1. Supplementary

Method	ImageNet-LT			
	Many	Medium	Few	Overall
Mixup [8]	67.3	41.0	11.1	46.3
Lifted Loss [6]	35.8	30.4	17.9	30.8
OLTR [5]	43.2	35.1	18.5	35.6
LWS [3]	60.2	47.2	30.3	49.9
TDE† [7]	62.5	47.8	29.9	51.0
τ-norm [3]	59.1	46.9	30.7	49.4
τ-norm + ELP-SR	**62.7**	**47.8**	**33.6**	**51.6**

Table 1. Comparison on ImageNet-LT. † denotes that the results are re-implemented by the author-provided code.

1.1. Long-tailed Visual Recognition on ImageNet-LT

We provide results on ImageNet-LT as shown in Table 1 additionally. ImageNet-LT, a sub-set from ImageNet [4], is a standard benchmark in the long-tailed recognition task from [5]. The largest 'head' class in ImageNet-LT contains 1280 images, and the smallest 'tail' class possesses only five images. This formulates a typical long-tailed problem and leads to a challenging benchmark. We compare methods in four conditions: (1). Many: the many-shot condition in which more than 100 samples per class; (2). Medium: the medium-shot condition in which 20 to 100 samples per class; (3). Few: the few-shot condition in which the number of samples per class is less than 20; (4). Overall: the overall dataset. We report top-1 accuracy of all conditions for various methods in Table 1.

After applying ELP-SR in training, 2.2% improvements occur based on [3], which is significant in this benchmark. Besides, without bells and whistles, our results are competitive to the recent method [7]. ELP-SR does not introduce any overhead in testing and even does not specifically consider the long-tail distribution. Improvements indicate that ELP-SR guides the networks to become more generalized.

1.2. Ablation for Fine-grained Visual Recognition

More ablation studies are provided for CAR and AIR datasets in fine-grained recognition as shown in Table 2. Similar to CUB dataset, the best performances occur when $\gamma = 3$ in both datasets. Meanwhile, the value of $\mathcal{I}$ should also be proper. Too large or too small values of $\mathcal{I}$ may slightly hinder the improvements from ELP-SR.

Besides, we also applied ELP with PMG [1] on CUB. The original PMG achieves the accuracy of 88.9% using a single branch. PMG with ELP achieves 89.3%, outperforming PMG by 0.4%. This reveals that ELP provides consistent improvements on complex structures.

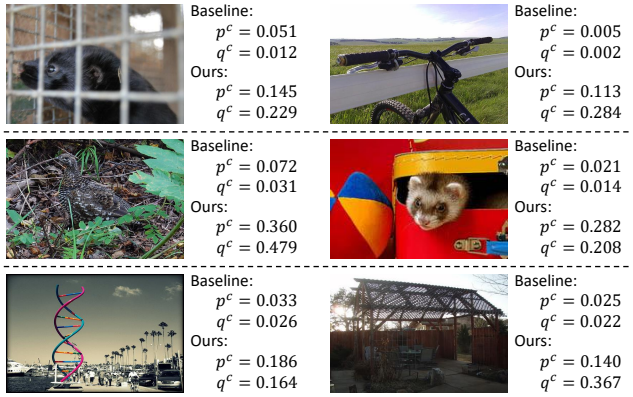

Figure 1. Visualization for cases that are correctly classified by ours but misclassified by the baseline method.

1.3. Visualization

Three kinds of visualized comparisons are provided. The first comparison is for samples. The influences of probabilities are presented with corresponding samples. The second comparison is for features. The intuitive improvements from features are revealed. The final comparison is for the weights of classifiers. The changing of classifiers' weights is shown.

Visualization for Samples: We showcase some examples in the testing set of ImageNet-1K. Based on ResNet-50, examples of the baseline method with or without ELP-SR are presented in Fig. 1. To compare the probabilities from baseline, we train the ELP classifier with baseline features but do not leverage ELP-SR to the training procedure.

We present the probabilities for the ground truth classes in the main classifier (p^c) and the ELP classifier (q^c) in baseline and ours, respectively. In comparison, most p^c values in the baseline are larger than q^c, revealing that the main classifier of baseline may overfit. Though the features are not discriminative enough and achieve relatively lower confidences from the ELP classifier, the main classifier presents higher probabilities. Moreover, our q^c values are larger than the baseline, reflecting that our features are more discriminative and easily classified by the simple ELP classifier.

Visualization for Features: We further provide the visualization for features to directly reveal the improvements from ELP-SR. We randomly sample 10 classes from CUB every time, apply PCA [2] to reduce the dimensions, and visualize them as in Fig. 2. Every column in Fig. 2 indicates the same classes sampled from CUB. The visualizations show that our method leads the features to become more discriminative.

Visualization for Classifiers' Weights: We visualized the classifiers' weights using tSNE. As shown in Figure 3, we show the distributions of classifiers after 1, 100, and 200 epochs. In the main classifier, the margin between classes

Parameter	CAR					AIR				
	$\mathcal{I}=1$	$\mathcal{I}=2$	$\mathcal{I}=3$	$\mathcal{I}=4$	$\mathcal{I}=5$	$\mathcal{I}=1$	$\mathcal{I}=2$	$\mathcal{I}=3$	$\mathcal{I}=4$	$\mathcal{I}=5$
$\gamma=1$	93.8	93.6	94.0	93.6	93.2	91.2	92.3	92.0	91.6	91.5
$\gamma=2$	93.9	**94.2**	94.0	93.7	93.3	92.5	92.2	91.5	91.5	91.2
$\gamma=3$	93.8	**94.2**	93.9	93.5	93.3	**92.7**	91.3	92.0	91.7	91.5
$\gamma=4$	93.8	**94.2**	94.1	93.6	93.1	91.6	92.2	92.1	91.5	91.2

Table 2. Ablation of parameters on CAR and AIR.

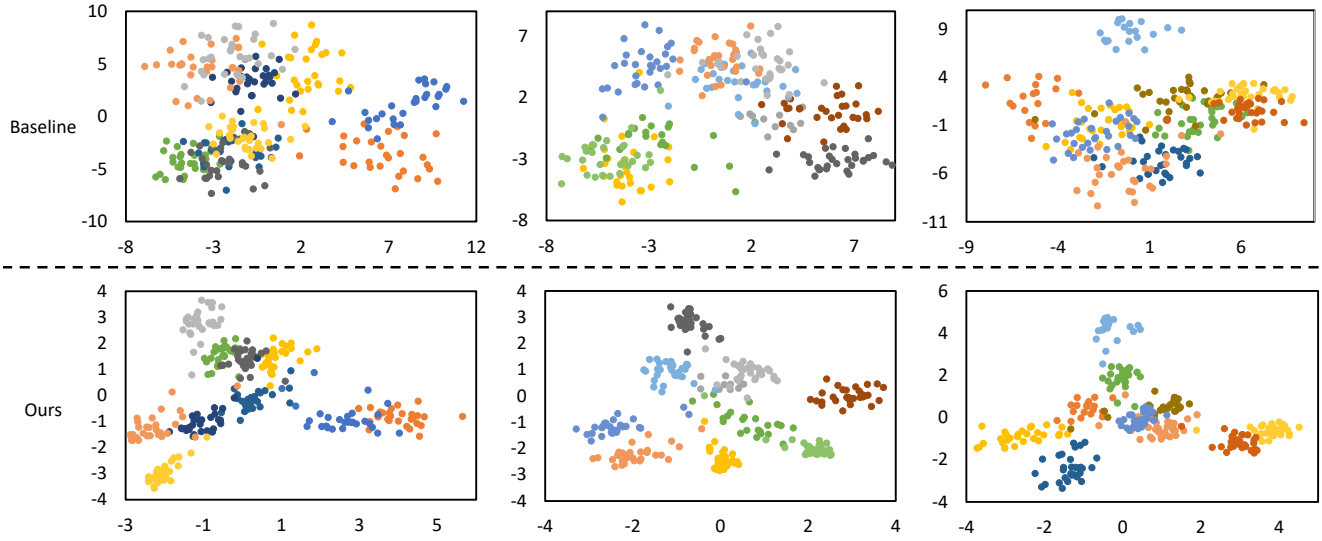

Figure 2. Visualization for features. Every column presents the same set of classes randomly sampled from the CUB dataset. The first raw is for thr baseline method, and the second raw is ours.

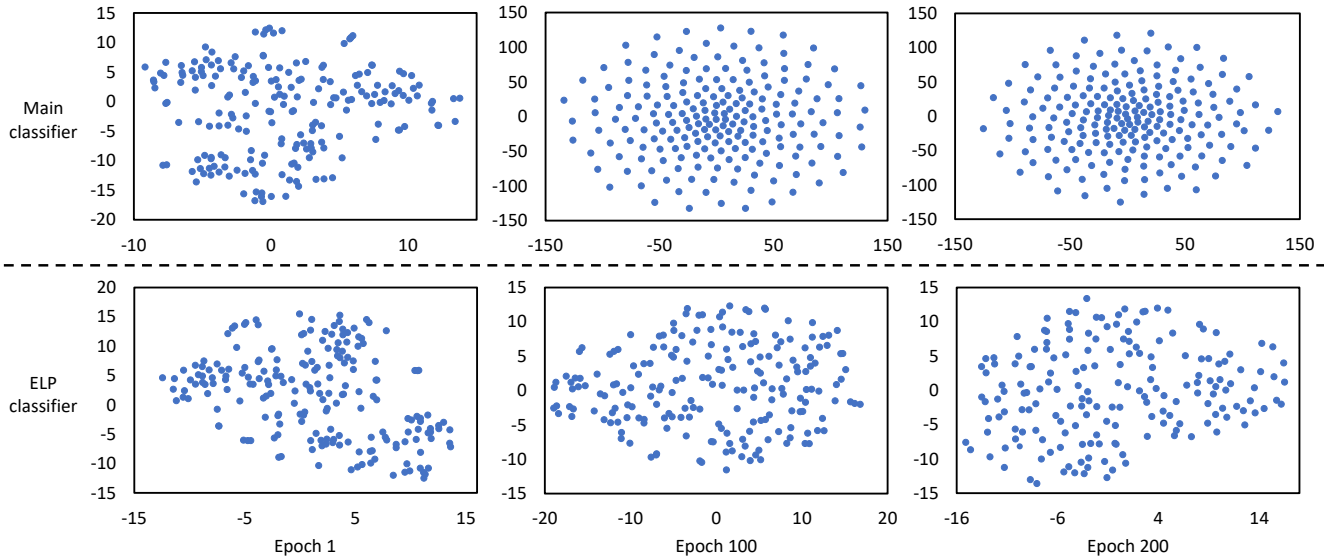

Figure 3. Visualization for classifiers' weights in CUB. Every point indicates the weights for the corresponding category in the classifier. Every column presents the weights in the same epoch. We visualize the weights after the 1st, 100th, and 200th epochs. The first raw is for the weights of the main classifier, and the second raw is the ELP classifier.

becomes more recognizable gradually. In the ELP classifier, the m argins are not as recognizable as the main classifier,

which indicates the ELP classifier can not overfit. Meanwhile, the weights from the ELP classifier become more discriminative after plenty of training steps. This phenomenon also reflects that the network achieves better immediate suitability with ELP.

2. Discussions for Limitation and Impact

Though ELP provides general improvements in recognition, the extensions for other tasks like regression, generation, etc., may be challenging. Since ELP needs to reflect simplicity, it is difficult for ELP to directly regress a particular value or fit a complex feature space. Besides, our method may also have a potential negative impact. Adversarial attacks targeting the ELP may significantly affect the model training. Moreover, since ELP reflects some properties of features in training, even if the backbone model is completely encapsulated and unavailable, the ELP layer may disclose information of training data.

References

[1] Ruoyi Du, Dongliang Chang, Ayan Kumar Bhunia, Jiyang Xie, Yi-Zhe Song, Zhanyu Ma, and Jun Guo. Fine-grained visual classification via progressive multi-granularity training of jigsaw patches. In *European Conference on Computer Vision*, 2020. 1

[2] Karl Pearson F.R.S. Liii. on lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 2(11):559–572, 1901. 1

[3] 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*, 2020. 1

[4] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Communications of the ACM*, 60(6):84–90, 2017. 1

[5] Ziwei Liu, Zhongqi Miao, Xiaohang Zhan, Jiayun Wang, Boqing Gong, and Stella X. Yu. Large-scale long-tailed recognition in an open world. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 1

[6] Hyun Oh Song, Yu Xiang, Stefanie Jegelka, and Silvio Savarese. Deep metric learning via lifted structured feature embedding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4004–4012, 2016. 1

[7] Kaihua Tang, Jianqiang Huang, and Hanwang Zhang. Long-tailed classification by keeping the good and removing the bad momentum causal effect. *arXiv preprint arXiv:2009.12991*, 2020. 1

[8] Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412*, 2017. 1