose a new strong baseline named CE++. It boosts the baseline-based methods on par with "bells and whistles" methods (Cui et al. 2021; Li et al. 2022a). Extensive experiments in four benchmarks justify the superiority of the proposed FSD cooperated with BCL.

In summary, our contributions are follows:

1) We pinpoint the essence of the conflict between re-sampling or feature mixing and BCL, which originates from the disturbance of label distribution prior.
2) We propose CE++, a new baseline that integrates the latest techniques in LTR to facilitate fair comparisons and align with recent progress in the long-tailed community.
3) We propose FSD, a novel framework to boost BCL to overcome the tail uncertainty and ensures model Fisher consistency. The final proposal achieves state-of-the-art performance on four challenging datasets.
4) We will make the proposed CE++ and FSD framework publicly available for research purposes.

## Related Work

**Rebalance Learning** can be divided into *feature-wise* and *loss-wise* from implementation perspective.

To avoid damaging model generalization severely from simply over/under sampling the tail/head classes, some methods (Han et al. 2005; Buda et al. 2018; He, Garcia et al. 2009; Kang et al. 2019; Zhang et al. 2021c) enrich the feature combination of tail samples with head ones' help (Kim et al. 2020; Zhang et al. 2021c; Wang et al. 2021a) or increase the tail frequency implicitly (Chou et al. 2020; Xu et al. 2021; Park et al. 2022). The two-stage approaches (Cao et al. 2019; Kang et al. 2019; Zhong et al. 2021) decouple feature learning from downstream tasks to reduce the bias on the classifier. Recent literature attempts to train the model in self-supervise manners (Chen et al. 2020; He et al. 2020; Khosla et al. 2020) to bypass the influence of imbalanced label distribution. SSP (Yang, Xu et al. 2020) and HybirdSC (Wang et al. 2021b) have shown that self-supervise or semi-supervise training can boost the performance through larger train epochs and GPU memory. Recent state-of-the-art (Cui et al. 2021; Wang et al. 2021b; Li et al. 2022c; Zhu et al. 2022) introduces fixed or learn-able proxy (or class center) to overcome the performance degradation due to the absence of label supervision.

To mitigate the inherent statistical bias in LTR, researchers have designed meticulous loss to learn larger *margins* among different classes (Cao et al. 2019; Menon et al. 2020; Hong et al. 2021; Ren et al. 2020; Xu et al. 2021; Jamal et al. 2020; Li et al. 2022b) or assign various *weights* for different classes based on the label frequencies (Cui et al. 2019; Ye et al. 2020; Tan et al. 2020; Zhong et al. 2021; Alshammari et al. 2022). In particular, the BCL (Menon et al. 2020; Ren et al. 2020; Hong et al. 2021; Xu et al. 2021) has been widely adopted by state-of-the-art methods for its effectiveness and conciseness (Li et al. 2022a; Cui et al. 2021; Zhu et al. 2022; Zhang et al. 2021b). However, it is not always complimentary with the above feature-wise methods for the inconsistency of the prior statistical bias.

**Multi-expert Learning.** To tackle the tail uncertainty (Iscen et al. 2021; Wang et al. 2020), multi-expert framework is increasingly valued, which can be divided into two stages, i.e., *single expert training* and *experts aggregation* (Zhou et al. 2020; Cai et al. 2021; Xiang et al. 2020; Wang et al. 2020; Zhang et al. 2021b). BBN (Zhou et al. 2020) trains two experts with instance sampling and inverse sampling respectively and aggregates their knowledge in a cumulative weighting manner. LFME (Xiang et al. 2020) trains different experts with different instances groups and weights each expert recommended logits as the final output. RIDE (Wang et al. 2020) enlarges the KL divergence to train single expert and cascades all the experts via decision gates for inference. TADE (Zhang et al. 2021b) trains experts by BCL with different assumed statistical prior and weights each expert's output, which is obtained via contrastive learning. Another feasible expert aggregation manner is knowledge distillation (Hinton et al. 2015). DiVE (He et al. 2021) shows the effectiveness of distillation in the LTR. SSD (Li, Wang, and Wu 2021) trains expert backbone via self-supervise and classifier via balanced sampling. CBD (Iscen et al. 2021)

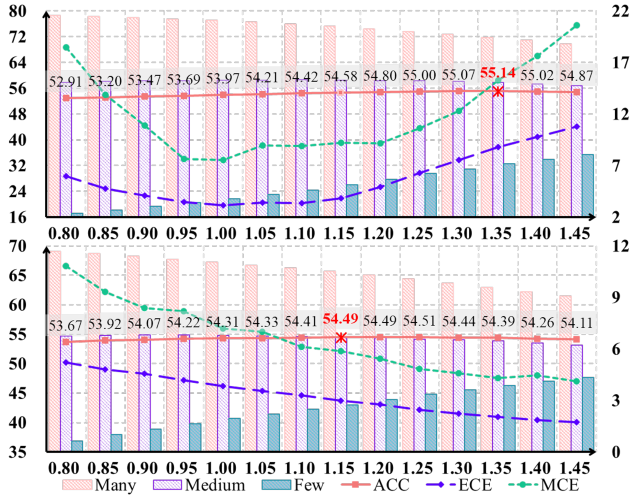



Figure 2: The post-hoc BCL on CIFAR-100-LT ($\gamma = 100$) (top) and ImageNet-LT (bottom). ECE and MCE describe model calibration (Ashukha et al. 2020; Xu et al. 2021). The best top-1 accuracy (red) is not achieved at the ideal $\tau = 1$, which means the bias that the classifier learned is inaccurate.

trains different teachers by different data augmentations and random seeds. Then, it trains student with balance sampling and CB Loss (Cui et al. 2019) through knowledge from the above teachers. NCL (Li et al. 2022a) trains experts in nested manner and adopts online self distillation with each other to reduce the tail uncertainty. However, these methods mainly make logits-level distillation, which may accumulate error from the inaccurate bias from teacher classifiers.

# Methodology

## Preliminaries

**Task Definition.** Given a $N$-sample dataset $\mathcal{D} = \{\mathcal{X}, \mathcal{Y}\}$ with $C$-class, each instance $\mathbf{x}_i \in \mathcal{X} := \{\mathbf{x}_1, \cdots, \mathbf{x}_n\}$ is labeled into $\mathcal{Y} := \{\mathbf{y}_1, \cdots, \mathbf{y}_n\}$, where $\mathbf{y}_i \in \mathcal{C} := \{1, \cdots, c\}$. We assume the dataset $\mathcal{D}$ is long-tailed distributed and denote each category as $\mathcal{C}_i$ and its instance number as $n_j = |\mathcal{C}_i|$. Furthermore, we consider a base classification model $\mathcal{M} := \{\mathscr{F}_\theta, \mathscr{W}_\phi\}$. It contains a *feature encoder* $\mathscr{F}_\theta$ and a *classifier* $\mathscr{W}_\phi$ with learnable parameters $\theta, \phi$ respectively. Given an input image $\mathbf{x}$, the encoder extracts the feature representation $\mathbf{v} := \mathscr{F}(\mathbf{x}|\theta) \in \mathbb{R}^d$. Then, the classifier (typically a fully connected layer) outputs the logits $\mathbf{z} := \mathscr{W}(\mathbf{v}|\phi) \in \mathbb{R}^C$. In this paper, we consider $k$-th teacher ($K$ experts in total) $\mathcal{M}_k^t := \{\mathscr{F}_{\theta_k}^t, \mathscr{W}_{\phi_k}^t\}$ and a student $\mathcal{M}^s := \{\mathscr{F}_\theta^s, \mathscr{W}_\phi^s\}$ for knowledge distillation, where all models adopt the same architecture $\mathcal{M}$.

**BCL & Statistical bias $\mathcal{B}$.** **B**ayesian-**C**ompensated **L**oss (BCL) is effective and widely adopted in LTR (Menon et al. 2020; Ren et al. 2020; Hong et al. 2021; Xu et al. 2021; Li et al. 2022a; Zhu et al. 2022). It compensates the statistical bias via logits adjustment on standard cross entropy loss. Consider the standard *softmax* operation:

$$p(\mathbf{y}_i|\mathbf{x};\theta,\phi) = \frac{e^{\mathcal{M}(\mathbf{x}|\theta,\phi)_{\mathbf{y}_i}}}{\sum_j e^{\mathcal{M}(\mathbf{x}|\theta,\phi)_{\mathbf{y}_j}}} \tag{1}$$

If the model $\mathcal{M}$ is supervised by cross entropy loss, we have:

$$\begin{aligned}\mathcal{L}_{\text{CE}}(\mathcal{M}(\mathbf{x}|\theta,\phi),\mathbf{y}_i) &= -\log(p(\mathbf{y}_i|\mathbf{x};\theta,\phi)) \\ &= \log[1 + \sum_{\mathbf{y}_j \neq \mathbf{y}_i} e^{\mathbf{z}_{\mathbf{y}_j} - \mathbf{z}_{\mathbf{y}_i}}],\end{aligned} \tag{2}$$

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

$$\begin{aligned}p_t(\mathbf{y}|\mathbf{x}) &= \frac{p_s(\mathbf{x}|\mathbf{y}) \cdot p_s(\mathbf{x})}{p_s(\mathbf{y})} \cdot \frac{p_t(\mathbf{y})}{p_t(\mathbf{x})} \propto \frac{p_s(\mathbf{x}|\mathbf{y}) \cdot p_t(\mathbf{y})}{p_s(\mathbf{y})} \\ &= \frac{\frac{p_t(\mathbf{y})}{p_s(\mathbf{y})} \cdot e^{\mathbf{z}_\mathbf{y}}}{\sum_{\mathbf{y}_j} \frac{p_t(\mathbf{y}_j)}{p_s(\mathbf{y}_j)} \cdot e^{\mathbf{z}_{\mathbf{y}_j}}} = \frac{e^{\mathbf{z}_\mathbf{y} + \log(p_t(\mathbf{y})) - \log(p_s(\mathbf{y}))}}{\sum_j e^{\mathbf{z}_{\mathbf{y}_j} + \log(p_t(\mathbf{y}_j)) - \log(p_s(\mathbf{y}_j))}},\end{aligned} \tag{3}$$

where $p_s(\mathbf{x})$ and $p_t(\mathbf{x})$ are regular terms (i.e., normal distribution). If the label in test set follows uniform distribution (i.e., $p_t(\mathbf{y}) \equiv \frac{1}{C}$), such bias can be compensated on standard CE. Combining Eq. 2 and Eq. 3, we have the BCL:

$$\mathcal{L}_{\text{BC}}(\mathcal{M}(\mathbf{x}|\theta,\phi),\mathbf{y}_i) = \log[1 + \sum_{\mathbf{y}_j \neq \mathbf{y}_i} e^{\mathcal{B}_{\mathbf{y}_j} - \mathcal{B}_{\mathbf{y}_i}} \cdot e^{\mathbf{z}_{\mathbf{y}_j} - \mathbf{z}_{\mathbf{y}_i}}], \tag{4}$$

where $\mathcal{B}_{\mathbf{y}_i} = p_s(\mathbf{y}_i) - p_t(\mathbf{y}_i)$ is statistical bias on class $\mathbf{y}_i$.

## Motivation

Although BCL is effective to mitigate the prior bias, it has two drawbacks worth considering. 1) *BCL fails to be collaborative with many feature-wise methods* commonly used in Multi-Expert framework. If balanced sampling is adopted, the input label prior $p_s(\mathbf{y})$ will be equal to $\frac{1}{C}$ and same to $p_t(\mathbf{y})$. In this case, $\mathcal{B}_{\mathbf{y}_i} = 0$ and thus BCL will be invalid. For feature-mixing strategies, the label distribution w.r.t category will be implicitly shifted. For example, CAM-sampling (Zhang et al. 2021c) creates a new sample via blending the foreground content of tail and the background of head. It is hard to deduce the label prior of the pseudo dataset concisely, which consists of mixed samples. 2) *The bias classifier learned may be inaccurate.* Consider implement BCL in a post-hoc processing way as a *softmax* variation (Eq. 5) (Menon et al. 2020; Hong et al. 2021; Xu et al. 2021). The best performance (see Fig. 2) achieves at $\tau \approx 1$ (0.9, 1.1 or others) instead of theoretical $\tau = 1$. Such an observation suggests that *the statistical bias learned by teacher classifier is unstable and inaccurate, which may be amplified and distorted further* if we do distillation on logits-level.

$$\begin{aligned}p(\mathbf{y}_i|\mathbf{x};\theta,\phi) &= \frac{e^{\mathbf{z}_{\mathbf{y}_i} - \tau\mathcal{B}_{\mathbf{y}_i}}}{\sum_{\mathbf{y}_j \neq \mathbf{y}_i} e^{\mathbf{z}_{\mathbf{y}_j} - \tau\mathcal{B}_{\mathbf{y}_j}}} \\ &= \frac{e^{\mathbf{z}_{\mathbf{y}_i} - \tau(p_s(\mathbf{y}_i) - p_t(\mathbf{y}_i))}}{\sum_{\mathbf{y}_j \neq \mathbf{y}_i} e^{\mathbf{z}_j - \tau(p_s(\mathbf{y}_j) - p_t(\mathbf{y}_j))}}\end{aligned} \tag{5}$$

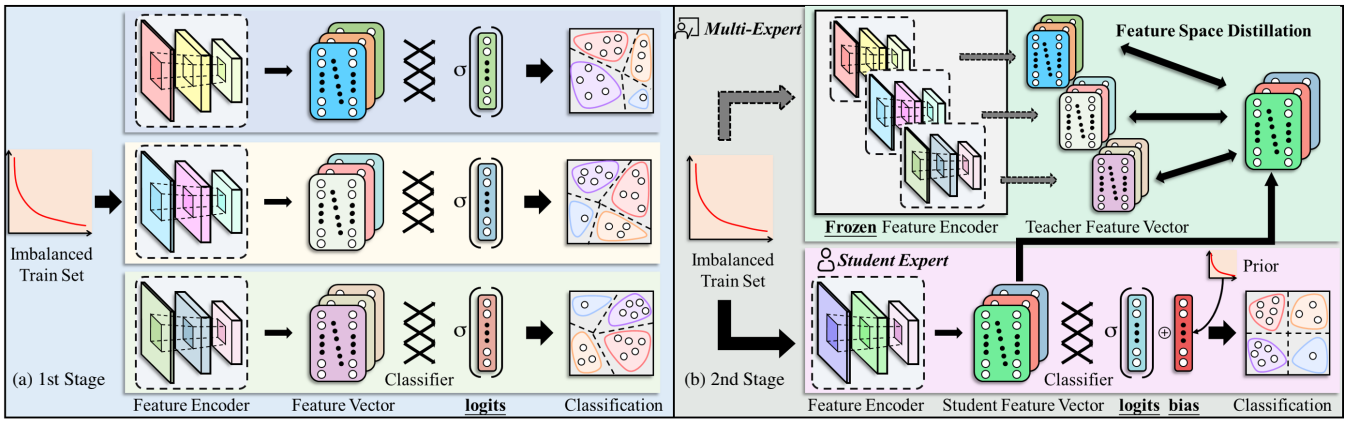



Figure 3: Overview of the proposed FSD. (a) *Teacher Expert Training*. Homomorphic experts are initialized with **different random seeds** and adopt class irrelevant augmentation to pretrain with CE Loss in parallel. (b) *Expert Knowledge Aggregation*. Student expert aggregates and distils teacher experts at **feature-level**, supervised by the Feature Distillation Loss and BCL. The overall FSD contributes to less bias and uncertainty simultaneously. $\sigma$ indicates *softmax* operation.

## Feature Space Distillation Framework

To take advantage of both BCL and ME, we propose feature space distillation based on the multi-expert framework, which can be divided into two parts as Fig. 3 illustrates.

**Teacher Expert training.** In the first stage, we train teacher experts with the goal that the teachers learn robust features without shifting its bias in Eq. 4. Hence, we train homomorphic experts with different random seeds and class irrelevant augmentation, e.g, Cutout (DeVries, Taylor et al. 2017). We adopt instance sampling to avoid disturbing the statistical bias explicitly. In this way, experts perform similarly on the head while varying significantly on the tail.

**Expert Knowledge Aggregation.** Previous works have shown that *the model bias that head overwhelms tail is mainly reflected on classifier* (Kang et al. 2019; Zhong et al. 2021; Xu et al. 2021; Iscen et al. 2021), while it is possible to achieve remarkable LTR ability by only adjusting or fine-tuning the classifier. Such observation inspires us to do distillation at the feature level. Considering $K$ *experts* with different random seed and *batch size* $B$:

$$\mathcal{L}_{\text{FD}}(\mathscr{F}^s(\mathbf{x}|\theta), \mathscr{F}^t(\mathbf{x}|\theta^t)) = \frac{1}{BK}\sum_{k=1}^{K}\sum_{i=1}^{B}\mathbb{M}(\mathbf{v}_i^s, \mathbf{v}_{k,i}^t), \tag{6}$$

where $\mathbf{v}$ represents the $d$-dim feature given by the feature encoder and $\mathbb{M}(\cdot,\cdot)$ is the proper metric to measure the feature-wise distance between student and teacher(s).

Considering different distance measures 1) cosine similarity (cos), 2) mean squared error (MSE), or 3) Kullback-Leibler divergence (KL), $\mathcal{L}_{\text{FD}}$ will have specific forms:

$$\mathcal{L}_{\text{FD}_{\cos}}(\mathbf{x}|\mathscr{F}^s, \mathscr{F}^t) = \frac{1}{BK}\sum_{k=1}^{K}\sum_{i=1}^{B}\left(1 - \frac{\mathbf{v}_i^s \cdot \mathbf{v}_{k,i}^t}{||\mathbf{v}_i^s|| \cdot ||\mathbf{v}_{k,i}^t||}\right), \tag{7}$$

$$\mathcal{L}_{\text{FD}_{\text{MSE}}}(\mathbf{x}|\mathscr{F}^s, \mathscr{F}^t) = \frac{1}{BK}\sum_{k=1}^{K}\sum_{i=1}^{B}\sum_{j=1}^{d}||\mathbf{v}_{i,j}^s - \mathbf{v}_{k,i,j}^t||_2, \tag{8}$$

$$\mathcal{L}_{\text{FD}_{\text{KL}}}(\mathbf{x}|\mathscr{F}^s, \mathscr{F}^t) = \frac{T^2}{BK}\sum_{k=1}^{K}\sum_{i=1}^{B}\text{KL}(\sigma(\frac{\mathbf{v}_i^s}{T}), \sigma(\frac{\mathbf{v}_{k,i}^t}{T})), \tag{9}$$

where $\sigma$ indicates the *softmax* operation (Eq. 1), $T$ is the temperature scale factor, and $\text{KL}(p,q) = \sum_i p_i \cdot \log(\frac{p_i}{q_i})$.

**Overall Loss Function.** With the carefully designed feature-wise distillation $\mathcal{L}_{\text{FD}}$, the student model will learn more stable and robust features with the supervision of the teacher experts, especially on the tail categories. Note that the student model adopts instance sampling and does not change the label frequency implicitly, the prior bias $\mathcal{B}$ given by Eq. 3 still works to collaborate with $\mathcal{L}_{\text{FD}}$. Hence, with a hyper-parameter $\lambda$ to leverage, we get the final loss:

$$\mathcal{L}(\mathcal{M}(\mathbf{x}|\theta,\phi), \mathbf{y}) = \lambda\mathcal{L}_{\text{FD}} + (1-\lambda)\mathcal{L}_{\text{BC}} \tag{10}$$

**Fisher consistency.** The Fisher consistency (Lin et al. 2004) means *the minimiser of the expected loss should result in a minimal balanced error*. Accounting to (Menon et al. 2020), the pair-wise loss like Eq. 11 is *Fisher consistent* with label weights $\alpha_{\mathbf{y}_i} = \delta_{\mathbf{y}_i}/p_s(\mathbf{y}_i)$ and pairwise label margins $\Delta_{\mathcal{B}_{\mathbf{y}_i,\mathbf{y}_j}} = \mathcal{B}_{\mathbf{y}_j} - \mathcal{B}_{\mathbf{y}_i} = \log(\delta_{\mathbf{y}_j}/\delta_{\mathbf{y}_i})$ for any $\delta \in \mathbb{R}_+^C$.

$$\mathcal{L}(\mathcal{M}(\mathbf{x}_i), \mathbf{y}_i) = \alpha_{\mathbf{y}_i}\log[1 + \sum_{\mathbf{y}_j \neq \mathbf{y}_i} e^{\mathcal{B}_{\mathbf{y}_j} - \mathcal{B}_{\mathbf{y}_i}} \cdot e^{\mathbf{z}_{\mathbf{y}_j} - \mathbf{z}_{\mathbf{y}_i}}] \tag{11}$$

Notice that $\mathcal{L}_{\text{BC}}$ (Eq. 4) is *Fisher consistent* if we set $\delta_{\mathbf{y}} = p_s(\mathbf{y})$ and $p_t(\mathbf{y}) \equiv \frac{1}{C}$ in balanced test data distribution.

For additional feature-level distillation loss $\mathcal{L}_{\text{FD}}$, we revisit Eq. 11 from the decoupling perspective. Regardless of the implementation form of $\mathbb{M}$, $\mathcal{L}_{\text{FD}}$ supervises the student model $\mathcal{M}^s$ to learn features as similar as teachers $\mathcal{M}^t$. Hence, considering the same structure and training manner of $\mathcal{M}^s$ and $\mathcal{M}^t$, it is reasonable to assume that $\mathcal{M}^s$ extracts the *expectation feature* for the model set $\{\mathcal{M}^s, \mathcal{M}_1^t, \cdots, \mathcal{M}_K^t\}$.

$$\hat{\mathbf{v}}_i = \mathbf{E}[\mathscr{F}^s(\mathbf{x}_i|\theta), \mathscr{F}^t(\mathbf{x}_i|\theta_k^t), k \in \{1, \cdots, K\}], \tag{12}$$

Table 1: Top-1 accuracy (%) on CIFAR-10/100-LT. IF: imbalance factor. Results are sorted according to publication time and method category. RW: re-weight wise methods. FW: feature improvement wise methods. ME: multi-expert frameworks. Underline: the best performance in each group. **Bold**: the best performance overall. We report the performance from original papers and reproduce results for unavailable settings according to their official repos. Our CE++ baseline shows powerful performance and our integral FSD + BCL achieves state-of-the-art by a large margin.

| Dataset | Type | Ref. | CIFAR100 | | | | CIFAR10 | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| IF | - | - | 10 | 50 | 100 | 200 | 10 | 50 | 100 | 200 |
| CE | - | NeurIPS19 | 55.70 | 44.02 | 38.32 | 34.56 | 86.39 | 74.94 | 70.36 | 66.21 |
| Focal Loss | RW | ICCV 17 | 55.80 | 44.30 | 38.40 | 35.62 | 86.55 | 76.71 | 70.43 | 68.85 |
| $\tau$ Norm | RW | ICLR20 | 59.10 | 48.23 | 43.60 | 39.30 | 87.80 | 82.78 | 75.10 | 70.30 |
| Causal Norm | RW | NeurIPS20 | 59.60 | 50.30 | 44.10 | - | 88.50 | 83.60 | 80.60 | - |
| LADE | RW | CVPR21 | 61.60 | 50.10 | 45.64 | 40.70 | 88.30 | 82.10 | 79.10 | 73.90 |
| TDE + IDR | RW | CVPR22 | - | 50.30 | 44.90 | - | - | 84.50 | 79.60 | - |
| M2m | FW | CVPR20 | 58.20 | - | 42.90 | - | 87.90 | - | 78.30 | - |
| CAM | FW | AAAI21 | - | 51.70 | 47.80 | - | - | 83.60 | 80.00 | - |
| DiVE | FW | ICCV21 | 62.00 | 51.13 | 45.35 | - | - | - | - | - |
| TSC | FW | CVPR22 | 59.00 | 47.40 | 43.80 | - | 88.70 | 82.90 | 79.70 | - |
| LDAM + DRW | RW+FW | CVPR19 | 58.70 | 46.60 | 42.00 | 38.45 | 88.16 | 81.27 | 77.03 | 74.74 |
| MiSLAS | RW+FW | CVPR21 | 63.20 | 52.30 | 47.00 | - | 90.00 | 85.70 | 82.10 | - |
| Prior-LT | RW+FW | NeurIPS21 | 61.25 | 51.11 | 45.45 | 42.07 | 89.66 | 84.32 | 82.75 | 78.48 |
| PaCo | RW+FW | ICCV21 | 64.2 | 56.00 | 52.00 | 47.75 | 91.50 | 85.43 | 87.96 | 82.29 |
| LTR-WD | RW+FW | CVPR22 | 68.67 | 57.71 | 53.35 | 46.80 | 91.70 | 84.00 | 82.60 | 80.71 |
| BCL | RW+FW | CVPR22 | 64.87 | 56.59 | 51.90 | - | 91.12 | 87.24 | 84.32 | - |
| GCL | RW+FW | CVPR22 | - | 53.55 | 48.71 | 44.88 | - | 85.46 | 82.68 | 79.03 |
| LFME | ME | ECCV20 | 57.77 | 47.21 | 42.30 | 39.01 | 87.13 | 81.45 | 75.33 | 72.88 |
| BBN | ME | CVPR20 | 59.10 | 47.00 | 42.60 | - | 88.32 | 82.20 | 79.80 | - |
| RIDE (4-expert) | ME | ICLR21 | 61.80 | 51.70 | 48.00 | 44.57 | 86.24 | 83.72 | 81.20 | 77.83 |
| Hybrid-SC | ME | CVPR21 | - | 51.90 | 46.70 | - | - | 85.40 | 81.40 | - |
| TADE | ME | ICCV21 | 63.60 | 53.90 | 49.80 | 44.71 | 89.99 | 85.77 | 82.87 | 78.01 |
| ACE (4-expert) | ME | ICCV21 | - | 51.90 | 49.60 | - | - | 84.90 | 81.40 | - |
| NCL | ME | CVPR22 | 63.84 | 58.20 | 54.2 | 49.52 | 91.10 | 87.30 | 85.50 | 82.18 |
| CE++ | - | - | 70.25 | 55.87 | 49.03 | 44.15 | 93.84 | 87.41 | 83.28 | 71.24 |
| + FSD | FW | - | 71.74 | 56.63 | 50.55 | 44.94 | 94.2 | 88.59 | 84.54 | 73.20 |
| + BCL | RW | - | 72.20 | 61.64 | 55.35 | 50.16 | 94.10 | 89.80 | 86.90 | 82.45 |
| + FSD + BCL | FW+RW+ME | - | **74.55** | **63.07** | **57.24** | **51.04** | **95.06** | **91.38** | **89.17** | **83.48** |

Then, for the pair-wise loss Eq. 11, we have:

$$\begin{aligned}\mathcal{L} &= \alpha_{\mathbf{y}_i} \log[1 + \sum_{\mathbf{y}_j \neq \mathbf{y}_i} e^{\mathcal{B}_{\mathbf{y}_j} - \mathcal{B}_{\mathbf{y}_i}} \cdot e^{\mathscr{W}(\hat{\mathbf{v}}_i|\phi)_{\mathbf{y}_j} - \mathscr{W}(\hat{\mathbf{v}}_i|\phi)_{\mathbf{y}_i}}] \\ &= \alpha_{\mathbf{y}_i} \log[1 + \sum_{\mathbf{y}_j \neq \mathbf{y}_i} e^{\mathcal{B}_{\mathbf{y}_j} - \mathcal{B}_{\mathbf{y}_i}} \cdot e^{\hat{\mathbf{z}}_{\mathbf{y}_j} - \hat{\mathbf{z}}_{\mathbf{y}_i}}]\end{aligned} \tag{13}$$

Notice that the label *weight* and pairwise label *margins* do not change with additional $\mathcal{L}_{\text{FD}}$, and thus the aforementioned conclusion still satisfies, which means *the final loss* (Eq. 10) *is still Fisher consistent*.

# Experiment

## Datasets

**CIFAR-10/100-LT.** CIFAR-10/100 (Krizhevsky, Hinton et al. 2009) have $10/100$ classes with $60,000$ images in $32 \times 32$ resolution. We follow (Cui et al. 2019; Cao et al. 2019) to sample train set of each class with exponential functions to create the long-tailed versions while remaining the validation set uniform distributed. The imbalance factor $\gamma$ indicates the skewness of the dataset, which is the ratio between the most and the least frequent classes. We employ $\gamma = [10, 50, 100, 200]$ for comprehensive comparisons.

**ImageNet-LT** is the subset of the large-scale ImageNet 2012 (Russakovsky et al. 2015), which is widely used in classification and localization tasks. The train samples in ImageNet-LT are sampled through Pareto distribution with power value $\alpha = 6$. It contains $115.8K$ images from $1,000$ classes. The most/least class number is $1,280/5$ respectively, and thus $\gamma = 256$. we utilize the balanced validation set constructed by (Cui et al. 2019) for fair comparisons.

**iNaturalist 2018** (Van Horn et al. 2018) is the large-scale real-world dataset for LTR. With over $437.5K$ images and $8,142$ classes ($\gamma = 500$), it suffers from extremely label long-tailed distribution and fine-grained challenges. We follow (Cao et al. 2019) to utilize the official splits of training and validation sets in our experiments.

## Implement Details & CE++

With the prosperous progress in LTR community, we notice the naïve ERM baseline fails to avoid implicit tricks to provide fair comparisons. Hence, we propose a novel baseline named CE++, which integrates recent advances (Loshchilov, Hutter et al. 2016; DeVries, Taylor et al. 2017; Chen et al. 2021; Cubuk et al. 2020; Alshammari et al. 2022) and re-normalizes the baseline in LTR community.

For CIFAR-LT, we follow LTR-WD (Alshammari et al. 2022) to set weight decay to $5e-3$ for ResNet-34 and use stochastic gradient descent with momentum $0.9$. All models are trained for 200 epochs with learning rate $0.01$ and

Table 2: Top-1 accuracy (%) on ImageNet-LT & iNaturalist 2018. Results are sorted by publication time. R-50: ResNet-50. RX-50: ResNeXt-50. Our integral FSD + BCL outperforms strong augmentation and long training time methods.

| Method | Ref. | ImageNet-LT | | iNat. 18 |
|---|---|---|---|---|
| | | R-50 | RX-50 | R-50 |
| CE | - | 38.88 | 44.40 | 60.88 |
| OLTR | CVPR19 | 40.36 | - | 63.90 |
| CB | CVPR19 | 40.85 | - | 63.50 |
| LDAM + DRW | NeurIPS19 | 45.75 | - | 68.00 |
| BBN | CVPR20 | 48.30 | 49.30 | 66.29 |
| NCM | ICLR20 | 44.30 | 47.30 | 63.10 |
| cRT | ICLR20 | 47.30 | 49.60 | 65.20 |
| $\tau$ Norm | ICLR20 | 46.7 0 | 49.40 | 65.6 |
| LWS | ICLR20 | 47.70 | 49.70 | 65.90 |
| BS | NeurIPS20 | 53.00 | - | 66.40 |
| RIDE | ICLR21 | 55.40 | 56.80 | 72.60 |
| DisAlign | CVPR21 | 52.90 | 53.40 | 70.60 |
| DiVE | ICCV21 | 53.10 | - | 71.71 |
| SSD | ICCV21 | - | 56.00 | 71.50 |
| ACE | ICCV21 | 54.70 | 56.60 | 72.90 |
| PaCo | ICCV21 | 56.12 | 57.23 | 72.13 |
| CBD | BMVC21 | 53.90 | - | 70.50 |
| TSC | CVPR22 | 52.40 | - | 69.70 |
| RIDE + CMO | CVPR22 | 56.20 | - | 72.80 |
| BCL | CVPR22 | 56.00 | 57.10 | 71.80 |
| CKT | CVPR22 | - | 54.20 | - |
| GCL | CVPR22 | - | 54.88 | 72.01 |
| NCL | CVPR22 | 57.65 | 58.11 | 73.14 |
| CE++ | - | 49.00 | 50.87 | 64.22 |
| + FSD | - | 51.11 | 52.24 | 65.98 |
| + BCL | - | 54.04 | 55.63 | 71.11 |
| + FSD + BCL | - | **57.91** | **58.52** | **73.59** |

mini-batch 64. The learning scheduler is Cosine Annealing (Loshchilov, Hutter et al. 2016) with an ending rate of 0. Further, Cutout (DeVries, Taylor et al. 2017) and AutoAug (Chen et al. 2021) are adopted to compensate origin data augmentation strategies (He et al. 2016).

For large-scale datasets, we follow LTR-WD (Alshammari et al. 2022) to set weight decay as $5e-4/1e-4$ for ImageNet-LT/iNaturalist 2018 and train 180/90 epochs respectively. We replace AutoAug with RandAug (Cubuk et al. 2020) while keeping other settings consistent with CIFAR-LT. Finally, we adopt horizontal flip as post-hoc augmentation. For our FSD, we set $K = 3$ experts and $\lambda = 0.4$ in default. The seed for each expert is randomly selected.

## Competing Methods

**Baselines.** The vanilla baseline (CE) conducts plain training with standard cross-entropy loss (Cui et al. 2019). The common networks are ResNet-32 (Idelbayev 2021) (CIFAR-10/100-LT), ResNet-50 (He et al. 2016) (ImageNet-LT, iNaturalist 2018) and ResNeXt-50 (Xie et al. 2017) (ImageNet-LT). In addition, to align with previous works that contain the additional proposal-independent tricks implicitly, we adopt our CE++ for additional fair comparisons.

**Feature-wise methods** modify the feature sampling or learning manners to cope with long-tailed datasets. M2m (Kim et al. 2020) generates pseudo samples for training and optimizing. CAM (Zhang et al. 2021c) and CMO (Park et al. 2022) enrich the training samples via feature combination. DiVE (He et al. 2021) adopts knowledge distillation and take the teacher feature as an additional training sample for the student model. Recent state-of-the-art (Li et al. 2022c; Cui et al. 2021; Zhu et al. 2022) adopts contrastive frameworks to improve representation learning.

Table 3: Ablation study of the proposed CE++ on CIFAR-100-LT ($\gamma = 100$). Each component contributes to performance gains and the overall baseline achieves the best performance. Group according to (Zhang et al. 2021b). Cos.Anne.: Cosine Annealing lr. WD: weight decay.

| Method | Many | Medium | Few | All |
|---|---|---|---|---|
| CE | 75.34 | 33.66 | 0.13 | 38.32 |
| + ResNet-34 | 75.46 | 42.11 | 9.20 | 44.77 |
| + Cos.Anne. | 75.69 | 44.29 | 12.7 | 45.10 |
| + Cutout | 77.14 | 47.17 | 15.23 | 48.08 |
| + AutoAug | 76.89 | 47.37 | **15.70** | 48.20 |
| + WD | 81.76 | 54.85 | 10.82 | 48.76 |
| CE++ (+ Flip) | **81.88** | **54.92** | 11.07 | **49.03** |

**Re-weight methods** focus on label weighting (Cui et al. 2019; Lin et al. 2017; Cao et al. 2019; Zhong et al. 2021) or logits adjusting (Menon et al. 2020; Hong et al. 2021; Ren et al. 2020; Xu et al. 2021; Samuel, Chechik et al. 2021; Yu et al. 2022; Li et al. 2022b) based on standard cross entropy loss. In addition, some methods (Kang et al. 2019; Tang, Huang, and Zhang 2020; Alshammari et al. 2022) are also effective by directly adjusting the classifier's weight.

**Multi-expert methods** have shown powerful generalization in LTR and can be classified into two categories. 1) Each expert learn *different* aspects of knowledge w.r.t. specific classes and then aggregates together (Xiang et al. 2020; Zhou et al. 2020; Wang et al. 2020; Cai et al. 2021; Wang et al. 2021b). 2) Each expert learns the *same* knowledge w.r.t. each class, and then reduces the uncertainty on minority classes (Zhang et al. 2021b; Li, Wang, and Wu 2021; Li et al. 2022a). Note that our FSD belongs to the latter one.

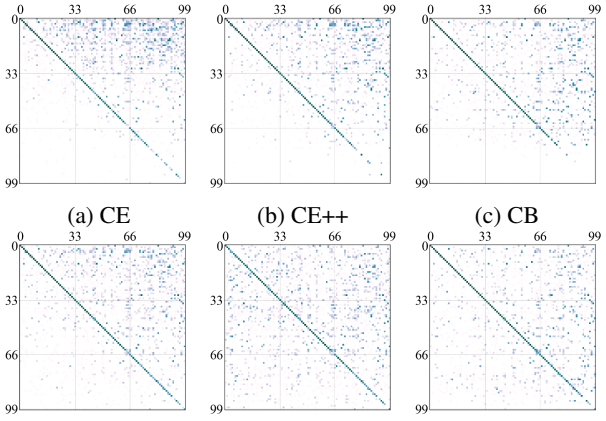
## Comparison with state-of-the-art

We conduct comprehensive comparisons on CIFAR-LT (Tab. 1) and large-scale datasets (Tab. 2). Our method uses 3 teacher experts and cosine distance (c.f. Eq. 7) in default. We present the performance of our CE++ baseline, with FSD, with BCL, and with both of them to show how each component collaborates together. For other methods, we report the performance in original papers and reproduce the missing settings through their official repos. For contrastive approaches (Cui et al. 2021; Li et al. 2022a), we keep the training epochs *consistent* with ours for fair comparisons.

Results show that our proposed baseline is powerful and even outperforms SOTA in some settings. The novel feature-wise distillation ameliorates BCL further and the whole FSD achieves state-of-the-art performance on all datasets by a large margin. It is worth mentioning that the self-supervision methods with more training epochs and larger GPU memory requisition are still competitive in large-scale datasets, e.g., PaCo (73.2%) and NCL (74.9%) on iNaturalist 2018. However, with fair settings, our proposal outperforms them by $\mathbf{1.46}\%$ and $\mathbf{0.45}\%$, where our FSD only requires a single student expert for inference, and thus it is light-weight compared to PaCo ($2\times$) and NCL ($3\times$).

Table 4: Results of reproduced methods based on our CE++ on CIFAR-100-LT ($\gamma = 100$). Each method gets improved to some extent compared to the original results in Tab. 1.

| Method | Ref. | Many | Medium | Few | All |
|---|---|---|---|---|---|
| CE++ | - | **81.76** | 54.85 | 10.82 | 48.76 |
| CB | CVPR19 | 75.94 | 50.00 | 18.27 | 50.28 |
| LDAM | NeurIPS19 | 74.33 | 51.11 | 19.87 | 50.34 |
| $\tau$ Norm | ICLR20 | 74.17 | 53.89 | 24.47 | 52.41 |
| RIDE | ICLR21 | 77.06 | 55.33 | 24.87 | 54.72 |
| LA | ICLR21 | 78.94 | 56.97 | 25.12 | 55.35 |
| MiSLAS | CVPR21 | 75.94 | 58.77 | 28.87 | 56.12 |
| TADE | ICCV21 | 81.26 | 57.74 | 19.53 | 55.73 |
| Prior-LT | NeurIPS21 | 79.17 | 58.14 | 26.60 | 56.22 |
| LTR-WD | CVPR22 | 77.26 | 56.39 | 22.53 | 54.32 |
| GCL | CVPR22 | 78.94 | 57.47 | 27.94 | 56.88 |
| Ours | - | 77.94 | **59.23** | **30.77** | **57.24** |




(a) CE (b) CE++ (c) CB

(d) LTR-WD (e) GCL (f) Ours

Figure 4: Visualized log-confusion matrix on CIFAR-100-LT ($\gamma = 100$). $x$-axis: ground truth. $y$-axis: predicted label.

## Further Analysis

**Ablation Studies of CE++.** We conduct detailed ablation study to investigate in the contribution of each component in CE++, see Tab. 3. Note that ResNet with proper WD will boost the model significantly (Alshammari et al. 2022). We add additional class irrelevant augmentations and change the learning rate scheduler. The CE++ is powerful enough to provide fair comparisons in the long-tailed community.

**Results based on CE++.** Recent methods adopt several training tricks or are implemented in different frameworks. To align with CE++, we reproduce recent state-of-the-art methods on CE++ to justify our novel baseline is easy to plug-and-play into existing methods, as in Tab. 4 illustrates. We follow (Zhang et al. 2021b) to set groups as Many ($> 100$), Medium ($20 \sim 100$), and Few ($< 20$) according to instance number. Our FSD ameliorates the few group by a large margin and make overall accuracy improved consistently. All approaches get improved using our novel baseline compared to the original implementation in Tab. 1. The visualized log-confusion matrix in Fig. 4 further verifies the effectiveness of FSD in tackling the head bias and tail variance, which makes more balanced predictions, i.e., the misclassified samples (non-diagonal elements) are balanced distributed. Compared to GCL (Li et al. 2022b) with BCL, FSD improves the tail accuracy remarkably (brighter diagonal elements) by abating the model variance on the tail.

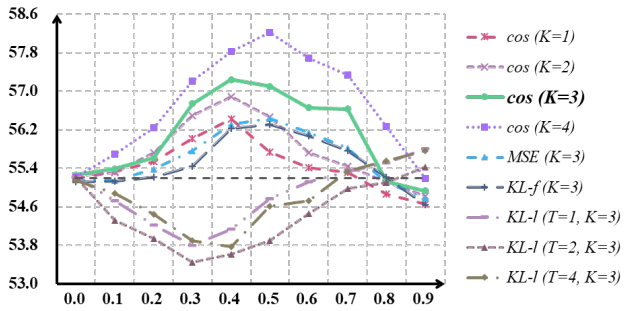



Figure 5: Ablation study w.r.t. $K$, $\mathbb{M}$ and $\lambda$. $x$-axis: value of $\lambda$. $y$-axis: Top-1 accuracy (%) on CIFAR-100-LT ($\gamma = 100$). *KL-f*: KL distance at feature-level. *KL-l*: KL distance at logits-level. $T$: distillation temperature.

**How to Select $K$, $\mathbb{M}$ and $\lambda$.** Fig. 5 presents the ablation study of all hyper-parameters in FSD, i.e., the number of teacher $K$, the distillation metric $\mathbb{M}$, and the weight of FSD loss $\lambda$, together with logits-level distillation who's teachers are trained with BCL as a comparison. Note that a larger $K$ (i.e., more experts) contributes to higher student performance, while it requires more computational overheads. Hence, we set $K = 3$ as a trade-off. For the distillation metric, cosine similarity is slightly better than the others. Note that for logits-level distillation, where the teachers are trained with BCL, the performance drops when $\lambda$ gets larger because the teacher experts learn inaccuracy Bayesian bias and the distillation distorts it further as we analyzed in Fig. 2.

**Complexity.** Intuitively, our FSD seems to require higher complexity compared with baseline and backbone-shared ME methods. However, similar to (Iscen et al. 2021), each teacher expert can be trained in parallel. Hence, FSD reduces the memory demand for training and occupies less storage when it comes to implementation compared to nested learning (prior SOTA). FSD only requires a single student model for inference while previous ensemble methods deploy all experts for the final aggregation.

## Conclusion

In this paper, we revisit BCL with ME framework in LTR and pinpoint the conflict between them lies in the essence that experts' train manner potentially shifts statistical bias while BCL strictly requires the precise prior for each expert. Therefore, we propose the novel Feature Space Distillation to aggregate different feature-level expert knowledge to eliminate distorted bias on teacher classifiers such that BCL is qualified on the student expert to learn an unbiased inference model. We also integrate latest conspicuous advances and propose an unprecedented strong baseline named CE++ to align with recent progress in long-tailed community. Extensive experiments demonstrate the power of CE++ and FSD, and the integral FSD improves robustness on the tail and achieves remarkable state-of-the-art performance. We leave it for future work to compensate the sacrificed head accuracy through robust class-irrelevant features, and simplify the two-stage framework to reduce the training complexity.

# References


Alshammari, S.; Wang, Y.-X.; Ramanan, D.; and Kong, S. 2022. Long-tailed recognition via weight balancing. In *CVPR*, 6897–6907.

Ashukha, A.; Lyzhov, A.; Molchanov, D.; and Vetrov, D. 2020. Pitfalls of in-domain uncertainty estimation and ensembling in deep learning. *arXiv preprint arXiv:2002.06470*.

Buda, M.; Maki, A.; Mazurowski, M. A.; et al. 2018. A systematic study of the class imbalance problem in convolutional neural networks. *Neural Networks*, 106: 249–259.

Cai, J.; Wang, Y.; Hwang, J.-N.; et al. 2021. Ace: Ally complementary experts for solving long-tailed recognition in one-shot. In *ICCV*, 112–121.

Cao, K.; Wei, C.; Gaidon, A.; Arechiga, N.; and Ma, T. 2019. Learning imbalanced datasets with label-distribution-aware margin loss. *NeurIPS*, 32.

Chen, T.; Kornblith, S.; Norouzi, M.; and Hinton, G. 2020. A simple framework for contrastive learning of visual representations. In *ICML*, 1597–1607. PMLR.

Chen, Y.; Li, Y.; Kong, T.; Qi, L.; Chu, R.; Li, L.; and Jia, J. 2021. Scale-aware Automatic Augmentation for Object Detection. In *CVPR*.

Chou, H.-P.; Chang, S.-C.; Pan, J.-Y.; Wei, W.; and Juan, D.-C. 2020. Remix: rebalanced mixup. In *ECCV*, 95–110. Springer.

Chu, P.; Bian, X.; Liu, S.; and Ling, H. 2020. Feature space augmentation for long-tailed data. In *ECCV*, 694–710. Springer.

Cubuk, E. D.; Zoph, B.; Shlens, J.; and Le, Q. V. 2020. Randaugment: Practical automated data augmentation with a reduced search space. In *CVPR workshops*, 702–703.

Cui, J.; Zhong, Z.; Liu, S.; Yu, B.; and Jia, J. 2021. Parametric contrastive learning. In *ICCV*, 715–724.

Cui, Y.; Jia, M.; Lin, T.-Y.; Song, Y.; and Belongie, S. 2019. Class-balanced loss based on effective number of samples. In *CVPR*, 9268–9277.

DeVries, T.; Taylor, G. W.; et al. 2017. Improved regularization of convolutional neural networks with cutout. *arXiv preprint arXiv:1708.04552*.

Garau, N.; Bisagno, N.; Bródka, P.; and Conci, N. 2021. DECA: Deep viewpoint-Equivariant human pose estimation using Capsule Autoencoders. In *CVPR*, 11677–11686.

Girshick, R.; Donahue, J.; Darrell, T.; and Malik, J. 2014. Rich feature hierarchies for accurate object detection and semantic segmentation. In *CVPR*, 580–587.

Han, H.; Wang, W.; Mao, B.; et al. 2005. Borderline-SMOTE: A New Over-Sampling Method in Imbalanced Data Sets Learning. In *ICIC*, volume 3644 of *Lecture Notes in Computer Science*, 878–887. Springer.

He, H.; Garcia, E. A.; et al. 2009. Learning from Imbalanced Data. *IEEE Trans. Knowl. Data Eng.*, 21(9): 1263–1284.

He, K.; Fan, H.; Wu, Y.; Xie, S.; and Girshick, R. 2020. Momentum contrast for unsupervised visual representation learning. In *CVPR*, 9729–9738.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *CVPR*, 770–778.

He, Y.-Y.; Wu, J.; Wei, X.-S.; et al. 2021. Distilling virtual examples for long-tailed recognition. In *ICCV*, 235–244.

Hinton, G.; Vinyals, O.; Dean, J.; et al. 2015. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2(7).

Hong, Y.; Han, S.; Choi, K.; Seo, S.; Kim, B.; and Chang, B. 2021. Disentangling Label Distribution for Long-Tailed Visual Recognition. In *CVPR*, 6626–6636. Computer Vision Foundation / IEEE.

Huang, C.-H. P.; Yi, H.; Höschle, M.; Safroshkin, M.; Alexiadis, T.; Polikovsky, S.; Scharstein, D.; and Black, M. J. 2022. Capturing and inferring dense full-body human-scene contact. In *CVPR*, 13274–13285.

Huang, G.; Liu, Z.; Van Der Maaten, L.; and Weinberger, K. Q. 2017. Densely connected convolutional networks. In *CVPR*, 4700–4708.

Idelbayev, Y. 2021. Proper ResNet Implementation for CIFAR10/CIFAR100 in PyTorch. https://github.com/akamaster/pytorch_resnet_cifar10.

Iscen, A.; Araujo, A.; Gong, B.; and Schmid, C. 2021. Class-Balanced Distillation for Long-Tailed Visual Recognition. In *BMVC*, 165. BMVA Press.

Jamal, M. A.; Brown, M.; Yang, M.-H.; Wang, L.; and Gong, B. 2020. Rethinking class-balanced methods for long-tailed visual recognition from a domain adaptation perspective. In *CVPR*, 7610–7619.

Kang, B.; Xie, S.; Rohrbach, M.; Yan, Z.; Gordo, A.; Feng, J.; and Kalantidis, Y. 2019. Decoupling representation and classifier for long-tailed recognition. *arXiv preprint arXiv:1910.09217*.

Khosla, P.; Teterwak, P.; Wang, C.; Sarna, A.; Tian, Y.; Isola, P.; Maschinot, A.; Liu, C.; and Krishnan, D. 2020. Supervised contrastive learning. *NeurIPS*, 33: 18661–18673.

Kim, J.; Jeong, J.; Shin, J.; et al. 2020. M2m: Imbalanced classification via major-to-minor translation. In *CVPR*, 13896–13905.

Krizhevsky, A.; Hinton, G.; et al. 2009. Learning multiple layers of features from tiny images.

Kwon, H.; Kim, M.; Kwak, S.; and Cho, M. 2020. Motionsqueeze: Neural motion feature learning for video understanding. In *ECCV*, 345–362. Springer.

Li, J.; Tan, Z.; Wan, J.; Lei, Z.; and Guo, G. 2022a. Nested Collaborative Learning for Long-Tailed Visual Recognition. In *CVPR*, 6949–6958.

Li, M.; Cheung, Y.-m.; Lu, Y.; et al. 2022b. Long-tailed Visual Recognition via Gaussian Clouded Logit Adjustment. In *CVPR*, 6929–6938.

Li, T.; Cao, P.; Yuan, Y.; Fan, L.; Yang, Y.; Feris, R. S.; Indyk, P.; and Katabi, D. 2022c. Targeted supervised contrastive learning for long-tailed recognition. In *CVPR*, 6918–6928.

Li, T.; Wang, L.; and Wu, G. 2021. Self supervision to distillation for long-tailed visual recognition. In *ICCV*, 630–639.

Lin, T.-Y.; Goyal, P.; Girshick, R.; He, K.; and Dollár, P. 2017. Focal loss for dense object detection. In *ICCV*, 2980–2988.

Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; and Zitnick, C. L. 2014. Microsoft coco: Common objects in context. In *ECCV*, 740–755. Springer.

Lin, Y.; et al. 2004. A note on margin-based loss functions in classification. *Statistics & probability letters*, 68(1): 73–82.

Loshchilov, I.; Hutter, F.; et al. 2016. Sgdr: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983*.

Mao, F.; Wu, X.; Xue, H.; and Zhang, R. 2018. Hierarchical video frame sequence representation with deep convolutional graph network. In *ECCV Workshops*, 0–0.

Menon, A. K.; Jayasumana, S.; Rawat, A. S.; Jain, H.; Veit, A.; and Kumar, S. 2020. Long-tail learning via logit adjustment. *arXiv preprint arXiv:2007.07314*.

Nakano, N.; Sakura, T.; Ueda, K.; Omura, L.; Kimura, A.; Iino, Y.; Fukashiro, S.; and Yoshioka, S. 2020. Evaluation of 3D markerless motion capture accuracy using OpenPose with multiple video cameras. *Frontiers in sports and active living*, 2: 50.

Park, S.; Hong, Y.; Heo, B.; Yun, S.; and Choi, J. Y. 2022. The Majority Can Help The Minority: Context-rich Minority Oversampling for Long-tailed Classification. In *CVPR*, 6887–6896.

Ren, J.; Yu, C.; Ma, X.; Zhao, H.; Yi, S.; et al. 2020. Balanced meta-softmax for long-tailed visual recognition. *NeurIPS*, 33: 4175–4186.

Ronneberger, O.; Fischer, P.; and Brox, T. 2015. U-net: Convolutional networks for biomedical image segmentation. In *MICCAI*, 234–241. Springer.

Russakovsky, O.; Deng, J.; Su, H.; Krause, J.; Satheesh, S.; Ma, S.; Huang, Z.; Karpathy, A.; Khosla, A.; Bernstein, M.; Berg, A. C.; and Fei-Fei, L. 2015. ImageNet Large Scale Visual Recognition Challenge. *IJCV*, 115(3): 211–252.

Samuel, D.; Chechik, G.; et al. 2021. Distributional robustness loss for long-tail learning. In *ICCV*, 9495–9504.

Tan, J.; Wang, C.; Li, B.; Li, Q.; Ouyang, W.; Yin, C.; and Yan, J. 2020. Equalization loss for long-tailed object recognition. In *CVPR*, 11662–11671.

Tang, K.; Huang, J.; and Zhang, H. 2020. Long-tailed classification by keeping the good and removing the bad momentum causal effect. *NeurIPS*, 33: 1513–1524.

Van Horn, G.; Mac Aodha, O.; Song, Y.; Cui, Y.; Sun, C.; Shepard, A.; Adam, H.; Perona, P.; and Belongie, S. 2018. The inaturalist species classification and detection dataset. In *CVPR*, 8769–8778.

Wang, J.; Lukasiewicz, T.; Hu, X.; Cai, J.; and Xu, Z. 2021a. RSG: A Simple but Effective Module for Learning Imbalanced Datasets. In *CVPR*, 3784–3793. Computer Vision Foundation / IEEE.

Wang, P.; Han, K.; Wei, X.-S.; Zhang, L.; and Wang, L. 2021b. Contrastive learning based hybrid networks for long-tailed image classification. In *CVPR*, 943–952.

Wang, X.; Lian, L.; Miao, Z.; Liu, Z.; and Yu, S. X. 2020. Long-tailed recognition by routing diverse distribution-aware experts. *arXiv preprint arXiv:2010.01809*.

Wu, C.-Y.; Girshick, R.; He, K.; Feichtenhofer, C.; and Krahenbuhl, P. 2020. A multigrid method for efficiently training video models. In *CVPR*, 153–162.

Xiang, L.; Ding, G.; Han, J.; et al. 2020. Learning from multiple experts: Self-paced knowledge distillation for long-tailed classification. In *ECCV*, 247–263. Springer.

Xie, S.; Girshick, R.; Dollár, P.; Tu, Z.; and He, K. 2017. Aggregated residual transformations for deep neural networks. In *CVPR*, 1492–1500.

Xu, Y.; Zhang, J.; Zhang, Q.; and Tao, D. 2022. ViTPose: Simple Vision Transformer Baselines for Human Pose Estimation. *arXiv preprint arXiv:2204.12484*.

Xu, Z.; Chai, Z.; Yuan, C.; et al. 2021. Towards calibrated model for long-tailed visual recognition from prior perspective. *NeurIPS*, 34: 7139–7152.

Yang, Y.; Xu, Z.; et al. 2020. Rethinking the value of labels for improving class-imbalanced learning. *NeurIPS*, 33: 19290–19301.

Ye, H.-J.; Chen, H.-Y.; Zhan, D.-C.; and Chao, W.-L. 2020. Identifying and compensating for feature deviation in imbalanced deep learning. *arXiv preprint arXiv:2001.01385*.

Yu, S.; Guo, J.; Zhang, R.; Fan, Y.; Wang, Z.; and Cheng, X. 2022. A Re-Balancing Strategy for Class-Imbalanced Classification Based on Instance Difficulty. In *CVPR*, 70–79.

Zeng, A.; Sun, X.; Huang, F.; Liu, M.; Xu, Q.; and Lin, S. 2020. Srnet: Improving generalization in 3d human pose estimation with a split-and-recombine approach. In *ECCV*, 507–523. Springer.

Zhang, X.; Wu, Z.; Weng, Z.; Fu, H.; Chen, J.; Jiang, Y.-G.; and Davis, L. S. 2021a. Videolt: Large-scale long-tailed video recognition. In *ICCV*, 7960–7969.

Zhang, Y.; Hooi, B.; Hong, L.; and Feng, J. 2021b. Test-agnostic long-tailed recognition by test-time aggregating diverse experts with self-supervision. *arXiv preprint arXiv:2107.09249*.

Zhang, Y.; Wei, X.-S.; Zhou, B.; and Wu, J. 2021c. Bag of tricks for long-tailed visual recognition with deep convolutional neural networks. In *AAAI*, 3447–3455.

Zhong, Z.; Cui, J.; Liu, S.; and Jia, J. 2021. Improving Calibration for Long-Tailed Recognition. In *CVPR*, 16489–16498. Computer Vision Foundation / IEEE.

Zhou, B.; Cui, Q.; Wei, X.-S.; and Chen, Z.-M. 2020. Bbn: Bilateral-branch network with cumulative learning for long-tailed visual recognition. In *CVPR*, 9719–9728.

Zhou, B.; Lapedriza, A.; Khosla, A.; Oliva, A.; and Torralba, A. 2017. Places: A 10 million Image Database for Scene Recognition. *IEEE TPAMI*.

Zhu, J.; Wang, Z.; Chen, J.; Chen, Y.-P. P.; and Jiang, Y.-G. 2022. Balanced Contrastive Learning for Long-Tailed Visual Recognition. In *CVPR*, 6908–6917.