Supplementary Material for Constructing Balance from Imbalance for Long-tailed Image Recognition

1 The Settings of Fig. 2 experiments

The separation model randomly sends samples to two groups with different probabilities: $(p, 1-p)$ for head classes and $(1-p, p)$ for tail classes. So larger p leads to higher accuracy. The classifier is a ResNet-50 trained with Adam optimizer Tail with lr=3e-4 and bz=512.

2 Proofs in Cluster Balancedness Loss

2.1 Equivalence of KL Divergence and Negative Entropy.

The KL divergence of $P(h|X)$ and discrete uniform distribution $u(h)$ is:

$$\begin{aligned} & KL\left[P(h|X) \| u(h)\right] \\ =& \sum_{k=1}^{K} P(h=k|X) \log \frac{P(h=k|X)}{u(h=k)} \\ =& \mathbb{E}_h\left[\log P(h|X)\right] - \sum_{k=1}^{K} P(h=k|X) \log \frac{1}{K} \\ =& -Ent\left[P(h|X)\right] + \log K. \end{aligned} \tag{1}$$

$Ent\left[P(h|X)\right]$ is the entropy of $P(h|X)$. Since $\log K$ is a constant, the optimization of KL divergence and negative entropy are equivalent in the *Cluster Balancedness Loss*.

2.2 Unbiasedness and Efficiency of Momentum Estimator.

We first convert the recursive formula of momentum estimator to closed-form:

$$\tilde{p}_t = \sum_{i=1}^{t} (1-\eta)\eta^{t-i} P(h=k|\mathcal{B}_i). \tag{2}$$

(1) Unbiasedness: Since $\hat{p} = P(h=k|\mathcal{B}_t)$ is unbiased $\mathbb{E}\left[\hat{p}\right] = P(h=k|X)$. Therefore we have

$$
\begin{aligned}
&\mathbb{E}\left[\frac{\tilde{p}_t}{1-\eta^t}\right] \\
=&\frac{1}{1-\eta^t}\mathbb{E}\left[\sum_{i=1}^{t}(1-\eta)\eta^{t-i}P(h=k|\mathcal{B}_i)\right] \\
=&\frac{1}{1-\eta^t}\sum_{i=1}^{t}(1-\eta)\eta^{t-i}P(h=k|X) \\
=&P(h=k|X)
\end{aligned}
\tag{3}
$$

is unbiased.
(2) Efficiency:

$$
\begin{aligned}
&\mathbb{D}\left[\frac{\tilde{p}_t}{1-\eta^t}\right] \\
=&\frac{(1-\eta)^2}{(1-\eta^t)^2}\mathbb{D}\left[\sum_{i=1}^{t}\eta^{t-i}P(h=k|\mathcal{B}_i)\right] \\
=&\frac{(1-\eta)^2}{(1-\eta^t)^2}\sum_{i=1}^{t}\eta^{2(t-i)}\mathbb{D}\left[\hat{p}\right] \\
=&\frac{(1-\eta)^2}{(1-\eta^t)^2}\cdot\frac{1-\eta^{2t}}{1-\eta^2}\mathbb{D}\left[\hat{p}\right] \\
=&\frac{(1-\eta)/(1+\eta)}{(1-\eta^t)/(1+\eta^t)}\mathbb{D}\left[\hat{p}\right].
\end{aligned}
\tag{4}
$$

Let $\phi(t)=\frac{1-\eta^t}{1+\eta^t}$, which is monotonic increasing when $0<\eta<1$. So $\mathbb{D}\left[\frac{\tilde{p}_t}{1-\eta^t}\right]=\frac{\phi(1)}{\phi(t)}\mathbb{D}\left[\hat{p}\right]\leq\mathbb{D}\left[\hat{p}\right]$ and $\frac{\tilde{p}_t}{1-\eta^t}$ is more efficient than $\hat{p}$.

3 Variance of Gaussian Mixture Centers

Following [4], the components of Gaussian Mixture are $\mathcal{N}(\mu_k, I)$, and each dimension of the center μ_k is sampled from $\mathcal{N}(0,\sigma^2)$. The σ is selected according to the distances between the generated centers. The clusters can overlap if the centers are too close, and samples may be stuck in the low-density area if the centers are far. The mean distance of two centers μ_p), μ_q is:

$$
\mathbb{E}\left(\|\mu_p-\mu_q\|^2\right)=\sum_{i=1}^{D}\mathbb{E}\left(\|\mu_{p,i}-\mu_{q,i}\|^2\right)=2D\sigma^2. \tag{5}
$$

If we expect the μ_p distributing around the *three-sigma borders* of μ_q ($mu_p-\mu_p=3$), then $\sigma=\sqrt{\frac{3}{2D}}\approx 0.04$ when the feature dimensionality is 1024. After experiments, we use $\sigma=0.05$ as the best choice.

4 Confusion Matrices of methods with DLSA

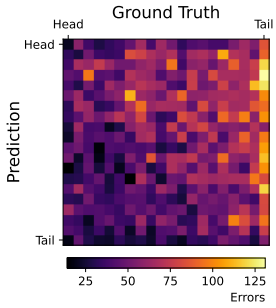

Fig. 1: PaCo + BalSoftmax on ImageNet w/o DLSA

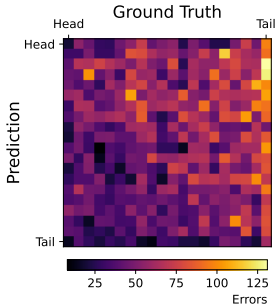

Fig. 2: PaCo + BalSoftmax on ImageNet w/ DLSA

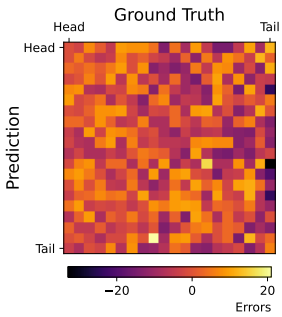

Fig. 3: Difference of PaCo w/o and w/ DLSA

The confusion matrix of PaCo+BalSoftmax w/ or w/o DLSA on ImageNet-LT are above. To illustrate the effect of DLSA, we also show the difference of Fig. 2 and Fig. 1 in Fig. 3. The darker colors indicate the reduction of errors and improvement of accuracy. In Fig. 3, the dark areas are mainly at the right top and middle since the DLSA reduces the error of misclassifying head classes to tail classes.

5 Hyper-parameters

We use cross-validation to select hyper-parameters. Different values are adopted on the datasets due to their different class number and granularity, unbalancedness and feature distribution.

Specifically, for ImageNet-LT, we use ResNet-50 [3] as the backbone. The backbone is trained from scratch with the feature learning methods following the previous methods. Each Flow Filter has 500 clusters and is learned with learning rate 0.2 and batch size 1024 for 50 epochs. The loss weights are $\lambda_{bal} = 1$, $\lambda_{pure} = 0.02$.

For Places-LT, we use ImageNet [2] pretrained ResNet-152 as the backbone. The Flow Filters are learned with learning rate 0.1 and batch size 512 for 60 epochs. The loss weights are $\lambda_{bal} = 2$, $\lambda_{pure} = 0.03$.

For iNaturalist18, we use ImageNet pretrained ResNet-50 as the backbone. The Flow Filters are learned with learning rate 0.2 and batch size 1024 for 30 epochs. The loss weights are $\lambda_{bal} = 1$, $\lambda_{pure} = 0.05$.

6 Results on iNaturalist18

The detailed results on iNaturalist18 (Many/Med/Few-show accuracy, MCC, NMI) are shown in Tab. 1.

Table 1: Results on iNaturalist18 [8] with ImageNet [2]-pretrained ResNet-50 [3].

Feature	Overall	Many	Medium	Few	MCC	NMI
LWS [5]	69.5	71.0	69.8	68.8	-	-
cRT [5]	68.2	73.2	68.8	66.1	-	-
PaCo+BalSoftmax [1]	71.8	73.9	71.2	71.8	71.1	94.3
PaCo+BalSoftmax [1]+ DLSA	**72.8**	**75.4**	**72.3**	**72.6**	**72.7**	**94.5**

7 Ablation study on Places-LT

Table 2: Ablation study on Places-LT [6] with PaCo [1]+BalSoftmax [7] model and ResNet-152 [3] backbone.

Method	Overall	Many	Medium	Few
Full model	**42.1**	**44.4**	**44.6**	**32.3**
w/o $\mathcal{L}_{MLE}$	40.8	43.7	43.2	30.1
w/o $\mathcal{L}_{bal}$	41.2	44.1	43.7	30.2
w/o $\mathcal{L}_{pure}$	41.5	44.0	43.9	31.2
300 clusters	41.8	44.0	**44.6**	31.5
1000 clusters	41.4	43.9	43.7	31.3

We extend the ablation study to Places-LT [6] on PaCo [1]+BalSoftmax [7] model. The results are shown in Tab. 2.

Objectives. Similar to ImageNet-LT, removing any loss ($\mathcal{L}_{MLE}$, $\mathcal{L}_{bal}$, $\mathcal{L}_{pure}$) leads to a significant performance drop. Among these losses, w/o $\mathcal{L}_{MLE}$ shows the greatest degradation since it controls the head-tail separation.

Cluster number. Models with less/more clusters perform worse than default 500 clusters. Larger cluster number results in slow training and inference too.

References

1. Cui, J., Zhong, Z., Liu, S., Yu, B., Jia, J.: Parametric contrastive learning. arXiv preprint arXiv:2107.12028 (2021)

2. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: 2009 IEEE conference on computer vision and pattern recognition. pp. 248–255. Ieee (2009)
3. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
4. Izmailov, P., Kirichenko, P., Finzi, M., Wilson, A.G.: Semi-supervised learning with normalizing flows. In: International Conference on Machine Learning. pp. 4615–4630. PMLR (2020)
5. Kang, B., Xie, S., Rohrbach, M., Yan, Z., Gordo, A., Feng, J., Kalantidis, Y.: Decoupling representation and classifier for long-tailed recognition. arXiv preprint arXiv:1910.09217 (2019)
6. Liu, Z., Miao, Z., Zhan, X., Wang, J., Gong, B., Yu, S.X.: Large-scale long-tailed recognition in an open world. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2537–2546 (2019)
7. Ren, J., Yu, C., Sheng, S., Ma, X., Zhao, H., Yi, S., Li, H.: Balanced meta-softmax for long-tailed visual recognition. arXiv preprint arXiv:2007.10740 (2020)
8. Van Horn, G., Mac Aodha, O., Song, Y., Cui, Y., Sun, C., Shepard, A., Adam, H., Perona, P., Belongie, S.: The inaturalist species classification and detection dataset. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 8769–8778 (2018)