Balanced Product of Calibrated Experts for Long-Tailed Recognition

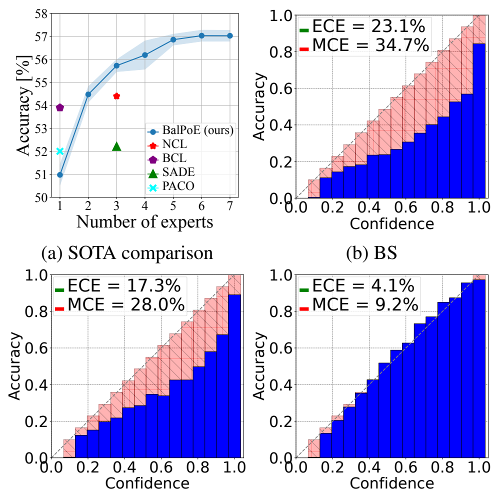
Emanuel Sanchez Aimar[1], Arvi Jonnarth[1,*], Michael Felsberg[1,†], Marco Kuhlmann[2]
[1]Department of Electrical Engineering, Linköping University, Sweden
[2]Department of Computer and Information Science, Linköping University, Sweden
{emanuel.sanchez.aimar,arvi.jonnarth,michael.felsberg,marco.kuhlmann}@liu.se

Abstract

Many real-world recognition problems are characterized by long-tailed label distributions. These distributions make representation learning highly challenging due to limited generalization over the tail classes. If the test distribution differs from the training distribution, e.g. uniform versus long-tailed, the problem of the distribution shift needs to be addressed. A recent line of work proposes learning multiple diverse experts to tackle this issue. Ensemble diversity is encouraged by various techniques, e.g. by specializing different experts in the head and the tail classes. In this work, we take an analytical approach and extend the notion of logit adjustment to ensembles to form a Balanced Product of Experts (BalPoE). BalPoE combines a family of experts with different test-time target distributions, generalizing several previous approaches. We show how to properly define these distributions and combine the experts in order to achieve unbiased predictions, by proving that the ensemble is Fisher-consistent for minimizing the balanced error. Our theoretical analysis shows that our balanced ensemble requires calibrated experts, which we achieve in practice using mixup. We conduct extensive experiments and our method obtains new state-of-the-art results on three long-tailed datasets: CIFAR-100-LT, ImageNet-LT, and iNaturalist-2018. Our code is available at https://github.com/emasa/BalPoE-CalibratedLT.

1. Introduction

Recent developments within the field of deep learning, enabled by large-scale datasets and vast computational resources, have significantly contributed to the progress in many computer vision tasks [33]. However, there is a discrepancy between common evaluation protocols in benchmark datasets and the desired outcome for real-world problems. Many benchmark datasets assume a balanced label distribution with a sufficient number of samples for each class. In this setting, empirical risk minimization (ERM) has been widely adopted to solve multi-class classification and is the key to many state-of-the-art (SOTA) methods, see Figure 1. Unfortunately, ERM is not well-suited for imbalanced or long-tailed (LT) datasets, in which *head classes* have many more samples than *tail classes*, and where an unbiased predictor is desired at test time. This is a common scenario for real-world problems, such as object detection [42], medical diagnosis [18] and fraud detection [50]. On the one hand, extreme class imbalance biases the classifier towards head classes [31, 57]. On the other hand, the paucity of data prevents learning good representations for less-represented classes, especially in few-shot data regimes [67]. Addressing class imbalance is also relevant from the perspective of algorithmic fairness, since incorporating unfair biases into the models can have life-changing consequences in real-world decision-making systems [46].



(a) SOTA comparison (b) BS (c) SADE (d) Our approach

Figure 1. (a) SOTA comparison. Many SOTA approaches employ Balanced Softmax (BS) [51] and its extensions to mitigate the long-tailed bias [12, 38, 70, 72]. However, we observe that single-expert bias-adjusted models, as well as multi-expert extensions, are still poorly calibrated by default. Reliability plots for (b) BS, (c) SADE [70] and (d) Calibrated BalPoE (our approach).

Previous work has approached the problem of class imbal-

*Affiliation: Husqvarna Group, Huskvarna, Sweden.
†Co-affiliation: University of KwaZulu-Natal, Durban, South Africa.

ance by different means, including *data re-sampling* [4, 8], *cost-sensitive learning* [1, 65], and *margin modifications* [6, 56]. While intuitive, these methods are not without limitations. Particularly, over-sampling can lead to overfitting of rare classes [13], under-sampling common classes might hinder feature learning due to the omitting of valuable information [31], and loss re-weighting can result in optimization instability [6]. Complementary to these approaches, ensemble learning has empirically shown benefits over single-expert models in terms of generalization [34] and predictive uncertainty [36] on balanced datasets. Recently, expert-based approaches [5, 61] also show promising performance gains in the long-tailed setting. However, the theoretical reasons for how expert diversity leads to better generalization over different label distributions remain largely unexplored.

In this work, we seek to formulate an ensemble where each expert targets different classes in the label distribution, and the ensemble as a whole is provably unbiased. We accomplish this by extending the theoretical background for logit adjustment to the case of learning diverse expert ensembles. We derive a constraint for the target distributions, defined in terms of expert-specific biases, and prove that fulfilling this constraint yields Fisher consistency with the *balanced error*. In our formulation, we assume that the experts are calibrated, which we find not to be the case by default in practice. Thus, we need to assure that the assumption is met, which we realize using mixup [69]. Our contributions can be summarized as follows:

- We extend the notion of logit adjustment based on label frequencies to balanced ensembles. We show that our approach is theoretically sound by proving that it is *Fisher-consistent* for minimizing the balanced error.
- Proper calibration is a necessary requirement to apply the previous theoretical result. We find that mixup is vital for expert calibration, which is not fulfilled by default in practice. Meeting the calibration assumption ensures Fisher consistency, and performance gains follow.
- Our method reaches new state-of-the-art results on three long-tailed benchmark datasets.

2. Related work

Data resampling and loss re-weighting. Common approaches resort to resampling the data or re-weighting the losses to achieve a more balanced distribution. Re-sampling approaches include under-sampling the majority classes [16], over-sampling the minority classes [8, 20], and class-balanced sampling [28, 62]. Re-weighting approaches are commonly based on the inverse class frequency [28], the *effective per-class number of samples* [13], or sample-level weights [41]. In contrast, we make use of theoretically-sound margin modifications without inhibiting feature learning.

Margin modifications. Enforcing large margins for minority classes has been shown to be an effective regularizer under class imbalance [6, 56]. Analogously, a posthoc *logit adjustment* (LA) can be seen as changing the class margins during inference time to favor tail classes [48, 57]. Particularly, the approach proposed by Menon *et al*. [48] is shown to be *Fisher-consistent* [43] for minimizing the balanced error. Finally, Hong *et al*. [26] generalize LA to accommodate an arbitrary, known target label distribution. We extend single-model LA to a multi-expert framework by parameterizing diverse target distributions for different experts and show how to formulate the margins such as to maintain Fisher consistency for the whole ensemble.

Calibration. Modern over-parameterized neural networks are notorious for predicting uncalibrated probability estimates [19], being wrongly overconfident in the presence of out-of-distribution data [40]. These issues are further exacerbated under class imbalance [71]. Mixup [69] and its extensions [60, 68] have been effective at improving confidence calibration [58], and to some extent, generalization [68, 69], in the balanced setting. However, mixup does not change the label distribution [66], thus several methods modify sampling [49, 66] and mixing components [9, 66] to boost the performance on tail classes. Zhong *et al*. [71] investigate the effect of mixup for *decoupled learning* [31], and find that although mixup can improve calibration when used for *representation learning*, it hurts *classifier learning*. Instead, MiSLaS relies on *class-balanced sampling* and *class-aware label smoothing* to promote good calibration during the second stage. Complementary to this approach, we show that *mixup non-trivially improves both calibration and accuracy by a significant margin when combined with logit adjustment*, in a single stage and without the need for data re-sampling.

Ensemble learning. The ensemble of multiple experts, e.g. Mixture of Experts [29, 30] and Product of Experts (PoE) [25], have empirically shown stronger generalization and better calibration [36] over their single-expert counterpart in the balanced setting [35, 55]. These benefits are typically attributed to model diversity, e.g. *making diverse mistakes* [15] or exploring different local minima in the loss landscape [17]. In the long-tailed setting, expert-based methods show promising results to improve the *head-tail trade-off* [5]. These approaches typically promote diversity by increasing the KL divergence between expert predictions [37, 61] or by learning on different groups of classes [5, 54, 63]. However, the former methods are not tailored to address the head bias, whereas the latter present limited scalability under low-shot regimes. Closer to our work, Li *et al*. [38] learn an ensemble of uniform experts with a combination of *balanced softmax* [51], nested knowledge distillation [24] and contrastive learning [21]. Recently, Zhang *et al*. [70] propose to address an unknown label distribution shift under the *transductive learning paradigm*, by learning a set of three skill-diverse experts during training time, and then aggregating them with *self-supervised test-time training*. We instead focus on the traditional setting, where our calibrated ensemble can consistently incorporate prior information about the target distribution if available, or by default, it is provably unbiased for minimizing the balanced error with naive aggregation.

3. Method

In this section, we start by describing the problem formulation in Section 3.1, and give an overview of logit adjustment in Section 3.2. Next, we introduce our Balanced Product of Experts in Section 3.3 and describe how we meet the calibration assumption in Section 3.4.

3.1. Problem formulation

In a multi-class classification problem, we are given an unknown distribution $\mathbb{P}$ over some data space $\mathcal{X} \times \mathcal{Y}$ of instances $\mathcal{X}$ and labels $\mathcal{Y} = \{1, 2, \ldots, C\}$, and we want to find a mapping $f : \mathcal{X} \to \mathbb{R}^C$ that minimizes the misclassification error $\mathbb{P}_{x,y}\left(y \neq \arg\max_{j\in\mathcal{Y}} f_j(x)\right)$, where we denote $f(x) \equiv [f_1(x), ..., f_C(x)]$. In practice, given a sample $\mathcal{D} = \{x_i, y_i\}_{i=1}^N \sim \mathbb{P}$, we intend to minimize the empirical risk [59], $R_\delta(f) = \frac{1}{N}\sum_{i=1}^N \ell_{0/1}(y_i, f(x_i))$, where n_j is the number of samples for class j, $N = \sum_{j=1}^C n_j$ and $\ell_{0/1}$ denotes the per-sample misclassification error, also known as $0/1$-loss. As $\ell_{0/1}$ is not differentiable, we typically optimize a surrogate loss, such as the *softmax cross-entropy* (CE),

$$\ell(y, f(x)) = -\log \frac{e^{f_y(x)}}{\sum_{j\in\mathcal{Y}} e^{f_j(x)}}. \tag{1}$$

However, in the long-tailed setting, the training class distribution $\mathbb{P}^{\text{train}}(y)$ is usually highly skewed and cannot be disentangled from the data likelihood $\mathbb{P}^{\text{train}}(x|y)$ by minimizing the misclassification error (or its surrogate loss) [26]. Following Hong *et al*. [26], hereafter we assume a *label distribution shift*, i.e. although training and test priors may be different, the likelihood shall remain unchanged,

$$\mathbb{P}^{\text{train}}(y) \neq \mathbb{P}^{\text{test}}(y) \qquad \mathbb{P}^{\text{train}}(x|y) = \mathbb{P}^{\text{test}}(x|y). \tag{2}$$

A special case is when all classes are equally relevant, i.e. $\mathbb{P}^{\text{test}}(y) \equiv \mathbb{P}^{\text{bal}}(y) = \frac{1}{C}$, where a more natural metric is the balanced error rate (BER) [3, 7]

$$\text{BER}(f) \equiv \frac{1}{C}\sum_{y\in\mathcal{Y}} \mathbb{P}_{x|y}\left(y \neq \arg\max_{j\in\mathcal{Y}} f_j(x)\right), \tag{3}$$

which is consistent with minimizing the misclassification error under a uniform label distribution.

3.2. Distribution shift by logit adjustment

Under a label distribution shift, we can write the training conditional distribution as $\mathbb{P}^{\text{train}}(y|x) \propto \mathbb{P}(x|y)\mathbb{P}^{\text{train}}(y)$, which motivates the use of logit adjustment to remove the long-tailed data bias [48, 51]

$$\exp\left[f_y(x) - \log \mathbb{P}^{\text{train}}(y)\right] \propto \frac{\mathbb{P}^{\text{train}}(y|x)}{\mathbb{P}^{\text{train}}(y)} \propto \mathbb{P}^{\text{bal}}(y|x), \tag{4}$$

leading to an *adjusted scorer* that is *Fisher-consistent* for minimizing BER [48]. Importantly, to obtain unbiased predictions via logit adjustment, an ideal (unadjusted) scorer must model the training conditional distribution, i.e. $e^{f_y(x)} \propto \mathbb{P}^{\text{train}}(y|x)$ [48]. We observe that, although this condition is unattainable in practice, a weaker but necessary requirement is *perfect calibration* [2] for the training distribution. Hereafter we refer to the aforementioned condition as the ***calibration assumption***. This observation explicitly highlights the importance of confidence calibration to obtain unbiased logit-adjusted models.

Hong *et al*. [26] generalize the logit adjustment to address an arbitrary label distribution shift by observing that we can swap the training prior for a desired test prior,

$$\exp\left[f_y(x) + \log \frac{\mathbb{P}^{\text{test}}(y)}{\mathbb{P}^{\text{train}}(y)}\right] \propto \mathbb{P}^{\text{test}}(y|x). \tag{5}$$

Based on the observation that diverse ensembles lead to stronger generalization [17, 61], we investigate in the following section ***how to learn an ensemble, which is not only diverse but also unbiased, in the sense of Fisher consistency, for a desired target distribution.***

3.3. Balanced Product of Experts

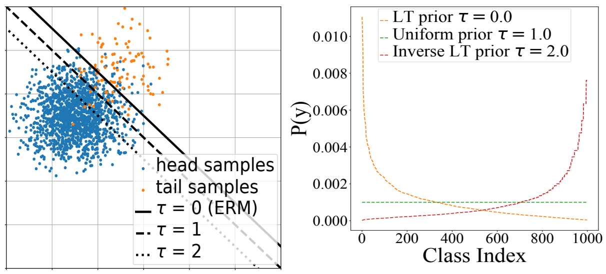
We will now introduce our main theoretical contribution. We start from the logit adjustment formulation and introduce a convenient parameterization of the distribution shift, as we intend to create an ensemble of logit-adjusted experts, biased toward different parts of the label distribution.

Parameterization of target distributions. Let us revisit the parametric logit adjustment formulation by Menon *et al*. [48]. For $\boldsymbol{\tau} \in \mathbb{R}^C$, we observe the following:

$$\begin{aligned}
&\arg\max_y f_y(x) - \boldsymbol{\tau}_y \log \mathbb{P}^{\text{train}}(y) \\
&= \arg\max_y f_y(x) - \log \mathbb{P}^{\text{train}}(y) + \log \mathbb{P}^{\text{train}}(y)^{1-\boldsymbol{\tau}_y}
\end{aligned} \tag{6}$$

$$= \arg\max_y f_y(x) + \log \frac{\mathbb{P}^{\boldsymbol{\lambda}}(y)}{\mathbb{P}^{\text{train}}(y)}, \tag{7}$$

where $\boldsymbol{\lambda}_y \equiv 1 - \boldsymbol{\tau}_y$ and $\mathbb{P}^{\boldsymbol{\lambda}}(y) \equiv \frac{\mathbb{P}^{\text{train}}(y)^{\boldsymbol{\lambda}_y}}{\sum_{j\in\mathcal{Y}} \mathbb{P}^{\text{train}}(j)^{\boldsymbol{\lambda}_j}}$. Note that adjusting the scorer $f_y(x)$ with $\boldsymbol{\tau}_y \log \mathbb{P}^{\text{train}}(y)$ in (6) can be interpreted as performing a distribution shift parameterized by $\boldsymbol{\lambda}$. Particularly, the adjusted model shall accommodate for a target distribution $\mathbb{P}^{\boldsymbol{\lambda}}(y)$. As we intend to incorporate



(a) (b)

Figure 2. (a) By increasing the pairwise class margin, the decision boundary smoothly moves towards the head class, resulting in (b) diverse label distributions parameterized by τ.

a controlled bias during training, we define the *generalized logit adjusted loss* (gLA), which is simply the softmax cross-entropy of the adjusted scorer in (6)

$$\ell_{\boldsymbol{\tau}}(y, f(x)) = -\log \frac{e^{f_y(x)+\boldsymbol{\tau}_y \log \mathbb{P}^{\text{train}}(y)}}{\sum_{j\in\mathcal{Y}} e^{f_j(x)+\boldsymbol{\tau}_j \log \mathbb{P}^{\text{train}}(j)}} \tag{8}$$

$$= \log \left[1 + \sum_{j\neq y} e^{f_j(x)-f_y(x)+\Delta^{\boldsymbol{\tau}}_{yj}} \right], \tag{9}$$

where $\Delta^{\boldsymbol{\tau}}_{yj} = \log \frac{\mathbb{P}^{\text{train}}(j)^{\boldsymbol{\tau}_j}}{\mathbb{P}^{\text{train}}(y)^{\boldsymbol{\tau}_y}}$ defines a pairwise class margin. Intuitively, by increasing the margin, the decision boundary moves away from minority classes and towards majority classes [48], as illustrated in Figure 2(a). For example, by setting $\boldsymbol{\lambda} = \mathbb{1}$ ($\boldsymbol{\tau} = 0$), we obtain the CE loss and a long-tailed distribution as a result. For $\boldsymbol{\lambda} = 0$ ($\boldsymbol{\tau} = \mathbb{1}$) we model the *uniform distribution*, with the so-called *balanced softmax* (BS) [48, 51]. Interestingly, we obtain an *inverse LT distribution* by further increasing the margin, e.g. for $\boldsymbol{\tau} = \mathbb{2}$ ($\boldsymbol{\lambda} = -\mathbb{1}$). Figure 2(b) depicts these scenarios for a long-tailed prior (Pareto distribution). In summary, the proposed loss provides a principled approach for learning specialized expert models and is the building block of the ensemble-based framework to be introduced in the next section.

Ensemble learning. To address the long-tailed recognition problem, we introduce a **Bal**anced **P**roduct **o**f **E**xperts (BalPoE), which combines multiple *logit-adjusted experts* in order to accommodate for a desired test distribution. Let us denote as $\mathbb{P}^{\boldsymbol{\lambda}}(x, y) \equiv \mathbb{P}(x|y)\mathbb{P}^{\boldsymbol{\lambda}}(y)$ the joint distribution associated to $\boldsymbol{\lambda} \in \mathbb{R}^C$.

Let S_λ be a multiset of $\boldsymbol{\lambda}$-vectors describing the parameterization of $|S_\lambda| \geq 1$ experts. Our framework is composed of a set of scorers $\{f^{\boldsymbol{\lambda}}\}_{\boldsymbol{\lambda}\in S_\lambda}$, where each $\boldsymbol{\lambda}$-expert is (ideally) distributed according to $\exp\left[f^{\boldsymbol{\lambda}}_y(x)\right] \propto \mathbb{P}^{\boldsymbol{\lambda}}(x, y)$. The proposed ensemble is then defined as the average of the expert scorers in log space,

$$\overline{p}(x, y) \equiv \exp\left[\overline{f}_y(x)\right] \equiv \exp\left[\frac{1}{|S_\lambda|}\sum_{\boldsymbol{\lambda}\in S_\lambda} f^{\boldsymbol{\lambda}}_y(x)\right]. \tag{10}$$

Now, let us denote as $\overline{\boldsymbol{\lambda}} \equiv \frac{1}{|S_\lambda|}\sum_{\boldsymbol{\lambda}\in S_\lambda} \boldsymbol{\lambda}$ the average of all expert parameterizations. We show, in Theorem 1, that an ensemble consisting of bias-adjusted experts accommodates a marginal distribution $\mathbb{P}^{\overline{\boldsymbol{\lambda}}}(y)$.

Theorem 1 (Distribution of BalPoE)**.** *Let S_λ be a multiset of $\boldsymbol{\lambda}$-vectors describing the parameterization of $|S_\lambda| \geq 1$ experts. Let us assume dual sets of training and target scorer functions, $\{s^{\boldsymbol{\lambda}}\}_{\boldsymbol{\lambda}\in S_\lambda}$ and $\{f^{\boldsymbol{\lambda}}\}_{\boldsymbol{\lambda}\in S_\lambda}$ with $s, f : \mathcal{X} \to \mathbb{R}^C$, respectively, s.t. they are related by*

$$f^{\boldsymbol{\lambda}}_y(x) \equiv s^{\boldsymbol{\lambda}}_y(x) - \log \mathbb{P}^{\text{train}}(y) + \boldsymbol{\lambda}_y \log \mathbb{P}^{\text{train}}(y). \tag{11}$$

*Assume that the **calibration assumption** holds for all training scorers, i.e.* $\exp s^{\boldsymbol{\lambda}}_y(x) \propto \mathbb{P}^{\text{train}}(y|x)$ *for $\boldsymbol{\lambda} \in S_\lambda$. Then, under a label distribution shift in (2), the product of experts as defined in (10) satisfies*

$$\overline{p}(x, y) \propto \mathbb{P}(x|y)\mathbb{P}^{\overline{\boldsymbol{\lambda}}}(y) \equiv \mathbb{P}^{\overline{\boldsymbol{\lambda}}}(x, y). \tag{12}$$

Proof. See supplementary material. □

In other words, **the proposed ensemble attains the average bias of all of its experts**. We can utilize this result to construct an ensemble of diverse experts which is Fisher-consistent for minimizing balanced error. All we need to make sure of is that $\mathbb{P}^{\overline{\boldsymbol{\lambda}}}(y) = \frac{1}{C}$ [48]. A simple constraint is then $\overline{\boldsymbol{\lambda}}_y = 0$ for all $y \in \mathcal{Y}$:

Corollary 1.1 (Fisher-consistency)**.** *If $\overline{\boldsymbol{\lambda}}_y = 0$ for all $y \in \mathcal{Y}$, then the BalPoE scorer $\overline{f}$ is fisher-consistent for minimizing the balanced error.*

Proof. From Theorem 1, we have that for $\overline{\boldsymbol{\lambda}} = 0$,

$$\arg\max_y \overline{f}_y(x) = \arg\max_y \log \mathbb{P}(x|y)\mathbb{P}^{\overline{\boldsymbol{\lambda}}}(y) \tag{13}$$

$$= \arg\max_y \log \mathbb{P}(x|y)\frac{1}{C} \tag{14}$$

$$= \arg\max_y \mathbb{P}(x|y). \tag{15}$$

From (7) in [48], it follows that BalPoE coincides with the *Bayes-optimal rule* for minimizing the balanced error. □

Thus, by defining a set of $\boldsymbol{\lambda}$-experts, such that the average bias parameterization $\overline{\boldsymbol{\lambda}} = 0$, we have shown that the resulting ensemble is unbiased. Note that our framework can accommodate for any known target distribution $\mathbb{P}^{\text{test}}(y)$, by ensuring $\overline{\boldsymbol{\lambda}}_y = \log \frac{\mathbb{P}^{\text{test}}(y)}{\mathbb{P}^{\text{train}}(y)}$.

Training of logit-adjusted experts. We propose to train each $\boldsymbol{\lambda}$-expert by minimizing its respective logit adjusted loss $\ell_{1-\boldsymbol{\lambda}}$. The loss of the overall ensemble is computed as the average of individual expert losses,

$$\ell^{\text{total}}(y, \overline{f}(x)) = \frac{1}{|S_\lambda|}\sum_{\boldsymbol{\lambda}\in S_\lambda} \ell_{1-\boldsymbol{\lambda}}(y, f^{\boldsymbol{\lambda}}(x)). \tag{16}$$

During inference, we average the logits of all experts before softmax normalization to estimate the final prediction.

Based on our theoretical observations, we argue that the calibration assumption needs to be met in order to guarantee Fisher consistency for logit-adjusted models. This is however not the case per default in practice, as can be seen in Figure 1(b) for BS (single-expert BalPoE) and in Table 2 for a three-expert BalPoE trained with ERM. Thus, we need to find an effective way to calibrate our ensemble, a matter we discuss in further detail in the next section.

3.4. Meeting the calibration assumption

Following Theorem 1, we observe that it is desirable for all experts to be well-calibrated for their target distribution in order to obtain an unbiased ensemble. Among calibration methods in the literature, we explore mixup [69] for two main reasons: (1) mixup has shown to improve calibration in the balanced setting [58] and to some extent in the long-tailed setting [71]. (2) mixup does not change the prior class distribution $\mathbb{P}^{\text{train}}(y)$ [66], which is crucial as it would otherwise perturb the necessary bias adjustments.

Mixup samples from a *vicinity distribution* $\{\tilde{x}_i, \tilde{y}_i\}_{i=1}^{N'} \sim \mathbb{P}_\nu$, where $\tilde{x}=\xi x_i + (1-\xi)x_j$ and $\tilde{y}=\xi y_i + (1-\xi)y_j$ are convex combinations of random input-label pairs, $\{(x_i, y_i), (x_j, y_j)\} \sim \mathbb{P}^{\text{train}}_{x,y}$, with $\xi \sim Beta(\alpha, \alpha)$ and $\alpha \in (0, \infty)$. The model is trained by minimizing $R_\nu(f) = \frac{1}{N'} \sum_{i=1}^{N'} \ell(\tilde{y}_i, f(\tilde{x}_i))$, known as *Vicinal Risk Minimization*. Zhong *et al*. [71] found that, even though calibration is improved, mixup does not guarantee an increase in test accuracy, and could even degrade classifier performance. However, we found this not to be the case for logit-adjusted models. We argue that this is due to the fact that mixup does not change the prior class distribution $\mathbb{P}^{\text{train}}(y)$, as proved by Xu *et al*. [66]. While the long-tailed prior is typically unfavorable, and something one strives to adjust, this property is crucial for methods based on logit adjustment, such as BalPoE. We assume $\mathbb{P}^{\text{train}}(y)$ is known, and in practice, estimate it from data. If the marginal distribution were to change, for instance by data resampling, the precomputed biases would no longer be accurate, and the ensemble would become biased. In summary, mixup is compatible with BalPoE as a means for improving calibration. In Section 4.2.2 we show that mixup significantly improves the calibration of our approach, leading to a *Balanced Product of Calibrated Experts*, see the reliability diagram in Figure 1(d). We empirically observe that mixup also non-trivially improves the balanced error, which we believe is attributed to the fact that it contributes to fulfilling the calibration assumption, thus guaranteeing a balanced ensemble.

4. Experiments

We validate our finding of Theorem 1 by measuring test accuracy when varying different aspects of the expert ensemble, including the number of experts, test-time target distributions in terms of λ, and backbone configuration. Furthermore, we measure model calibration of our expert ensemble and investigate how this is affected by mixup. Finally, we compare our method with current state-of-the-art methods on several benchmark datasets.

4.1. Experimental setup

4.1.1 Long-tailed datasets

CIFAR-100-LT. Following previous work [6, 13], a long-tailed version of the balanced CIFAR-100 dataset [32] is created, by discarding samples according to an exponential profile. The dataset contains 100 classes, and the class imbalance is controlled by the imbalance ratio (IR) $\rho = \frac{\max_i n_i}{\min_i n_i}$, i.e. the ratio between the number of instances for the most and least populated classes. We conduct experiments with $\rho \in \{10, 50, 100\}$. For experiments on CIFAR-10-LT [32], see the supplementary material.

ImageNet-LT. Built from ImageNet-2012 [14] with 1K classes, this long-tailed version is obtained by sampling from a Pareto distribution with $\alpha = 6$ [44]. The resulting dataset consists of 115.8K training, 20K validation, and 50K test images. The categories in the training set contain between 5 and 1280 samples, with an imbalanced ratio of $\rho = 256$.

iNaturalist. The iNaturalist 2018 dataset [27] is a large-scale species classification dataset containing 437K images and 8142 classes, with a natural imbalance ratio of $\rho = 500$.

4.1.2 Implementation details

For the experiments on CIFAR, following Cao *et al*. [6], we train a ResNet-32 backbone [22] for 200 epochs with SGD, initial learning rate (LR) of 0.1, momentum rate of 0.9, and a batch size of 128. A multi-step learning rate schedule decreases the LR by a factor of 10 at the 160th and 180th epochs. For large-scale datasets, we follow [61, 70]. For ImageNet-LT, we train ResNet-50 and ResNeXt-50 [64] backbones for 180 epochs with a batch size of 64. For iNaturalist, we train a ResNet-50 for 100 epochs with a batch size of 512. We use a cosine annealing scheduler [45] for ImageNet-LT and iNaturalist, with initial LR of 0.025 and 0.2, respectively. For mixup experiments, we set α to 0.4, 0.3, and 0.2 for CIFAR-100-LT, ImageNet-LT and iNaturalist, respectively, unless stated otherwise. By default, we use 3 experts for BalPoE. We set S_λ to $\{1, 0, -1\}$, $\{1, -0.25, -1.5\}$ and $\{2, 0, -2\}$, for CIFAR-100-LT, ImageNet-LT and iNaturalist, respectively, unless stated otherwise. See the supplementary material for additional details.

In addition, we conduct experiments with the longer training schedule using stronger data augmentation, following Cui *et al*. [12]. All models are trained for 400 epochs, using RandAugment [11] in the case of ImageNet and iNaturalist, and AutoAugment [10] for CIFAR-100-LT. In the latter case, the learning rate is decreased at epochs 320 and 360.

4.1.3 Evaluation protocol

Following Cao *et al*. [6], we report accuracy on a balanced test set. Whenever confidence intervals are shown, they are acquired from five runs. We also report accuracy for three common data regimes, where classes are grouped by their number of samples, namely, many-shot (> 100), medium-shot (20-100), and few-shot (< 20).

In addition, we investigate model calibration by estimating the expected calibration error (ECE), maximum calibration error (MCE), and visualizing reliability diagrams. See the supplementary material for definitions of these metrics.

Finally, for a more realistic benchmark, we follow [26,70] and evaluate our approach under a diverse set of shifted target distributions, which simulate varying levels of class imbalance, from *forward long-tailed* distributions that resemble the training data to extremely different *backward long-tailed* cases. For more details on how these distributions are created, see the supplementary material.

4.2. State-of-the-art comparison

In this section, we compare our method, BalPoE, with state-of-the-art methods, in terms of accuracy and calibration, both on the uniform and non-uniform test distributions.

4.2.1 Comparison under the balanced test distribution

Table 1 compares accuracy under the balanced test distribution on CIFAR-100-LT, ImageNet-LT, and iNaturalist 2018. For ensemble methods, we report the performance for a three-expert model. For the standard setting, we observe significant improvements over previous methods across most benchmarks. For CIFAR-100-LT, we gain +1.5 and +2.2 accuracy points (pts) over the previous state-of-the-art for imbalance ratios 10 and 100, respectively. On ImageNet-LT, we improve by +3.6 and +1.1 for ResNet50 and ResNeXt50 backbones over RIDE and SADE, respectively, which highlights the benefits of BalPoE compared to other ensemble-based competitors that are not fisher-consistent by design. On iNaturalist, we outperform SADE by +2.1 pts and match the previous SOTA, NCL (+0.1 pts), which trains for longer (4x epochs) with strong data augmentation. Lastly, our skill-diverse ensemble considerably outperforms its single-expert counterpart (LA), while also surpassing an *all-uniform ensemble* ($S_\lambda = \{0, 0, 0\}$) on challenging large-scale datasets.

Similar to PaCo [12], BalPoE also benefits from stronger data augmentation and longer training, commonly required by contrastive learning approaches [12,38,72], setting a new state-of-the-art across all benchmarks under evaluation.

Table 1. Test accuracy (%) on CIFAR-100-LT, ImageNet-LT, and iNaturalist 2018 for different imbalance ratios (IR) and backbones (BB). Notation: R32=ResNet32, R50=ResNet50, RX50=ResNeXt50. DA denotes data augmentation. $\star$: reproduced results. **: reproduced with mixup. †: From [66]. ‡: From [70].

	CIFAR-100-LT			ImageNet-LT		iNat
IR →	10	50	100	256	256	500
Method ↓ BB →	R32	R32	R32	R50	RX50	R50
CE*	57.2	43.9	38.8	47.2	48.0	65.2
CB-Focal [13]	58.0	45.3	39.6	-	-	-
LDAM-DRW [6]	58.7	48.0†	42.0	45.8†	-	68.0
BS [51]	59.9†	49.8†	43.9	51.1	-	68.4
LA** [47]	-	-	47.0	-	55.2	69.9
LADE [26]	61.7	50.5	45.4	-	53.0	70.0
MiSLAS [71]	63.2	52.3	47.0	52.7	51.4‡	71.6
RIDE [61]	61.8‡	51.7‡	48.0	54.9	56.4	72.2
DiVE [23]	62.0	51.1	45.4	53.1	-	71.7
SSD [39]	62.3	50.5	46.0	-	56.0	71.5
DRO-LT [53]	63.4	**57.6**	47.3	-	53.5	69.7
ACE [5]	-	51.9	49.6	54.7	56.6	72.9
UniMix+Bayias [66]	61.3	51.1	45.5	48.4	-	69.2
CMO+BS [49]	62.3	51.4	46.6	-	-	70.9
TLC [37]	-	-	49.0	54.6	-	-
SADE [70]	63.6	53.9	49.8	-	58.8	72.9
Uniform BalPoE (ours)	-	-	52.0	-	59.5	74.7
BalPoE (ours)	**65.1**	56.3	**52.0**	**58.5**	**59.8**	**75.0**
Stronger DA						
BCL [72]	64.9	56.6	51.9	56.0	57.1	71.8
BalPoE (ours)	**66.3**	**58.7**	**54.7**	**59.7**	**61.6**	**73.5**
Longer training						
PaCo [12]	64.2	56.0	52.0	57.0	58.2	73.2
CMO+BS [49]	65.3	56.7	51.7	58.0	-	74.0
BCL [72]	-	-	53.9	-	-	-
NCL [38]	-	58.2	54.2	59.5	60.5	74.9
SADE [70]	65.3	57.3	52.2	-	61.2	74.5
BalPoE (ours)	**68.1**	**60.1**	**55.9**	**60.8**	**62.0**	**76.9**

Table 2. Expected calibration error (ECE), maximum calibration error (MCE), and test accuracy (ACC) on CIFAR-100-LT-100. $\star$: reproduced results. †: from [66]. ‡: trained with ERM.

	CIFAR-100-LT-100		
Method ↓	ECE ↓	MCE ↓	ACC ↑
CE*	32.0±0.4	47.3±1.8	38.8±0.6
Bayias [66]	24.3	39.7	43.5
TLC [37]	22.8	-	49.0
BalPoE‡ (ours)	17.6±0.4	28.9±0.9	49.2±0.5
Mixup* [69]	9.6±0.8	15.9±1.5	40.8±0.7
Remix† [9]	33.6	51.0	41.9
UniMix+Bayias [66]	23.0	37.4	45.5
MiSLAS [71]	**4.8**	-	47.0
BalPoE (ours)	**4.9**±1.0	**11.3**±1.6	**52.0**±0.5

4.2.2 Calibration comparison

In this section, we provide a comparison with previous methods for confidence calibration under the LT setting. See Table 2 for ECE, MCE, and accuracy computed over CIFAR-100-LT-100 balanced test set. First, we observe that *Fisher-consistent* approaches trained with ERM, such as Bayias [66] and BalPoE, significantly improve calibra-

Table 3. Test accuracy (%) on multiple target distributions for CIFAR-100-LT-100 and Imagenet-LT (ResNeXt50). *Prior*: test class distribution is used. $*$: Prior implicitly estimated from test data by self-supervised learning. †: results from [70].

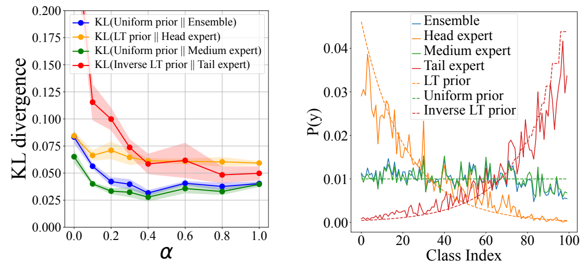
		CIFAR-100-LT-100					Imagenet-LT				
		Fwd-LT		Uni	Bwd-LT		Fwd-LT		Uni	Bwd-LT	
Method	prior	50	5	1	5	50	50	5	1	5	50
Softmax†	✗	63.3	52.5	41.4	30.5	17.5	66.1	56.6	48.0	38.6	27.6
MiSLAS†	✗	58.8	53.0	46.8	40.1	32.1	61.6	56.3	51.4	46.1	39.5
LADE†	✗	56.0	51.0	45.6	40.0	34.0	63.4	57.4	52.3	46.8	40.7
RIDE†	✗	63.0	53.6	48.0	38.1	29.2	**67.6**	61.7	56.3	51.0	44.0
SADE	✗	58.4	53.1	49.4	42.6	35.0	65.5	62.0	58.8	54.7	49.8
BalPoE	✗	**65.1**	**54.8**	**52.0**	**44.6**	**36.1**	**67.6**	**63.3**	**59.8**	**55.7**	**50.8**
LADE†	✓	62.6	52.7	45.6	41.1	41.6	65.8	57.5	52.3	48.8	49.2
SADE	*	65.9	54.8	49.8	44.7	42.4	69.4	63.0	58.8	55.5	53.1
BalPoE	✓	**70.3**	**59.3**	**52.0**	**46.9**	**46.1**	**72.5**	**64.6**	**59.8**	**57.2**	**56.9**

tion over the standard CE. In this setting, our approach achieves lower calibration errors compared to single-expert and multi-expert competitors, namely, Bayias and TLC [37]. Second, we notice that mixup ($\alpha = 0.4$) is surprisingly effective for improving calibration under the LT setting. Although Remix [9] and UniMix [66] improve tail performance by modifying mixup, they tend to sacrifice model calibration, as shown in Table 2. Differently from these methods, we show that the regular mixup effectively regularizes BalPoE, simultaneously boosting its generalization and calibration performance. We hypothesize that logit adjustment might benefit from a smoother decision boundary induced by mixup [60]. Finally, although MiSLAS [71] leverages mixup for classifier learning, it relies on *decoupled learning* [31] and *class-aware label smoothing* for improving calibration, while our approach trains end-to-end without incurring in data re-sampling. Remarkably, BalPoE improves generalization without sacrificing confidence calibration, effectively keeping the best of both worlds.

4.2.3 Results under diverse test distributions

In this section, we compare model generalization with previous methods under multiple test distributions. We report test accuracy for CIFAR-100-LT-100 and ImageNet-LT in Table 3. Without any knowledge of the target distribution, our approach surpasses strong expert-based baselines, such as RIDE and SADE, in most cases, demonstrating the benefits of our Fisher-consistent ensemble.

When knowledge about $\mathbb{P}^{\text{test}}$ is available, our framework addresses the distribution shift by ensuring $\overline{\boldsymbol{\lambda}}_y = \log \frac{\mathbb{P}^{\text{test}}(y)}{\mathbb{P}^{\text{train}}(y)}$. For an efficient evaluation, we first train an unbiased ensemble and then perform post-hoc logit adjustment to incorporate the test bias [26], instead of training a specialized ensemble for each target distribution. For CIFAR-100-LT, BalPoE surpasses SADE by +2.2 pts under the uniform distribution and by nearly +4.0 pts in the case of highly skewed forward and backward distributions. For the challenging ImageNet-LT benchmark, BalPoE outperforms SADE by similar margins, with a +5.1 and +3.8 pts difference for forward50 and backward50, respectively. See the supplementary material for additional results on CIFAR-100-LT, ImageNet-LT, and iNaturalist datasets.



(a) (b)

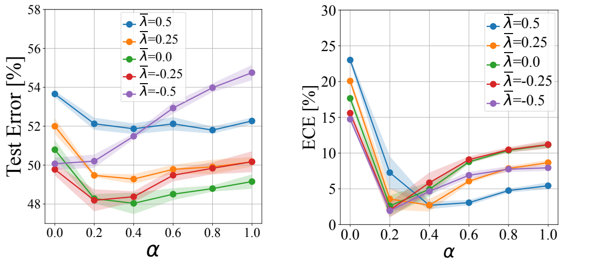
Figure 3. (a) KL divergence of target priors vs expected marginal. (b) Target priors vs expected marginals. Marginals estimated by averaging predictions over CIFAR-100-LT-100 test set.

4.3. Ablation study and further discussion

Effect of mixup. As seen in Figure 1, mixup reduces the calibration error for logit-adjusted models. We further investigate the effect of mixup by comparing the expected prior (estimated from data) to the ideal prior of each expert. We train BalPoE by setting S_λ to $\{1, 0, -1\}$ on CIFAR-100-LT-100, and vary the mixup parameter α. Figure 3(a) shows the KL divergence between expert-specific biases, estimated by averaging the predictive confidence of a given λ-expert over the balanced test data, against the corresponding parametric prior $\mathbb{P}^\lambda(y)$. The prior of the ensemble is compared to the uniform distribution. We find that mixup regularization decreases the divergence for all experts, as well as for the ensemble up to $\alpha = 0.4$, where the divergence attains its minimum for the uniform distribution. This provides further evidence for the fact that well-regularized experts are better calibrated for their respective target distributions. Figure 3(b) shows that estimated marginal distributions are reasonable approximations of the ideal priors for $\alpha = 0.4$.

Do we need a *balanced ensemble*? To verify the validity of Corollary 1.1 in practical settings, we vary the average bias $\overline{\lambda}$ and report the test error in Figure 4(a). As expected, the optimal choice for $\overline{\lambda}$ is near zero, particularly for well-regularized models. For ERM ($\alpha = 0$), the best $\overline{\lambda}$ is -0.25. In this case, the calibration assumption might be violated, thus unbiased predictions cannot be guaranteed.

Effect of the number of experts. We investigate the scalability of our approach in Figure 5, where we plot accuracy over the number of experts. We set different λ configurations equidistantly spaced, with minimum and maximum values at -1.0 and 1.0 respectively, and $\lambda = 0$ for the single-expert case. This ensures that the average is $\overline{\lambda} = 0$. We observe an increase in accuracy as more experts are added to the ensemble. As shown in Figure 1(a), a two-expert BalPoE is a cost-effective trade-off, which surpasses contrastive-learning



(a) (b)

Figure 4. (a) Test error and (b) expected calibration error (ECE) computed over CIFAR-100-LT-100 balanced test set.

approaches, e.g. BCL and PACO, as well as more complex expert approaches, such as NCL and SADE, which rely on knowledge distillation and test-time training, respectively. Interestingly, the performance for even-numbered ensembles interpolates seamlessly, which indicates that a *uniform specialist* is not strictly required. Our approach achieves a peak performance of 57% for seven experts, which represents a relative increase of 11% over its single-expert counterpart. Notably, BalPoE provides tangible benefits for many-, medium-, and especially, few-shot classes.

Connection to recent approaches. Our framework generalizes several recent works based on logit adjustment for single-expert models [26, 48, 52, 66]. LADE [26] introduces a regularizer based on the Donsker-Varadhan representation, which empirically improves single-expert calibration for the uniform distribution, whereas we use mixup regularization to meet the calibration assumption. Within our theoretical framework, we observe that SADE [70], which learns forward, uniform, and backward experts (with different softmax losses), is neither well-calibrated (see Figure 1(c)) nor guaranteed to be Fisher-consistent, as in general, a backward bias (based on flipped class probabilities) does not cancel out with a forward bias, particularly without test-time aggregation. Finally, NCL [38] learns an ensemble of uniform experts with a combination of balanced softmax, contrastive learning, and nested distillation, but does not enforce the calibration assumption required to achieve Fisher consistency.

5. Conclusion

In this paper, we extend the theoretical foundation for logit adjustment to be used for training a balanced product of experts (BalPoE). We show that the ensemble is Fisher-consistent for the balanced error, given that a constraint for the expert-specific biases is fulfilled. We find that model calibration is vital for achieving an unbiased ensemble since the experts need to be weighed against each other in a proper way. This is achieved using mixup. Our BalPoE sets a new state-of-the-art on several long-tailed benchmark datasets.

Limitations. First, we assume $\mathbb{P}^{\text{train}}(x|y) = \mathbb{P}^{\text{test}}(x|y)$, which is a fair assumption but may be violated in practice,

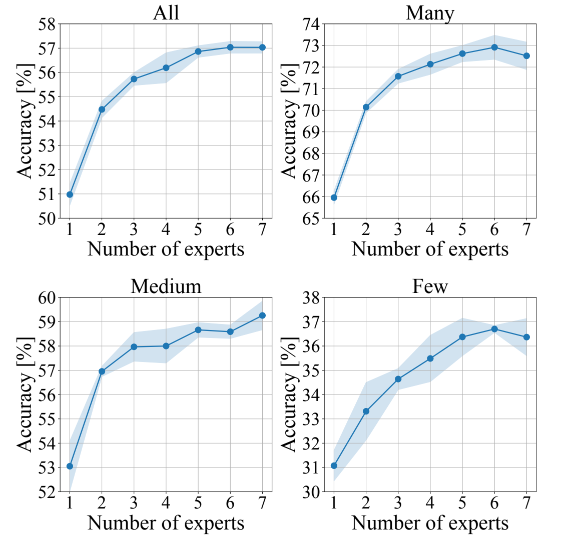


Figure 5. Test accuracy vs number of experts on CIFAR-100-LT-100 for all, many-shot, medium-shot and few-shot classes.

e.g. in autonomous driving applications, where the model might be exposed to out-of-distribution data. Second, the prior $\mathbb{P}^{\text{train}}(y)$ is estimated empirically based on the number of training samples, which can be suboptimal for few-shot classes. To address this issue, considering the effective number of samples [13] could be an interesting venue for future research. Finally, our findings are limited to single-label multi-class classification, and extending our balanced product of experts to multi-label classification and other detection tasks is left to future work.

Societal impact. We believe that the societal impacts of this work are mainly positive. Our proposed method reduces biases caused by label imbalance in the training data, which is important from an algorithmic fairness point of view. Additionally, our model is calibrated, which improves trustworthiness, and usefulness in downstream tasks. However, it cannot account for out-of-distribution data or cases where a view of a tail class appears at test time, which is not captured in the training data. Thus, there is a risk that the model is being overtrusted. Finally, we utilize multiple models, which increases computational cost and electricity consumption, especially for large-scale datasets.

6. Acknowledgements

This work was supported by the Wallenberg Artificial Intelligence, Autonomous Systems and Software Program (WASP), funded by Knut and Alice Wallenberg Foundation. The computational resources were provided by the National Academic Infrastructure for Supercomputing in Sweden (NAISS), partially funded by the Swedish Research Council through grant agreement no. 2022-06725, and by the Berzelius resource, provided by the Knut and Alice Wallenberg Foundation at the National Supercomputer Centre.

References

[1] Rehan Akbani, Stephen Kwek, and Nathalie Japkowicz. Applying support vector machines to imbalanced datasets. In *European conference on machine learning*, pages 39–50. Springer, 2004. 2

[2] Jochen Bröcker. Reliability, sufficiency, and the decomposition of proper scores. *Quarterly Journal of the Royal Meteorological Society: A journal of the atmospheric sciences, applied meteorology and physical oceanography*, 135(643):1512–1519, 2009. 3

[3] Kay Henning Brodersen, Cheng Soon Ong, Klaas Enno Stephan, and Joachim M Buhmann. The balanced accuracy and its posterior distribution. In *2010 20th international conference on pattern recognition*, pages 3121–3124. IEEE, 2010. 3

[4] Mateusz Buda, Atsuto Maki, and Maciej A Mazurowski. A systematic study of the class imbalance problem in convolutional neural networks. *Neural Networks*, 106:249–259, 2018. 2

[5] Jiarui Cai, Yizhou Wang, and Jenq-Neng Hwang. Ace: Ally complementary experts for solving long-tailed recognition in one-shot. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 112–121, 2021. 2, 6

[6] Kaidi Cao, Colin Wei, Adrien Gaidon, Nikos Arechiga, and Tengyu Ma. Learning imbalanced datasets with label-distribution-aware margin loss. *Advances in Neural Information Processing Systems*, 32:1567–1578, 2019. 2, 5, 6

[7] Philip K Chan and Salvatore J Stolfo. Learning with non-uniform class and cost distributions: Effects and a distributed multi-classifier approach. In *In Workshop Notes KDD-98 Workshop on Distributed Data Mining*. Citeseer, 1998. 3

[8] Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16:321–357, 2002. 2

[9] Hsin-Ping Chou, Shih-Chieh Chang, Jia-Yu Pan, Wei Wei, and Da-Cheng Juan. Remix: rebalanced mixup. In *European Conference on Computer Vision*, pages 95–110. Springer, 2020. 2, 6, 7

[10] Ekin D. Cubuk, Barret Zoph, Dandelion Mane, Vijay Vasudevan, and Quoc V. Le. Autoaugment: Learning augmentation strategies from data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019. 6

[11] Ekin D Cubuk, Barret Zoph, Jonathon Shlens, and Quoc V Le. Randaugment: Practical automated data augmentation with a reduced search space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 702–703, 2020. 5

[12] Jiequan Cui, Zhisheng Zhong, Shu Liu, Bei Yu, and Jiaya Jia. Parametric contrastive learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 715–724, 2021. 1, 5, 6

[13] Yin Cui, Menglin Jia, Tsung-Yi Lin, Yang Song, and Serge Belongie. Class-balanced loss based on effective number of samples. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9268–9277, 2019. 2, 5, 6, 8

[14] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 5

[15] Thomas G Dietterich. Ensemble methods in machine learning. In *International workshop on multiple classifier systems*, pages 1–15. Springer, 2000. 2

[16] Chris Drummond, Robert C Holte, et al. C4. 5, class imbalance, and cost sensitivity: why under-sampling beats over-sampling. In *Workshop on learning from imbalanced datasets II*, volume 11, pages 1–8. Citeseer, 2003. 2

[17] Stanislav Fort, Huiyi Hu, and Balaji Lakshminarayanan. Deep ensembles: A loss landscape perspective. *arXiv preprint arXiv:1912.02757*, 2019. 2, 3

[18] Jerzy W Grzymala-Busse, Linda K Goodwin, Witold J Grzymala-Busse, and Xinqun Zheng. An approach to imbalanced data sets based on changing rule strength. In *Rough-neural computing*, pages 543–553. Springer, 2004. 1

[19] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. In *International Conference on Machine Learning*, pages 1321–1330. PMLR, 2017. 2

[20] Hui Han, Wen-Yuan Wang, and Bing-Huan Mao. Borderline-smote: a new over-sampling method in imbalanced data sets learning. In *International conference on intelligent computing*, pages 878–887. Springer, 2005. 2

[21] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9729–9738, 2020. 2

[22] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 5

[23] Yin-Yin He, Jianxin Wu, and Xiu-Shen Wei. Distilling virtual examples for long-tailed recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 235–244, 2021. 6

[24] Geoffrey Hinton, Oriol Vinyals, Jeff Dean, et al. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2(7), 2015. 2

[25] Geoffrey E Hinton. Training products of experts by minimizing contrastive divergence. *Neural computation*, 14(8):1771–1800, 2002. 2

[26] Youngkyu Hong, Seungju Han, Kwanghee Choi, Seokjun Seo, Beomsu Kim, and Buru Chang. Disentangling label distribution for long-tailed visual recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6626–6636, 2021. 2, 3, 6, 7, 8

[27] Grant Van Horn, Oisin Mac Aodha, Yang Song, Yin Cui, Chen Sun, Alex Shepard, Hartwig Adam, Pietro Perona, and Serge Belongie. The inaturalist species classification and detection dataset. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 8769–8778, 7 2017. 5

[28] Chen Huang, Yining Li, Chen Change Loy, and Xiaoou Tang. Learning deep representation for imbalanced classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5375–5384, 2016. 2

[29] Robert A Jacobs, Michael I Jordan, Steven J Nowlan, and Geoffrey E Hinton. Adaptive mixtures of local experts. *Neural computation*, 3(1):79–87, 1991. 2

[30] Michael I Jordan and Robert A Jacobs. Hierarchical mixtures of experts and the em algorithm. *Neural computation*, 6(2):181–214, 1994. 2

[31] Bingyi Kang, Saining Xie, Marcus Rohrbach, Zhicheng Yan, Albert Gordo, Jiashi Feng, and Yannis Kalantidis. Decoupling representation and classifier for long-tailed recognition. In *International Conference on Learning Representations*, 2019. 1, 2, 7

[32] Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, 2009. 5

[33] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25, 2012. 1

[34] Ludmila I Kuncheva. *Combining pattern classifiers: methods and algorithms*. John Wiley & Sons, 2014. 2

[35] Thanard Kurutach, Ignasi Clavera, Yan Duan, Aviv Tamar, and Pieter Abbeel. Model-ensemble trust-region policy optimization. In *International Conference on Learning Representations (ICLR)*, 2018. 2

[36] Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. *Advances in neural information processing systems*, 30, 2017. 2

[37] Bolian Li, Zongbo Han, Haining Li, Huazhu Fu, and Changqing Zhang. Trustworthy long-tailed classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6970–6979, 2022. 2, 6, 7

[38] Jun Li, Zichang Tan, Jun Wan, Zhen Lei, and Guodong Guo. Nested collaborative learning for long-tailed visual recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6949–6958, 2022. 1, 2, 6, 8

[39] Tianhao Li, Limin Wang, and Gangshan Wu. Self supervision to distillation for long-tailed visual recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 630–639, 2021. 6

[40] Zhizhong Li and Derek Hoiem. Improving confidence estimates for unfamiliar examples. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2686–2695, 2020. 2

[41] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017. 2

[42] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European Conference on Computer Vision*, pages 740–755. Springer, 2014. 1

[43] Yi Lin. A note on margin-based loss functions in classification by. 2002. 2

[44] Ziwei Liu, Zhongqi Miao, Xiaohang Zhan, Jiayun Wang, Boqing Gong, and Stella X Yu. Large-scale long-tailed recognition in an open world. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2537–2546, 2019. 5

[45] Ilya Loshchilov and Frank Hutter. SGDR: stochastic gradient descent with warm restarts. In *5th International Conference on Learning Representations, ICLR 2017*. OpenReview.net, 2017. 5

[46] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. A survey on bias and fairness in machine learning. *ACM Computing Surveys (CSUR)*, 54(6):1–35, 2021. 1

[47] Aditya Menon, Harikrishna Narasimhan, Shivani Agarwal, and Sanjay Chawla. On the statistical consistency of algorithms for binary classification under class imbalance. In *International Conference on Machine Learning*, pages 603–611. PMLR, 2013. 6

[48] Aditya Krishna Menon, Sadeep Jayasumana, Ankit Singh Rawat, Himanshu Jain, Andreas Veit, and Sanjiv Kumar. Long-tail learning via logit adjustment. In *International Conference on Learning Representations (ICLR)*, 2021. 2, 3, 4, 8

[49] Seulki Park, Youngkyu Hong, Byeongho Heo, Sangdoo Yun, and Jin Young Choi. The majority can help the minority: Context-rich minority oversampling for long-tailed classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6887–6896, 2022. 2, 6

[50] K Philip and SJS Chan. Toward scalable learning with non-uniform class and cost distributions: A case study in credit card fraud detection. In *Proceeding of the Fourth International Conference on Knowledge Discovery and Data Mining*, pages 164–168, 1998. 1

[51] Jiawei Ren, Cunjun Yu, shunan sheng, Xiao Ma, Haiyu Zhao, Shuai Yi, and hongsheng Li. Balanced meta-softmax for long-tailed visual recognition. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 4175–4186. Curran Associates, Inc., 2020. 1, 2, 3, 4, 6

[52] Mengye Ren, Wenyuan Zeng, Bin Yang, and Raquel Urtasun. Learning to reweight examples for robust deep learning. In *International Conference on Machine Learning*, pages 4334–4343. PMLR, 2018. 8

[53] Dvir Samuel and Gal Chechik. Distributional robustness loss for long-tail learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9495–9504, 2021. 6

[54] Saurabh Sharma, Ning Yu, Mario Fritz, and Bernt Schiele. Long-tailed recognition using class-balanced experts. In *DAGM German Conference on Pattern Recognition*, pages 86–100. Springer, 2020. 2

[55] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with

convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–9, 2015. 2

[56] Jingru Tan, Changbao Wang, Buyu Li, Quanquan Li, Wanli Ouyang, Changqing Yin, and Junjie Yan. Equalization loss for long-tailed object recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11662–11671, 2020. 2

[57] Kaihua Tang, Jianqiang Huang, and Hanwang Zhang. Long-tailed classification by keeping the good and removing the bad momentum causal effect. *Advances in Neural Information Processing Systems*, 33:1513–1524, 2020. 1, 2

[58] Sunil Thulasidasan, Gopinath Chennupati, Jeff A Bilmes, Tanmoy Bhattacharya, and Sarah Michalak. On mixup training: Improved calibration and predictive uncertainty for deep neural networks. *Advances in Neural Information Processing Systems*, 32, 2019. 2, 5

[59] Vladimir Vapnik. Principles of risk minimization for learning theory. *Advances in neural information processing systems*, 4, 1991. 3

[60] Vikas Verma, Alex Lamb, Christopher Beckham, Amir Najafi, Ioannis Mitliagkas, David Lopez-Paz, and Yoshua Bengio. Manifold mixup: Better representations by interpolating hidden states. In *International Conference on Machine Learning*, pages 6438–6447. PMLR, 2019. 2, 7

[61] Xudong Wang, Long Lian, Zhongqi Miao, Ziwei Liu, and Stella Yu. Long-tailed recognition by routing diverse distribution-aware experts. In *International Conference on Learning Representations*, 2020. 2, 3, 5, 6

[62] Yu-Xiong Wang, Deva Ramanan, and Martial Hebert. Learning to model the tail. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pages 7032–7042, 2017. 2

[63] Liuyu Xiang, Guiguang Ding, and Jungong Han. Learning from multiple experts: Self-paced knowledge distillation for long-tailed classification. In *European Conference on Computer Vision*, pages 247–263. Springer, 2020. 2

[64] Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. Aggregated residual transformations for deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1492–1500, 2017. 5

[65] Yu Xie and Charles F Manski. The logit model and response-based samples. *Sociological Methods & Research*, 17(3):283–302, 1989. 2

[66] Zhengzhuo Xu, Zenghao Chai, and Chun Yuan. Towards calibrated model for long-tailed visual recognition from prior perspective. *Advances in Neural Information Processing Systems*, 34, 2021. 2, 5, 6, 7, 8

[67] Han-Jia Ye, Hong-You Chen, De-Chuan Zhan, and Wei-Lun Chao. Identifying and compensating for feature deviation in imbalanced deep learning. *arXiv preprint arXiv:2001.01385*, 2020. 1

[68] Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoon Yoo. Cutmix: Regularization strategy to train strong classifiers with localizable features. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6023–6032, 2019. 2

[69] Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. *International Conference on Learning Representations (ICLR)*, 2018. 2, 5, 6

[70] Yifan Zhang, Bryan Hooi, Hong Lanqing, and Jiashi Feng. Self-supervised aggregation of diverse experts for test-agnostic long-tailed recognition. *Advances in Neural Information Processing Systems*, 35, 2022. 1, 2, 5, 6, 7, 8

[71] Zhisheng Zhong, Jiequan Cui, Shu Liu, and Jiaya Jia. Improving calibration for long-tailed recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16489–16498, 2021. 2, 5, 6, 7

[72] Jianggang Zhu, Zheng Wang, Jingjing Chen, Yi-Ping Phoebe Chen, and Yu-Gang Jiang. Balanced contrastive learning for long-tailed visual recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6908–6917, 2022. 1, 6