Supplementary Material for
Balanced Product of Calibrated Experts for Long-Tailed Recognition

A. Theoretical results

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

$$f^{\boldsymbol{\lambda}}_y(x) \equiv s^{\boldsymbol{\lambda}}_y(x) - \log \mathbb{P}^{\text{train}}(y) + \boldsymbol{\lambda}_y \log \mathbb{P}^{\text{train}}(y). \tag{1}$$

Assume that the **calibration assumption** holds for all training scorers, i.e.

$$\exp s^{\boldsymbol{\lambda}}_y(x) \propto \mathbb{P}^{\text{train}}(y|x) \quad \forall \boldsymbol{\lambda} \in S_\lambda. \tag{2}$$

Then, under a label distribution shift, our product of experts satisfies

$$\overline{p}(x, y) \propto \mathbb{P}(x|y)\mathbb{P}^{\overline{\boldsymbol{\lambda}}}(y) \equiv \mathbb{P}^{\overline{\boldsymbol{\lambda}}}(x, y). \tag{3}$$

Proof of Theorem 1 Given $x \in \mathcal{X}$ and $y \in \mathcal{Y}$, for each $\boldsymbol{\lambda} \in S_\lambda$ and its respective (training) scorer $s^{\boldsymbol{\lambda}}$, we have that

$$\mathbb{P}^{\text{train}}(y|x) = \frac{\exp s^{\boldsymbol{\lambda}}_y(x)}{Z^{\boldsymbol{\lambda}}_x}, \qquad (2) \tag{4}$$

where $Z^{\boldsymbol{\lambda}}_x \in \mathbb{R}$ is an (unknown) normalizing factor. Then, our mean scorer $\overline{f}$ satisfies

$$\begin{aligned}
\overline{f}_y(x) &= \frac{1}{|S_\lambda|} \sum_{\boldsymbol{\lambda} \in S_\lambda} \left[s^{\boldsymbol{\lambda}}_y(x) + (\boldsymbol{\lambda}_y - 1) \log \mathbb{P}^{\text{train}}(y) \right] \qquad (1) & (5)\\
&= \frac{1}{|S_\lambda|} \sum_{\boldsymbol{\lambda} \in S_\lambda} s^{\boldsymbol{\lambda}}_y(x) + \left[\frac{1}{|S_\lambda|} \sum_{\boldsymbol{\lambda} \in S_\lambda} \boldsymbol{\lambda}_y - 1 \right] \log \mathbb{P}^{\text{train}}(y) & (6)\\
&= \frac{1}{|S_\lambda|} \sum_{\boldsymbol{\lambda} \in S_\lambda} \log \left[\mathbb{P}^{\text{train}}(y|x) Z^{\boldsymbol{\lambda}}_x \right] + (\overline{\boldsymbol{\lambda}}_y - 1) \log \mathbb{P}^{\text{train}}(y) \quad (4) & (7)\\
&= \log \frac{\mathbb{P}^{\text{train}}(y|x)}{\mathbb{P}^{\text{train}}(y)} + \log \mathbb{P}^{\text{train}}(y)^{\overline{\boldsymbol{\lambda}}_y} + \frac{1}{|S_\lambda|} \sum_{\boldsymbol{\lambda} \in S_\lambda} \log Z^{\boldsymbol{\lambda}}_x & (8)\\
&= \log \mathbb{P}^{\text{train}}(x|y) + \log \mathbb{P}^{\overline{\boldsymbol{\lambda}}}(y) + \overline{C}^{\boldsymbol{\lambda}}_x \qquad \text{(see definition of } \overline{C}^{\boldsymbol{\lambda}}_x \text{ below)} & (9)\\
&= \log \left[\mathbb{P}(x|y)\mathbb{P}^{\overline{\boldsymbol{\lambda}}}(y) \right] + \overline{C}^{\boldsymbol{\lambda}}_x & (10)\\
&= \log \mathbb{P}^{\overline{\boldsymbol{\lambda}}}(x, y) + \overline{C}^{\boldsymbol{\lambda}}_x, & (11)
\end{aligned}$$

where $\overline{C}^{\boldsymbol{\lambda}}_x = -\log \mathbb{P}^{\text{train}}(x) + \log \left[\sum_{j \in \mathcal{Y}} \mathbb{P}^{\text{train}}(j)^{\overline{\boldsymbol{\lambda}}_j} \right] + \frac{1}{|S_\lambda|} \sum_{\boldsymbol{\lambda} \in S_\lambda} \log Z^{\boldsymbol{\lambda}}_x$ hides terms that are constant w.r.t. y. By re-arranging terms in (11) and applying *softmax*, $\overline{C}^{\boldsymbol{\lambda}}_x$ is cancelled out, obtaining

$$\begin{aligned}
\frac{\mathbb{P}^{\overline{\boldsymbol{\lambda}}}(x, y)}{\sum_{j \in \mathcal{Y}} \mathbb{P}^{\overline{\boldsymbol{\lambda}}}(x, j)} &= \frac{\exp \left[\overline{f}_y(x) - \overline{C}^{\boldsymbol{\lambda}}_x \right]}{\sum_{j \in \mathcal{Y}} \exp \left[\overline{f}_j(x) - \overline{C}^{\boldsymbol{\lambda}}_x \right]} & (12)\\
&= \frac{\exp \overline{f}_y(x)}{\sum_{j \in \mathcal{Y}} \exp \overline{f}_j(x)} & (13)\\
&= \frac{\overline{p}(x, y)}{\sum_{j \in \mathcal{Y}} \overline{p}(x, j)}. & (14)
\end{aligned}$$

From (14) it follows that our BalPoE is proportional to a joint (target) distribution parameterized by $\overline{\boldsymbol{\lambda}}$, i.e. $\overline{p}(x, y) \propto \mathbb{P}^{\overline{\boldsymbol{\lambda}}}(x, y)$.

B. Implementation details

B.1. Dataset summary

In Table 1 we include additional information for the datasets used in this work.

Table 1. Summary of long-tailed datasets.

Dataset	# classes	# samples	IR
CIFAR-LT [3]	10 / 100	60K	{10,50,100}
ImageNet-LT [12]	1K	186K	256
iNaturalist 2018 [9]	8K	437K	500

B.2. Training details

Following previous LT approaches [3, 12], we use cosine classifier, which is defined as $\psi(z, y) = \frac{\kappa w_y^T z}{||w_y||||z||}$, where w_y are learnable weights for class y, z denotes the output of a neural network and κ is a hyperparameter (set to 32). We use weight decay with its hyperparameter set to $5 \cdot 10^{-4}$, $5 \cdot 10^{-4}$ and $2 \cdot 10^{-4}$ for CIFAR-LT, ImageNet-LT, and iNaturalist datasets, respectively. For the CIFAR-LT experiments, we use a warmup period of 5 and 10 epochs for standard and longer training schedules, respectively. We use (up to) four Nvidia A100 40GB GPUs to train our models in an internal cluster. Following [2, 15], our expert architecture comprises an extensive shared backbone and small expert heads (one and two ResNet blocks for large-scale and CIFAR experiments, respectively).

C. Experiments

Here we present additional experiments and an extended analysis to further validate our approach.

C.1. Extended state-of-the-art comparison

In this section, we provide a more detailed comparison with previous state-of-the-art approaches, by reporting test accuracy for many-, medium- and few-shot classes, separately. Tables 2, 3 and 4 present results for CIFAR-100-LT-100, ImageNet-LT and Inaturalist, respectively. For CIFAR-100-LT-100, see Table 2, we observe that our balanced product of calibrated experts significantly improves generalization under few-shot and medium-shot regimes, with only a slight drop in head performance, effectively mitigating the elusive head-tail trade-off. Under the standard setting, we surpass all baselines in medium-, few-shot, and overall performance, while also retaining competitive performance in many-shot classes. As discussed in the main paper, BalPoE can effectively tackle large-scale datasets. We achieve a new state-of-the-art for few-shot, medium-shot, and overall performance for Inaturalist, see Table 4. Finally, for ImageNet-LT we obtain very strong results, on medium- and few-shot classes on-par with current SOTA approaches, while achieving the best head-tail trade-off in overall performance, as shown in Table 3.

Mixup encourages expert specialization We plot the test accuracy for CIFAR-100-LT100 as a function of α in Figure 1. Results are shown for the final ensemble as well as for the different experts separately, and on different data regimes. We observe that mixup promotes expert specialization, especially for the tail expert which becomes a specialist in few-shot classes. Expert regularization boosts the performance of the ensemble, attaining its peak performance at $\alpha = 0.2$-0.4. This observation is consistent with the study of mixup under the balanced setting [17], and previous findings suggesting that large α values may lead to underfitting, due to *manifold intrusion* [7].

Results on CIFAR-10-LT. Table 5 presents results for CIFAR-10-LT, which includes 10 classes, under different imbalance ratios. By default, we train BalPoE with mixup regularization ($\alpha = 0.8$). We observe that our approach promotes a consistent boost in performance under less extreme scenarios, where there are a few classes with arguably enough data. Moreover, we demonstrate that, despite the lower difficulty of this task, BalPoE can still benefit from stronger data augmentation and more extended training, pushing the state-of-the-art on CIFAR-10-LT across several levels of class imbalance.

Table 2. Test accuracy (%) of ResNet-32 trained on CIFAR-100-LT-100 for methods under comparison. ⋆: Our reproduced results. ‡: ACE trained for 400 epochs with regular data augmentation.

Methods	CIFAR-100-LT-100			
	Many	Medium	Few	All
CE*	67.6±1.0	36.7±1.2	7.6±0.6	38.8±0.6
LDAM-DRW [3]	-	-	-	39.6
BS [13]	59.5	45.4	30.7	46.1
LADE [8]	58.7	45.8	29.8	45.6
MiSLAS [19]	60.4	49.6	26.6	47.0
RIDE [15]	68.1	49.2	23.9	48.0
UniMix+Bayias [16]	-	-	-	48.4
DRO-LT [14]	64.7	50.0	23.8	47.3
TLC [10]	**70.9**	47.9	28.1	49.0
SADE [18]	65.4	49.3	29.3	49.8
BalPoE (ours)	67.7±0.3	**54.2**±**0.9**	**31.0**±**0.6**	**52.0**±**0.5**
Longer training				
ACE‡ [2]	66.1	55.7	23.5	49.4
PaCo [5]	-	-	-	52.0
BCL [21]	69.7	53.8	35.5	53.9
NCL [11]	-	-	-	54.2
SADE [18]	-	-	-	52.2
BalPoE (ours)	**71.4**±**0.6**	**58.0**±**0.7**	**35.4**±**0.4**	**55.9**±**0.4**

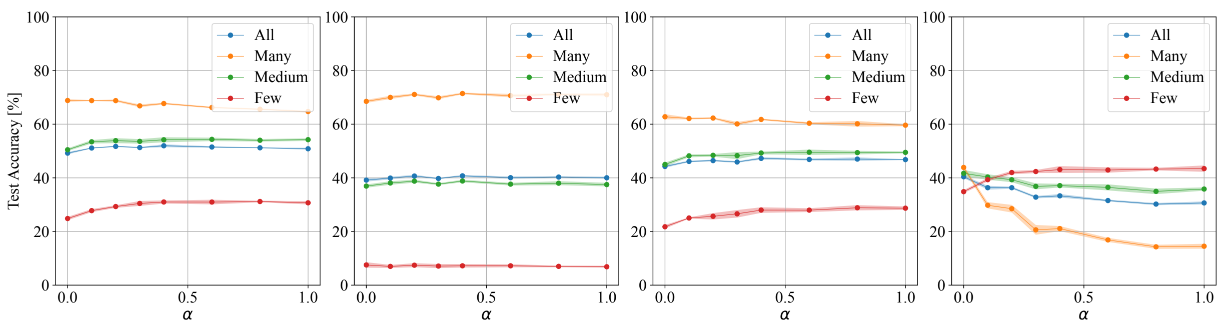
Table 3. Test accuracy (%) of ResNet-50 / ResNeXt-50 trained on ImageNet-LT for methods under comparison. ⋆: Our reproduced results.

Methods	ImageNet-LT							
	ResNet50				ResNeXt50			
	Many	Medium	Few	All	Many	Medium	Few	All
CE*	66.5	40.5	15.9	47.2	68.1	41.5	14.0	48.0
BS [13]	-	-	-	-	64.1	48.2	33.4	52.3
LADE [8]	-	-	-	-	65.1	48.9	33.4	53.0
MiSLAS [19]	61.7	51.3	35.8	52.7	-	-	-	-
RIDE [15]	66.2	51.7	34.9	54.9	67.6	53.5	35.9	56.4
ACE [2]	-	-	-	54.7	-	-	-	56.6
TLC [10]	**69.3**	**56.7**	37.9	54.6	-	-	-	-
SADE [18]	-	-	-	-	66.5	57.0	43.5	58.8
BalPoE (ours)	66.0	**56.7**	**43.6**	**58.5**	**68.2**	**57.2**	**44.9**	**59.8**
Longer training								
PaCo [5]	65.0	55.7	38.2	57.0	67.5	56.9	36.7	58.2
NCL [11]	-	-	-	59.5	-	-	-	60.5
SADE [18]	-	-	-	-	67.3	**60.4**	**46.4**	61.2
BalPoE (ours)	67.8	**59.2**	**46.5**	**60.8**	**70.8**	59.5	**46.4**	**62.0**

C.2. Extended calibration comparison

Table 4. Test accuracy (%) of ResNet-50 trained on Inaturalist-2018 for methods under comparison. $\star$: Our reproduced results.

	Inaturalist			
Methods	Many	Medium	Few	All
CE*	**76.4**	66.8	60.1	65.2
LDAM-DRW [3]	-	-	-	68.0
BS [13]	70.9	70.7	70.4	70.6
LADE [8]	68.9	68.7	70.2	69.3
MiSLAS [19]	73.2	72.4	70.4	71.6
RIDE [15]	70.2	72.2	72.7	72.2
ACE [2]	-	-	-	72.9
SADE [18]	74.5	72.5	73.0	72.9
BalPoE (ours)	73.2	**75.5**	**74.7**	**75.0**
Longer training				
PaCo [5]	70.3	73.2	73.6	73.2
NCL [11]	72.7	75.6	74.5	74.9
SADE [18]	**75.5**	73.7	75.1	74.5
BalPoE (ours)	75.0	**77.4**	**76.9**	**76.9**



(a) BalPoE (b) Head expert (c) Medium expert (d) Tail expert

Figure 1. Test accuracy on CIFAR-100-LT with IR=100, as a function of the mixup parameter α, for (a) BalPoE with three experts, (b) head expert, (c) medium expert, and (d) tail expert.

Definition of calibration. Intuitively, calibration is the measure of how well the model confidence reflects the true probability, i.e. when the model predicts a class with 90% confidence, it should be the correct class in 90% of the cases on average. Formally, a model h is said to be *calibrated* [1] if

$$\mathbb{P}(Y = y | h(X) = \boldsymbol{p}) = \boldsymbol{p}_y \qquad \forall \boldsymbol{p} \in \Delta, \tag{15}$$

where $\Delta = \{p \in [0,1]^C | \sum_{y \in \mathcal{Y}} p_y = 1\}$ is a (C-1)-dimensional simplex. A strictly weaker, but more useful, condition is *argmax calibration* [6], which requires

$$\mathbb{P}(Y = \arg\max h(X) | \max h(X) = p) = p \qquad \forall p \in [0,1]. \tag{16}$$

In practice, we empirically estimate the disagreement between the two sides of (16) over a discrete set of samples. Given a dataset $\mathcal{D} = \{x_i, y_i\}_{i=1}^N$, denote $\hat{p}_i$ the predicted confidence of sample x_i. [6] propose to group predictions into M discrete intervals, and then calculate accuracy and confidence over the respective batch of samples. Let B_m denote the batch of indices in the m interval, we define the average accuracy of B_m as $acc(B_m) = \frac{1}{|B_m|} \sum_{i \in B_m} \mathbf{1}(\hat{y}_i = y_i)$. Similarly, the average confidence of B_m is defined as $conf(B_m) = \frac{1}{|B_m|} \sum_{i \in B_m} \hat{p}_i$. We estimate the Expected Calibration Error (ECE) as

Table 5. Test accuracy (%) of ResNet32 on CIFAR-10-LT for different imbalance ratios (IR). $\star$: Our reproduced results. †: From [16]. ‡: From [18]. §: From [20].

Method ↓	IR →	CIFAR-10-LT 10	50	100
CE*		87.2±0.3	77.3±0.4	71.3±0.9
LDAM-DRW [3]		88.2	81.8†	77.1
BS [13]		**90.9**±0.4	-	83.1±0.4
MiSLAS [19]		90.0	<u>85.7</u>	82.1
RIDE‡ [15]		89.7	-	81.6
ACE [2]		-	84.3	81.2
UniMix+Bayias [16]		89.7	84.3	82.7
TLC [10]		-	-	80.3
SADE [18]		<u>90.8</u>	-	<u>83.8</u>
BalPoE (ours)		90.2±0.2	**86.2**±0.2	**84.2**±0.3
Longer training				
NCL [11]		-	87.3	85.5
BalPoE (ours)		**91.9**±0.1	**88.5**±0.2	**86.8**±0.2

Table 6. Expected calibration error (ECE), maximum calibration error (MCE), and test accuracy (ACC) on CIFAR-10-LT-100. $\star$: Our reproduced results, where mixup is trained with $\alpha = 0.8$. †: from [16]. ‡: our approach trained with ERM.

Method ↓	CIFAR-10-LT-100 ECE ↓	MCE ↓	ACC ↑
CE*	19.1±0.9	33.9±2.0	71.3±0.9
Bayias [16]	**11.0**	**23.7**	78.7
TLC [10]	13.1	-	<u>80.3</u>
BalPoE‡ (ours)	**11.0**±0.3	<u>27.7</u>±2.7	**80.5**±0.3
Mixup* [17]	**3.7**±0.4	**12.9**±4.9	72.9±0.7
Remix† [4]	15.4	28.0	75.4
MiSLAS [19]	**3.7**	-	82.1
UniMix+Bayias [16]	10.2	25.5	<u>82.7</u>
BalPoE (ours)	<u>6.3</u>±0.7	<u>15.8</u>±4.7	**84.2**±0.3

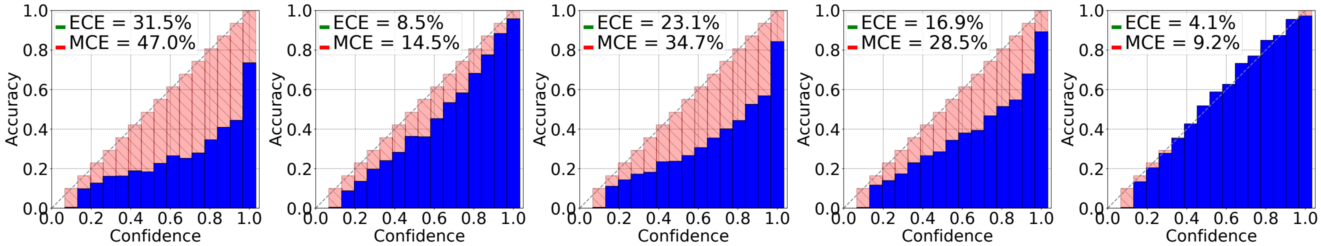
a weighted average of the batch's differences between accuracy and confidence, i.e.

$$ECE = \sum_{m=1}^{M} \frac{|B_m|}{n} |acc(B_m) - conf(B_m)|, \tag{17}$$

where n denotes the number of samples in each equally-spaced interval. Analogously, the Maximum Calibration Error (MCE) describes the maximum difference between accuracy and confidence, i.e.

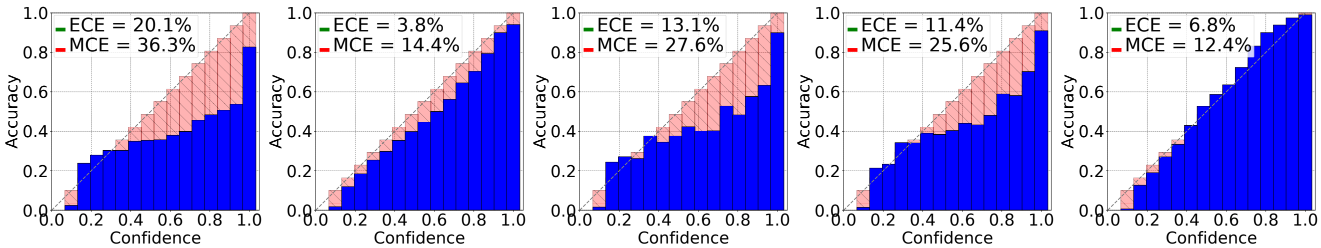
$$MCE = \max_{m \in \{1,..,M\}} |acc(B_m) - conf(B_m)|. \tag{18}$$

Extended discussion. We present reliability diagrams in Figure 2 and Figure 3 for CIFAR-100-LT-100 and CIFAR-10-LT-100, respectively, where we plot the accuracy as a function of the model confidence. Ideally, for samples where the confidence is C, the rate at which the prediction is correct should be the same, namely C. This is highlighted by the diagonal line in the diagram, which corresponds to a perfectly calibrated model. For CIFAR-100-LT-100, the ECE for a single model trained with ERM is 31.5%, which is reduced to 23.1% for BS (equivalent to $\lambda = 0$), and further reduced to 16.9% with a BalPoE of 3 experts ($\lambda = \{1, 0, -1\}$). Remarkably, mixup can further improve the calibration of our approach, leading to an ECE of 4.1%. We observe similar gains for CIFAR-10-LT-100 in terms of calibration, see Figure 3, and generalization performance, as shown in Table 6. We conclude that meeting the calibration assumption is vital for our logit-adjusted expert framework, which we argue explains the large performance gains obtained by using mixup.



(a) CE (b) Mixup (c) BS (d) Uncal. BalPoE (e) BalPoE

Figure 2. Reliability plots for (a) CE, (b) mixup, (c) BS, (d) uncalibrated BalPoE (trained with ERM) and (d) BalPoE (trained with mixup). Computed over **CIFAR-100-LT-100** test set.



(a) CE (b) Mixup (c) BS (d) Uncal. BalPoE (e) BalPoE

Figure 3. Reliability plots for (a) CE, (b) mixup, (c) BS, (d) uncalibrated BalPoE (trained with ERM), and (e) BalPoE (trained with mixup). Computed over **CIFAR-10-LT-100** test set.

C.3. Extended comparison under diverse test distributions

Definition of shifted long-tailed datasets. Following [8, 18], we evaluate our approach under various test class distributions with different imbalance ratios, in order to simulate the diversity of real-world situations. We group these datasets into forward long-tailed distributions, the uniform distribution, and backward long-tailed distributions. For forward distributions, the classes are sorted in decreasing order according to the number of training samples, whereas for backward distributions the class order is flipped. See [8, 18] for a comprehensive description of these benchmarks.

Extended discussion. In Tables 7, 8 and 9 we present additional results of multiple shifted target distributions for CIFAR-100-LT with a training imbalance ratio of 100, 50 and 10, respectively. Across different distributions, our approach provides a significantly better *head-tail trade-off* than other expert-based frameworks, outperforming SADE and RIDE by notable margins at forward and backward scenarios, respectively. Remarkably, the benefits of our unbiased ensemble can also be appreciated as more training data becomes available, particularly for IR=50 and IR=10. Our extensive evaluation further corroborates our early hypothesis: *an ensemble of well-calibrated experts can be a more robust long-tailed classifier than single-expert (often uncalibrated) logit-adjusted approaches*, such as BS and LADE. Tables 10 and 11 show additional results for ImageNet-LT and iNaturalist datasets, respectively, where we observe the effectiveness of our framework to tackle LT recognition under challenging large-scale datasets.

Table 7. Test accuracy (%) on multiple test distributions for model trained on CIFAR-100-LT-100. †: results from [18]. *Prior*: test class distribution is used. $*$: Prior implicitly estimated from test data by self-supervised learning.

		CIFAR-100-LT-100										
		Forward-LT					Unif.	Backward-LT				
Method	prior ↓ IR →	50	25	10	5	2	1	2	5	10	25	50
Softmax†	✗	63.3	62.0	56.2	52.5	46.4	41.4	36.5	30.5	25.8	21.7	17.5
BS†	✗	57.8	55.5	54.2	52.0	48.7	46.1	43.6	40.8	38.4	36.3	33.7
MiSLAS†	✗	58.8	57.2	55.2	53.0	49.6	46.8	43.6	40.1	37.7	33.9	32.1
LADE†	✗	56.0	55.5	52.8	51.0	48.0	45.6	43.2	40.0	38.3	35.5	34.0
RIDE†	✗	63.0	59.9	57.0	53.6	49.4	48.0	42.5	38.1	35.4	31.6	29.2
SADE	✗	58.4	57.0	54.4	53.1	50.1	49.4	45.2	42.6	39.7	36.7	35.0
BalPoE	✗	**65.1**	**63.1**	**60.8**	**58.4**	**54.8**	**52.0**	**48.6**	**44.6**	**41.8**	**38.0**	**36.1**
LADE†	✓	62.6	60.2	55.6	52.7	48.2	45.6	43.8	41.1	41.5	40.7	41.6
SADE	*	65.9	62.5	58.3	54.8	51.1	49.8	46.2	44.7	43.9	42.5	42.4
BalPoE	✓	**70.3**	**66.8**	**62.7**	**59.3**	**54.8**	**52.0**	**49.2**	**46.9**	**46.2**	**45.4**	**46.1**

Table 8. Test accuracy (%) on multiple test distributions for model trained on CIFAR-100-LT-50. †: results from [18]. *Prior*: test class distribution is used. $*$: Prior implicitly estimated from test data by self-supervised learning.

		CIFAR-100-LT-50										
		Forward-LT					Unif.	Backward-LT				
Method	prior ↓ IR →	50	25	10	5	2	1	2	5	10	25	50
Softmax†	✗	64.8	62.7	58.5	55.0	49.9	45.6	40.9	36.2	32.1	26.6	24.6
BS†	✗	61.6	60.2	58.4	55.9	53.7	50.9	48.5	45.7	43.9	42.5	40.6
MiSLAS†	✗	60.1	58.9	57.7	56.2	53.7	51.5	48.7	46.5	44.3	41.8	40.2
LADE†	✗	61.3	60.2	56.9	54.3	52.3	50.1	47.8	45.7	44.0	41.8	40.5
RIDE†	✗	62.2	61.0	58.8	56.4	52.9	51.7	47.1	44.0	41.4	38.7	37.1
SADE	✗	59.5	58.6	56.4	54.8	53.2	53.8	50.1	48.2	46.1	44.4	43.6
BalPoE	✗	**66.5**	**64.8**	**62.8**	**60.9**	**58.3**	**56.3**	**53.8**	**51.0**	**48.9**	**46.6**	**45.3**
LADE†	✓	65.9	62.1	58.8	56.0	52.3	50.1	48.3	45.5	46.5	46.8	47.8
SADE	*	67.2	64.5	61.2	58.6	55.4	53.9	51.9	50.9	51.0	51.7	52.8
BalPoE	✓	**71.1**	**68.3**	**64.8**	**61.8**	**58.2**	**56.3**	**54.4**	**53.4**	**53.4**	**53.8**	**55.4**

Table 9. Test accuracy (%) on multiple test distributions for model trained on CIFAR-100-LT-10. †: results from [18]. *Prior*: test class distribution is used. $*$: Prior implicitly estimated from test data by self-supervised learning.

		CIFAR-100-LT-10										
		Forward-LT					Unif.	Backward-LT				
Method	prior ↓ IR →	50	25	10	5	2	1	2	5	10	25	50
Softmax†	✗	**72.0**	**69.6**	66.4	65.0	61.2	59.1	56.3	53.5	50.5	48.7	46.5
BS†	✗	65.9	64.9	64.1	63.4	61.8	61.0	60.0	58.2	57.5	56.2	55.1
MiSLAS†	✗	67.0	66.1	65.5	64.4	63.2	62.5	61.2	60.4	59.3	58.5	57.7
LADE†	✗	67.5	65.8	65.8	64.4	62.7	61.6	60.5	58.8	58.3	57.4	57.7
RIDE†	✗	67.1	65.3	63.6	62.1	60.9	61.8	58.4	56.8	55.3	54.9	53.4
SADE	✗	66.3	64.5	64.1	62.7	61.6	63.6	60.2	59.7	59.8	58.7	58.6
BalPoE	✗	69.1	68.2	**67.4**	**66.8**	**65.7**	**65.1**	**63.8**	**63.0**	**62.3**	**61.8**	**61.3**
LADE†	✓	71.2	69.3	67.1	64.6	62.4	61.6	60.4	61.4	61.5	62.7	64.8
SADE	*	71.2	69.4	67.6	66.3	64.4	63.6	62.9	62.4	61.7	62.1	63.0
BalPoE	✓	**74.9**	**72.4**	**70.0**	**68.1**	**66.0**	**65.1**	**64.1**	**64.3**	**65.0**	**66.3**	**67.8**

Table 10. Test accuracy (%) on multiple test distributions for ResNeXt50 trained on Imagenet-LT. †: results from [18]. *Prior*: test class distribution is used. $*$: Prior implicitly estimated from test data by self-supervised learning.

		Imagenet-LT										
		Forward-LT					Unif.	Backward-LT				
Method	prior $\downarrow$ IR $\rightarrow$	50	25	10	5	2	1	2	5	10	25	50
Softmax†	✗	66.1	63.8	60.3	56.6	52.0	48.0	43.9	38.6	34.9	30.9	27.6
BS†	✗	63.2	61.9	59.5	57.2	54.4	52.3	50.0	47.0	45.0	42.3	40.8
MiSLAS†	✗	61.6	60.4	58.0	56.3	53.7	51.4	49.2	46.1	44.0	41.5	39.5
LADE†	✗	63.4	62.1	59.9	57.4	54.6	52.3	49.9	46.8	44.9	42.7	40.7
RIDE†	✗	**67.6**	**66.3**	64.0	61.7	58.9	56.3	54.0	51.0	48.7	46.2	44.0
SADE	✗	65.5	64.4	63.6	62.0	60.0	58.8	56.8	54.7	53.1	51.1	49.8
BalPoE	✗	**67.6**	**66.3**	**65.2**	**63.3**	**61.5**	**59.8**	**58.1**	**55.7**	**54.3**	**52.2**	**50.8**
LADE†	✓	65.8	63.8	60.6	57.5	54.5	52.3	50.4	48.8	48.6	49.0	49.2
SADE	$*$	69.4	67.4	65.4	63.0	60.6	58.8	57.1	55.5	54.5	53.7	53.1
BalPoE	✓	**72.5**	**70.2**	**67.3**	**64.6**	**61.8**	**59.8**	**58.3**	**57.2**	**56.6**	**56.6**	**56.9**

Table 11. Test accuracy (%) on multiple test distributions for ResNet50 trained on iNaturalist-2018. †: results from [18]. *Prior*: test class distribution is used. $*$: Prior implicitly estimated from test data by self-supervised learning.

		Inaturalist				
		Forward-LT		Unif.	Backward-LT	
Method	prior $\downarrow$ IR $\rightarrow$	3	2	1	2	3
Softmax†	✗	65.4	65.5	64.7	64.0	63.4
BS†	✗	70.3	70.5	70.6	70.6	70.8
MiSLAS†	✗	70.8	70.8	70.7	70.7	70.2
LADE†	✗	68.4	69.0	69.3	69.6	69.5
RIDE†	✗	71.5	71.9	71.8	71.9	71.8
SADE	✗	-	72.4	72.9	73.1	-
BalPoE	✗	**74.3**	**75.0**	**75.0**	**75.1**	**74.7**
LADE†	✓	-	69.1	69.3	70.2	-
SADE	$*$	72.3	72.5	72.9	73.5	73.3
BalPoE	✓	**74.7**	**75.4**	**75.0**	**75.6**	**75.3**

References

[1] 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. 4

[2] 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, 3, 4, 5

[3] 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, 3, 4, 5

[4] 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. 5

[5] 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. 3, 4

[6] 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. 4

[7] Hongyu Guo, Yongyi Mao, and Richong Zhang. Mixup as locally linear out-of-manifold regularization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 3714–3722, 2019. 2

[8] 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. 3, 4, 6

[9] 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. 2

[10] 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. 3, 5

[11] 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. 3, 4, 5

[12] 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. 2

[13] 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. 3, 4, 5

[14] 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. 3

[15] 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, 4, 5

[16] 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. 3, 5

[17] 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

[18] 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. 3, 4, 5, 6, 7, 8

[19] 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. 3, 4, 5

[20] Boyan Zhou, Quan Cui, Xiu-Shen Wei, and Zhao-Min Chen. Bbn: Bilateral-branch network with cumulative learning for long-tailed visual recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9719–9728, 2020. 5

[21] 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. 3