JUNE 18-22, 2023 CVPR VANCOUVER, CANADA

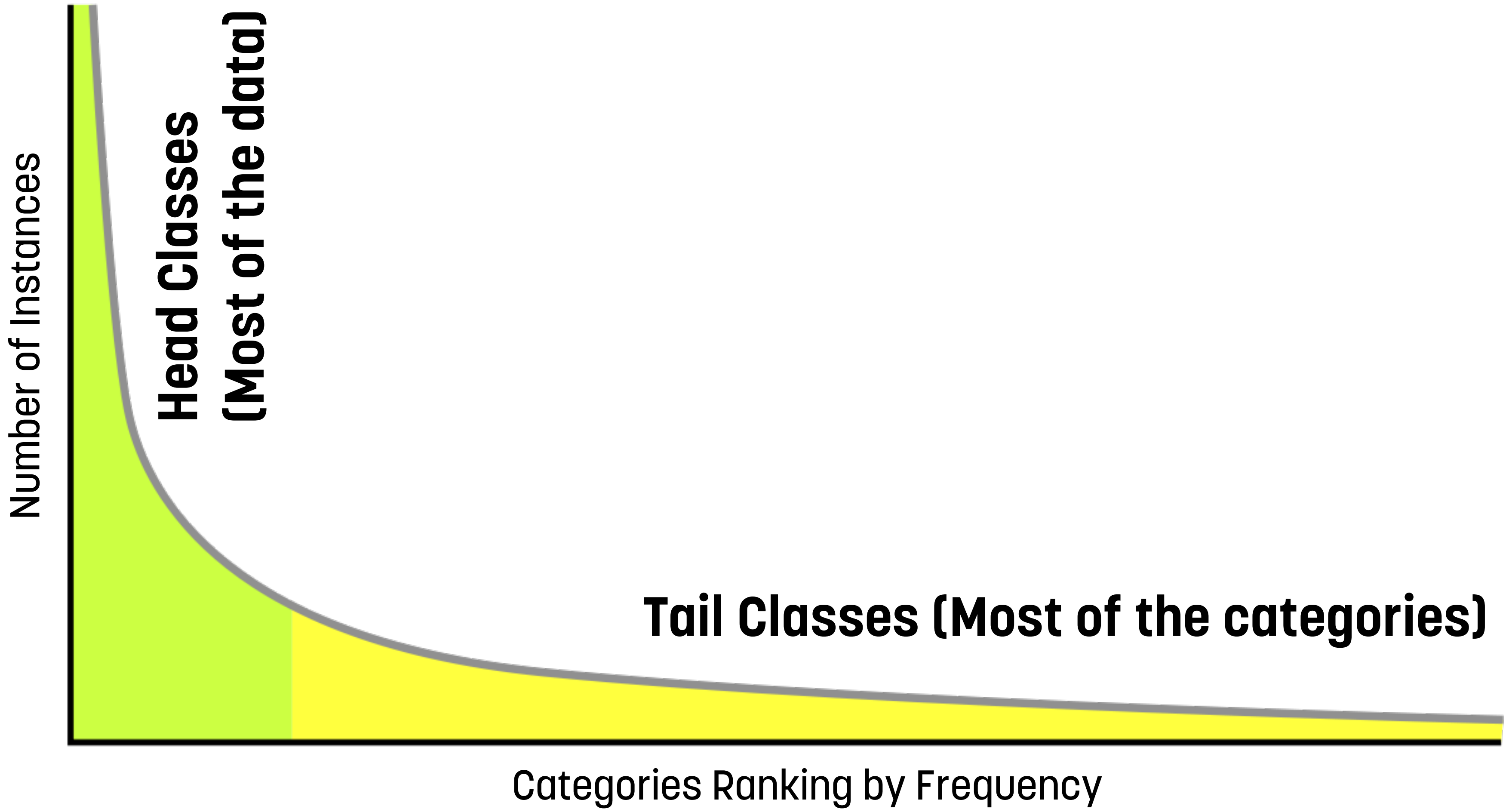
Balanced Product of Calibrated Experts for Long-Tailed Recognition

Emanuel Sanchez Aimar[1], Arvi Jonnarth[1,2], Michael Felsberg[1], Marco Kuhlmann[1]

[1]Linköping University, Sweden [2]Husqvarna Group, Sweden

CVPR 2023 THU-AM-331

What is long-tailed recognition?



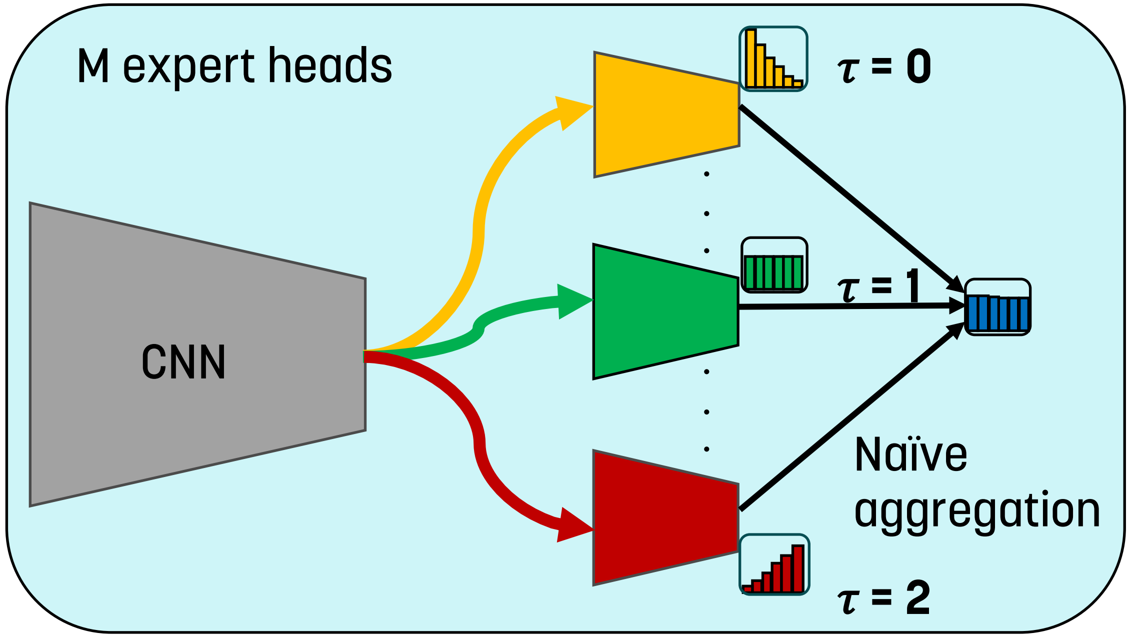
Why is it challenging?

- **Label distribution shift**: training and test distributions tend to be different (long-tailed vs balanced)
- Leads to **uncalibrated models**

Balanced Product of Calibrated Experts (BalPoE)

We propose a simple ensemble-based framework for LT recognition

- Diverse
- Unbiased
- Well-calibrated

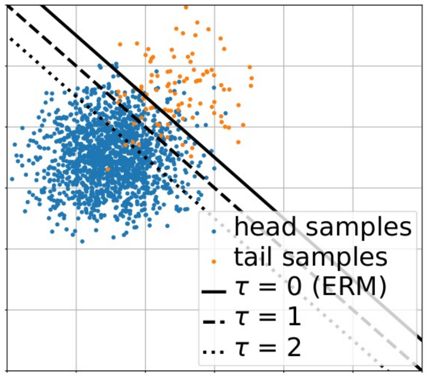


How to learn multiple experts while ensuring the desired properties?

Expert learning

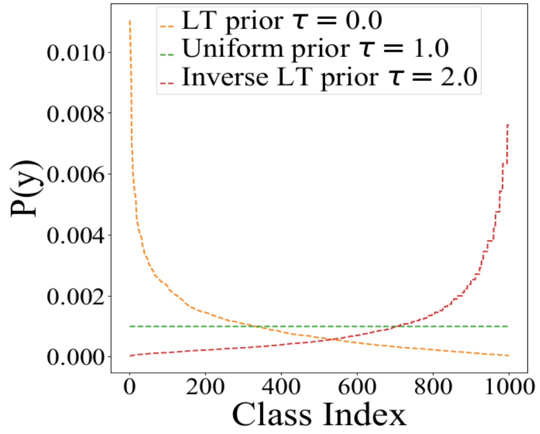
- Logit-adjusted loss for $\boldsymbol{\lambda}$-expert with $\boldsymbol{\lambda}_y \equiv 1 - \boldsymbol{\tau}_y$

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



- Overall ensemble loss:

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



Expert aggregation

- Naïve fusion with uniform weights:

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

- BalPoE attains the average bias of all its experts,
  - Controlled by $\overline{\boldsymbol{\lambda}} \equiv \frac{1}{|S_\lambda|}\sum_{\boldsymbol{\lambda}\in S_\lambda} \boldsymbol{\lambda}$

- **Simple constraint**: $\overline{\lambda}$ is zero => unbiased ensemble

JUNE 18-22, 2023 CVPR VANCOUVER, CANADA

Revisiting our assumptions

- Logit adjustment assumes a **bayes-optimal classifier** for the training distribution (Menon'20):

$$e^{f_y(x)} \propto \mathbb{P}^{\text{train}}(y|x)$$

- A weaker but necessary requirement is **perfect calibration** (Bröcker'09)

- We leverage Mixup (Zhang'18) to meet the calibration assumption

Menon et al. "Long-tail learning via logit adjustment." (ICLR 2020)
Bröcker et al. Reliability, sufficiency, and the decomposition of proper scores. (Quarterly Journal of the Royal Meteorological Society 2009).
Zhang et al. Mixup: Beyond empirical risk minimization. (ICLR 2018)

Mixup promotes expert calibration

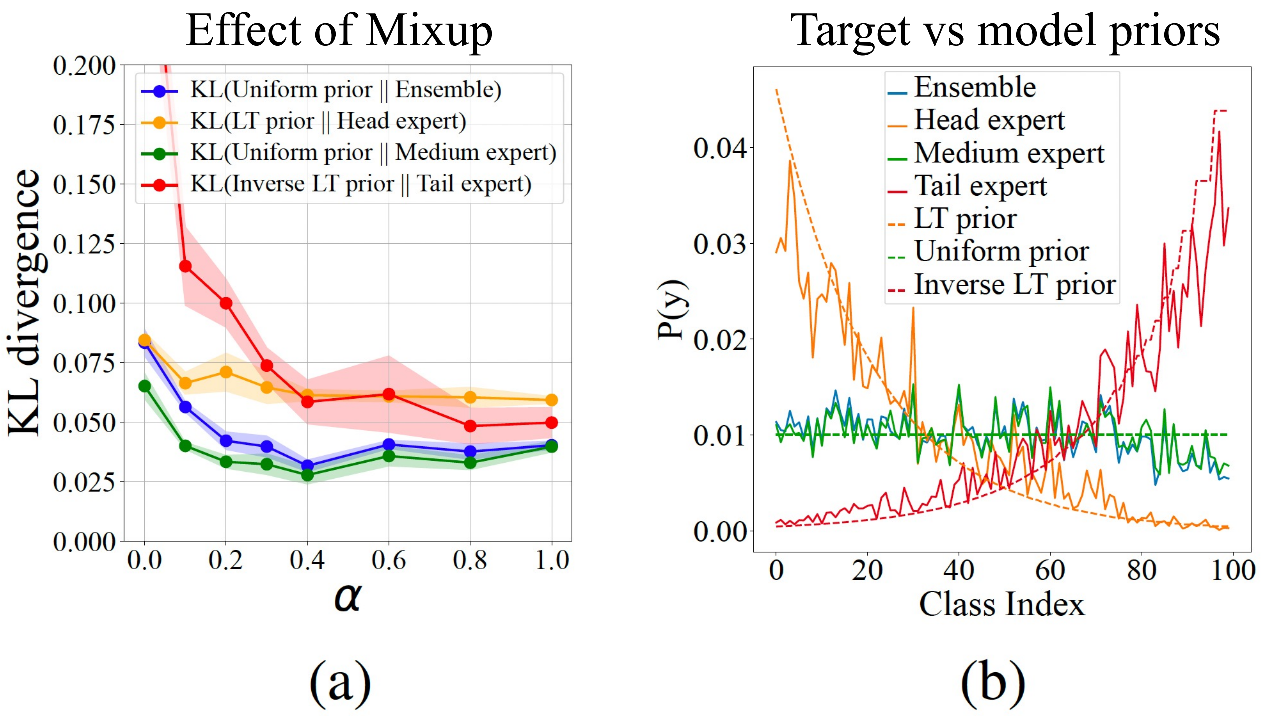


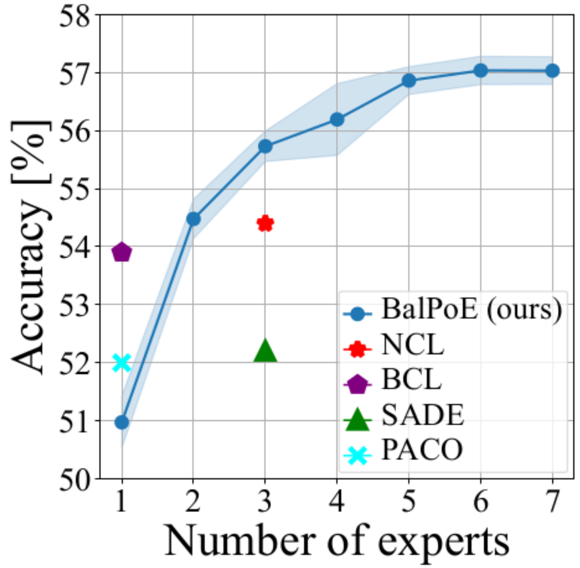
Figure 3. (a) KL divergence of target priors vs expected marginal. (b) Target priors vs expected marginals. Marginals estimated by averaging predictions over CIFAR-100-LT-100 test set.

Experiments

SOTA accuracy under the balanced test set

		CIFAR-100-LT			ImageNet-LT		iNat
	IR →	10	50	100	256	256	500
Method ↓	BB →	R32	R32	R32	R50	RX50	R50
Stronger DA							
BCL [71]	(CVPR'22)	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]	(ICCV'21)	64.2	56.0	52.0	57.0	58.2	73.2
CMO+BS [48]	(CVPR'22)	<u>65.3</u>	56.7	51.7	58.0	-	74.0
BCL [71]	(CVPR'22)	-	-	53.9	-	-	-
NCL [38]	(CVPR'22)	-	<u>58.2</u>	<u>54.2</u>	<u>59.5</u>	60.5	<u>74.9</u>
SADE [69]	(NeurIPS'22)	<u>65.3</u>	57.3	52.2	-	<u>61.2</u>	74.5
BalPoE (ours)		**68.1**	**60.1**	**55.9**	**60.8**	**62.0**	**76.9**

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. ⋆: reproduced results. **: reproduced with mixup. †: From [66]. ‡: From [70].



Improvements across the board

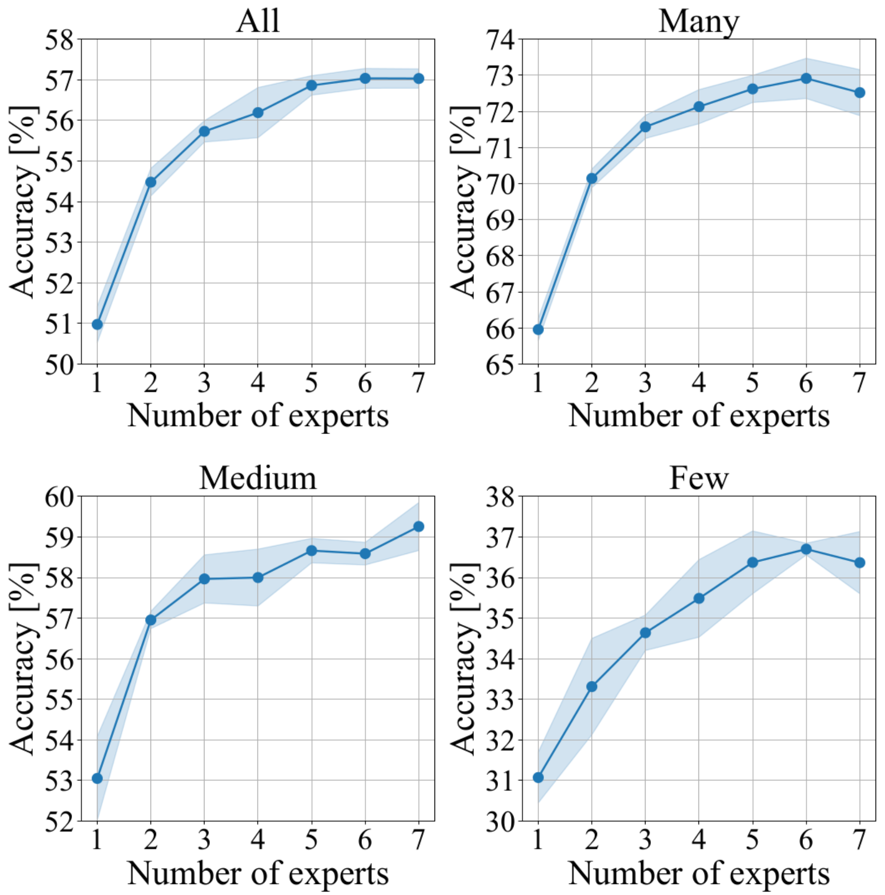


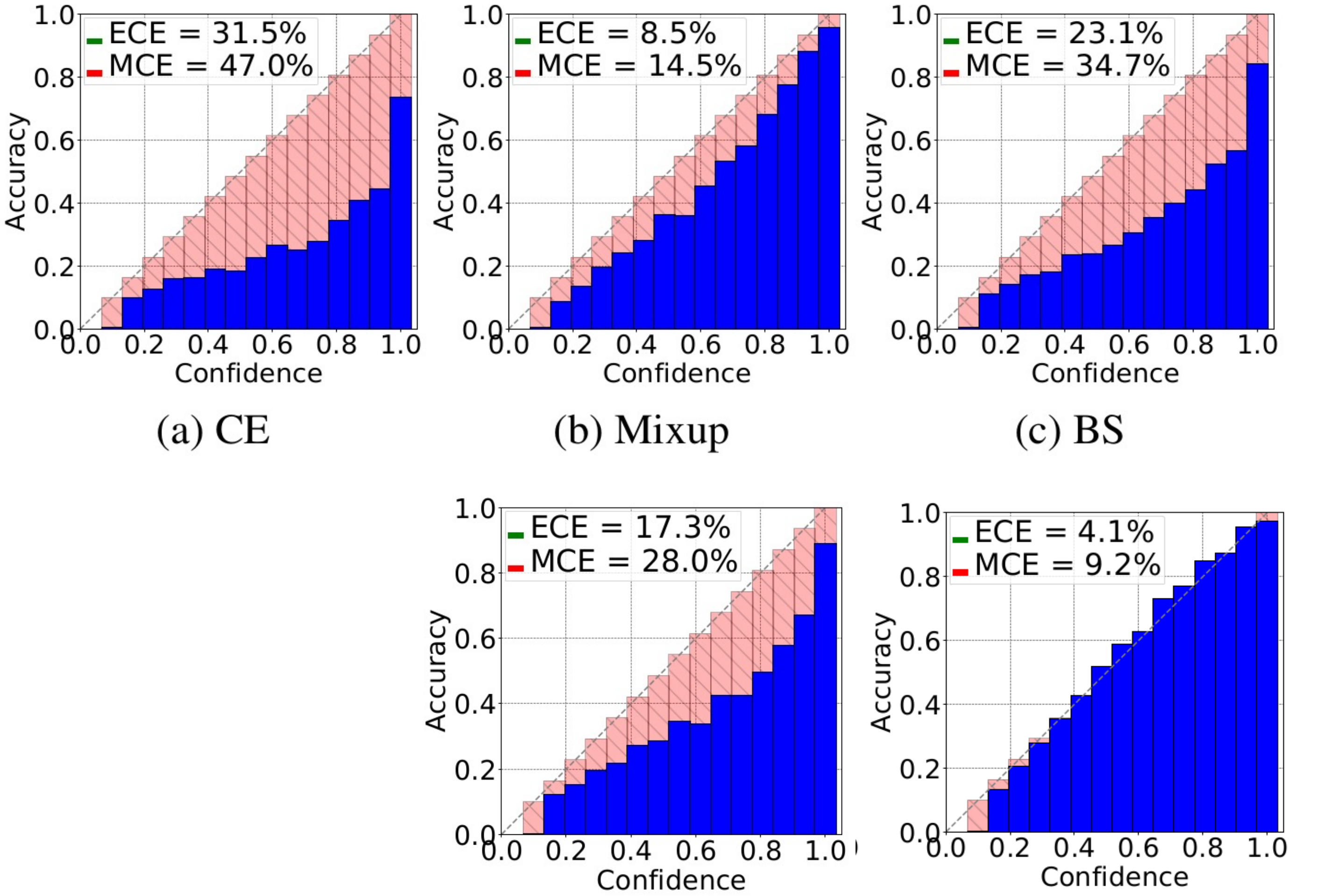
Figure 5. Test accuracy vs number of experts on CIFAR-100-LT-100 for all, many-shot, medium-shot and few-shot classes.

Robustness under varying degrees of distribution shift

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**

Confidence Calibration



(a) CE (b) Mixup (c) BS (d) SADE (e) BalPoE

Figure 2. Reliability plots for (a) CE, (b) mixup, (c) BS, (d) SADE, and (e) BalPoE. Computed over **CIFAR-100-LT-100** test set. We also report the Expected Calibration Error (ECE) and Maximum Calibration Error (MCE).

JUNE 18-22, 2023 CVPR VANCOUVER, CANADA

SOTA calibration under the balanced test distribution

Table 2. Expected calibration error (ECE), maximum calibration error (MCE), and test accuracy (ACC) on CIFAR-100-LT-100. ⋆: Our reproduced results. †: from [65]. ‡: 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 [65]	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⋆ [68]	9.6±0.8	15.9±1.5	40.8±0.7
Remix† [9]	33.6	51.0	41.9
UniMix+Bayias [65]	23.0	37.4	45.5
MiSLAS [70]	**4.8**	-	47.0
BalPoE (ours)	**4.9**±1.0	**11.3**±1.6	**52.0**±0.5

Conclusions

- We extend the logit adjustment formulation for training a balanced product of experts
- We show that the ensemble is Fisher-consistent for the balanced error, given that a simple constraint is fulfilled
- We find that expert calibration is important for achieving an unbiased ensemble using naïve expert aggregation

Thank you for listening!

SCAN ME