Appendix

6.1. Proof to Theorem 1

Assume $p_t(y|x)$ to be the target conditional probability and $p_s(y|x)$ to be the source conditional probability. We start with $p_s(y|x)$ formulated with logits $f_\theta(x)[y]$:

$$p_s(y|x) = \frac{e^{f_\theta(x)[y]}}{\sum_c e^{f_\theta(x)[c]}}. \tag{22}$$

By applying the log function on both sides,

$$\begin{aligned} f_\theta(x)[y] &= \log p_s(y|x) + C_x \\ &= \log\left(\frac{p_s(y)p_s(x|y)}{\sum_c p_s(c)p_s(x|c)}\right) + C_x \\ &= \log(p_s(y)p_s(x|y)) + C'_x \\ &= \log(p_s(y)p_t(x|y)) + C'_x \\ &= \log(p_t(y)p_t(x|y)) + \log p_s(y) \\ &\quad - \log p_t(y) + C'_x, \end{aligned} \tag{23}$$

where C_x and C'_x can be regarded as constants for a fixed x as follows:

$$C_x = \log\left(\sum_c e^{f_\theta(x)[c]}\right), \tag{24}$$

$$C'_x = C_x - \log\left(\sum_c p_s(c)p_s(x|c)\right). \tag{25}$$

Let us derive the post-compensated logit f_θ^{PC} (Definition 3.1) from f_θ:

$$\begin{aligned} &\log(p_t(y)p_t(x|y)) \\ &\quad = f_\theta(x)[y] - \log p_s(y) + \log p_t(y) - C'_x \\ &\quad = f_\theta^{PC}(x)[y] - C'_x. \end{aligned} \tag{26}$$

Re-calculating the Softmax function yields:

$$\begin{aligned} \frac{e^{f_\theta^{PC}(x)[y]}}{\sum_c e^{f_\theta^{PC}(x)[c]}} &= \frac{e^{f_\theta^{PC}(x)[y]-C'_x}}{\sum_c e^{f_\theta^{PC}(x)[c]-C'_x}} \\ &= \frac{p_t(y)p_t(x|y)}{\sum_c p_t(c)p_t(x|c)} \\ &= p_t(y|x), \end{aligned} \tag{27}$$

which ends the proof.

6.2. Implementation details

For all the experiments over multiple datasets, we use the SGD optimizer with momentum $\gamma = 0.9$ and weight decay $5 \cdot 10^{-4}$ to optimize the network if not specified. We use the same random seed throughout the whole experiment for a fair comparison. For image classification on CIFAR-100-LT and ImageNet-LT, we follow most of the details from [55], and on Places-LT and iNaturalist 2018, we follow [28]. All the models are trained on 4 GPUs, except CIFAR-100-LT, where we use 1 GPU. We find the optimal hyperparameters based on a grid search with the validation set. However, as the iNaturalist 2018 dataset does not contain the validation set, we use the same λ and α searched on the ImageNet-LT dataset since it has a similar number of classes and samples compared to the iNaturalist 2018 dataset. Detailed experiment settings for LADE are summarized in Table 7.

Table 7: Experimental settings on four benchmark datasets when using LADE. *IB* stands for the imbalance ratio.

Dataset	λ	α	Batch size
CIFAR-100-LT (*IB* 10)	0.01	0.01	256
CIFAR-100-LT (*IB* 50)	0.01	0.01	256
CIFAR-100-LT (*IB* 100)	0.01	0.1	256
Places-LT	0.1	0.005	128
ImageNet-LT	0.5	0.05	256
iNaturalist 2018	0.5	0.05	256

CIFAR-100-LT [30] On the CIFAR-100-LT dataset, we use ResNet-32 [22] as the backbone network for all the experiments, following the implementation of [55]. We train for 200 epochs and apply the linear warm-up learning rate schedule [19] to the first five epochs. The learning rate is initialized as 0.2, and it is decayed at the 120th and 160th epoch by 0.01.

Places-LT [64] We use ResNet-152 [22] as the backbone network with pretraining on the ImageNet-2012 [14] dataset. We use 0.05 and 0.001 for the initial learning rate of the classifier and the feature extractor. We train for 30 epochs with a learning rate decay of 0.1 every 10 epochs.

ImageNet-LT [14] On the ImageNet-LT dataset, we utilize ResNeXt-50-32x4d [61] as the backbone network for all the experiments. We use the cosine learning rate schedule [39] decaying from 0.05 to 0.0 during 180 epochs.

iNaturalist 2018 [57] For the iNaturalist 2018 dataset, we use ResNet-50 [22] as the backbone network for all experiments. We use cosine learning rate scheduling [39] decaying from 0.1 to 0.0 during 200 epochs, following [28].

Data Pre-processing We follow [38] for the details on image preprocessing. For the training set, images are resized to 256 × 256 and randomly cropped to 224 × 224. After cropping, we augment images with random horizontal flip with probability $p = 0.5$ and apply random color jitter. For validation and test set, images are center cropped to 224 × 224 without any augmentation.

6.3. Ablation study

To verify the effectiveness of the regularizer term for DV representation (Equation 9) and LADER (Equation 16), we conduct an ablation test. Table 8 shows how the top-1 accuracy changes when removing the regularizer term for the DV representation ($\lambda = 0$) or removing LADER ($\alpha = 0$), respectively.

Table 8: Ablation study for LADE on the long-tailed benchmark datasets. LADE (Ours) shows the best evaluation performance, and $\lambda = 0$ and $\alpha = 0$ denote the performance with the same settings except for the DV representation regularization or LADER, respectively.

Dataset	LADE (Ours)	$\lambda = 0$	$\alpha = 0$
CIFAR-100-LT (*IB* 10)	**61.7**	61.5	61.6
CIFAR-100-LT (*IB* 50)	**50.5**	49.5	49.9
CIFAR-100-LT (*IB* 100)	**45.4**	45.2	45.1
Places-LT	**38.8**	38.5	38.6
ImageNet-LT	**53.0**	47.0	52.1
iNaturalist 2018	**70.0**	58.3	69.8

[12] introduces λ to control the instability induced from directly using the DV representation. The model suffers a severe performance drop on ImageNet-LT and iNaturalist 2018 when the regularizer term for DV representation is not used ($\lambda = 0$). α represents the regularization strength of LADER on logits, as mentioned in Section 4.4. Without LADER ($\alpha = 0$), performance degradation is observed, demonstrating the efficacy of LADER.

6.4. Additional results on variously shifted test label distributions

In Section 4.3, we show that our LADE achieves state-of-the-art performance on variously shifted test label distribution with ImageNet-LT, which is the large-scale long-tailed dataset. We further conduct experiments on the small-scale dataset, CIFAR-100-LT, to ensure the consistent effectiveness of our LADE loss. For the training set, we use CIFAR-100-LT with an imbalance ratio of 50. The shifted test set is constructed by the same setting in Section 4.3. As shown in Table 9, LADE outperforms all the other methods, which is consistent with the results on ImageNet-LT (Table 6). We can also reconfirm the effectiveness of the PC strategy. These results from CIFAR-100-LT and ImageNet-LT imply that our PC strategy and LADE work well on both small-scale and large-scale datasets.

6.5. Additional confidence calibration results

We report the additional results of LADE against other methods in the perspective of confidence calibration, using the same datasets from the section above, CIFAR-100-LT with an imbalance ratio of 50 for the small-scale dataset and ImageNet-LT for the large-scale dataset. Following [53, 31], we estimate the quality of calibration on two datasets with four metrics:

- **Expected Calibration Error**

$$\text{ECE} = \sum_{m=1}^{M} \frac{|B_m|}{N} \cdot |acc(B_m) - conf(B_m)|, \tag{28}$$

- **Classwise Expected Calibration Error**

$$\text{Classwise-ECE} \\ = \frac{1}{C} \sum_{j=1}^{C} \sum_{m=1}^{M} \frac{|B_{m,j}|}{N} \cdot |acc(B_{m,j}) - conf(B_{m,j})| \tag{29}$$

- **Brier Score**

$$\text{Brier} = \sum_{i=1}^{N} \sum_{c=1}^{C} (p(y_i = c|x_i;\theta) - \mathbb{1}(y_i = c))^2, \tag{30}$$

- **Negative Log Likelihood**

$$\text{NLL} = -\sum_{i=1}^{N} \log p(y_i|x_i;\theta), \tag{31}$$

where N is the total number of test samples (x_i, y_i), C is the total number of classes, $M(= 20)$ is the total number of bins, each bin B_m is the set of indices of test samples where $\frac{m-1}{M} < p(y_i|x_i;\theta) \leq \frac{m}{M}$, $|B_m|$ is the total number of samples inside the bin B_m, $acc(B_m) = \frac{1}{|B_m|} \sum_{i \in B_m} \mathbb{1}(\arg\max_{y_j} p(y_j|x_i;\theta) = y_i)$, and $conf(B_m) = \frac{1}{|B_m|} \sum_{i \in B_m} p(y_i|x_i;\theta)$. The bin $B_{m,j}$ is the set of indices of test samples where the class for the samples is j, and the other definitions $|B_{m,j}|$, $acc(B_{m,j})$ and $conf(B_{m,j})$ are exactly same as the above.

Table 10 and 11 summarize the calibration results on CIFAR-100-LT and ImageNet-LT datasets, respectively. For all the evaluation metrics, LADE shows better overall calibration results than baseline methods. These observations demonstrate that our proposed LADE is effective in terms of calibration on both small-scale (CIFAR-100-LT) and large-scale (ImageNet-LT) datasets.

Table 9: Top-1 accuracy over all classes on test time shifted CIFAR-100-LT with imbalance ratio of 50.

Dataset	Forward					Uniform	Backward				
Imbalance ratio	50	25	10	5	2	1	2	5	10	25	50
Causal Norm	63.7	61.6	58.7	55.9	51.5	48.1	44.7	41.2	38.3	35.6	33.6
Balanced Softmax	59.6	58.5	56.9	54.8	52.2	49.9	47.5	45.1	42.7	40.9	39.9
Softmax	65.9	63.4	59.7	55.6	50.1	45.5	40.8	35.2	30.5	26.8	23.9
PC Causal Norm	66.1	62.9	58.8	55.6	51.2	48.1	45.7	44.2	43.4	44.3	44.9
PC Balanced Softmax	65.9	63.1	59.5	**56.3**	52.2	49.9	47.9	46.9	46.4	47.3	48.4
PC Softmax	66.0	63.2	59.2	55.9	52.4	49.5	47.5	46.7	46.2	47.4	49.0
LADE	**67.4**	**64.7**	**60.2**	**56.3**	**52.8**	**50.5**	**48.2**	**47.4**	**46.6**	**48.1**	**49.4**

Table 10: Confidence calibration results on CIFAR-100-LT with imbalance ratio of 50.

Method	Accuracy	ECE	Classwise ECE ($\times$1000)	Brier	NLL
Causal Norm	48.1	0.150	4.830	0.689	2.13
Balanced Softmax	49.9	0.168	4.607	0.673	2.07
Softmax	45.5	0.249	6.798	0.769	2.50
PC Softmax	49.5	0.174	4.723	0.678	2.10
LADE	**50.5**	**0.148**	**4.339**	**0.658**	**2.02**

Table 11: Confidence calibration results on ImageNet-LT.

Method	Accuracy	ECE	Classwise ECE ($\times$1000)	Brier	NLL
Causal Norm	52.0	0.108	0.4615	0.634	2.42
Balanced Softmax	52.1	0.061	0.4065	0.621	2.20
Softmax	48.2	0.140	0.6027	0.688	2.47
PC Softmax	52.8	0.057	0.4113	0.615	**2.17**
LADE	**53.0**	**0.035**	**0.4063**	**0.611**	2.18

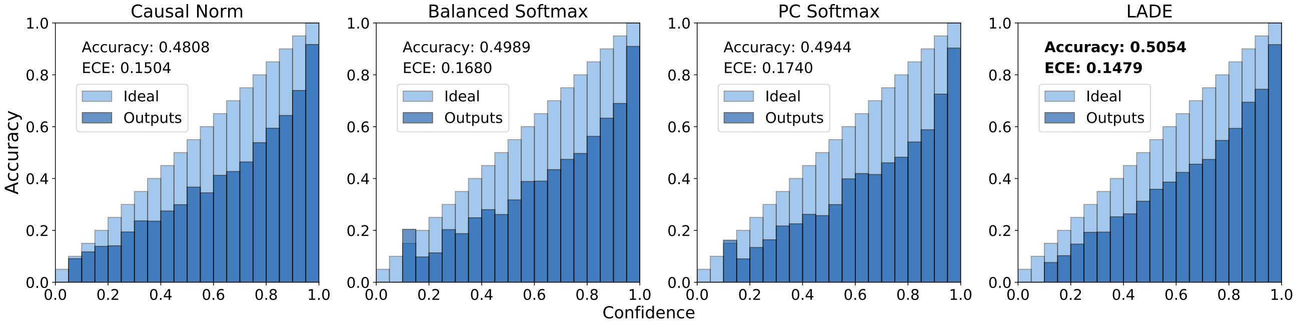

Figure 6: Reliability diagrams of ResNet-32 [22] on CIFAR-100-LT with imbalance ratio of 50.