Supplementary Materials: Parameter-Efficient Complementary Expert Learning for Long-Tailed Visual Recognition

1 DETAILED EXPERIMENTAL SETTINGS

The detailed experimental settings on CIFAR100-LT [4], ImageNet-LT [1], Places-LT [17] and iNaturalist 2018 [13] are presented in Table 1 and 2. † denotes lower bottle dimensions in PEFT blocks. As shown in Table 1, except for the bottleneck dimension, CIFAR100-LT, ImageNet-LT and Places-LT share the same settings. Experiments on iNaturalist 2018 adopt the AdamW [7] optimizer with longer training epochs.

Table 1: The default experimental settings on CIFAR100-LT [4], ImageNet-LT [1], Places-LT [17]. † denotes a setting with a lower bottleneck dimension.

	CIFAR100-LT	ImageNet-LT	Places-LT
optimizer	SGD	SGD	SGD
training epochs	10	10	10
learning rate	1e-2	1e-2	1e-2
weight decay	5e-4	5e-4	5e-4
batch size	128	128	128
expert number	3	3	3
weight factors τ	[0.5, 1.0, 1.5]	[0.5, 1.0, 1.5]	[0.5, 1.0, 1.5]
bias term ϵ	0.1	0.1	0.1
bottleneck dim †	1	32	4
bottleneck dim	14	64	24

Table 2: The default experimental settings on iNaturalist 2018 [13]. † denotes a setting with a lower bottleneck dimension.

	iNaturalist 2018
optimizer	AdamW
training epochs	20
learning rate	5e-4
weight decay	5e-4
batch size	128
expert number	3
weight factors τ	[0.5, 1.0, 1.5]
bias term ϵ	0.1
bottleneck dim †	128
bottleneck dim	224

2 MORE EXPERIMENTAL ANALYSIS

2.1 Scale factors of Different Experts

In Figure 1, we present the learned scale factors in different AdaptFormer blocks of PECEL. A larger scale factor means that the output features of this PEFT block are more significant. As shown in Figure 1, different experts in PECEL typically exhibit the same significance tendency in different layers. Since the resolutions of images in CIFAR100-LT and training images for CLIP [9] are different, the scale factors in shallow layers are higher than the factors in mediate layers. Besides, due to the difference of semantic categories, the deeper layers typically exhibit higher scale factors to learn representations with better semantic discriminability on all three datasets.

2.2 ϵ in Sample-aware Logit Adjustment

In Figure 2, we present the impact of bias value ϵ in the sample-aware logit adjustment on CIFAR100-LT. Figure 2 shows that the accuracy on the few-shot class increases when increasing the bias value ϵ, since a larger ϵ encourages the model to focus more on the misclassified samples. We select $\epsilon = 0.1$ in our experiments since it achieves the highest overall accuracy.

2.3 Regularization Loss

The intention of regularization loss is to reduce the representation redundancy of shared parameters in PECEL. To verify this, in Figure 3, we present the mean cosine similarities of the shared parameters of training PECEL with and without the regularization loss on ImageNet-LT. As presented in Figure 3, training PECEL without the regularization loss (PECEL w/o Reg) naturally decreases the self-similarity of parameters, since the parameter sharing strategy inherently encourages to better exploit the shared parameters and reduces their redundancy. Due to the soft orthogonal regularization, training with the regularization loss (PECEL w/ Reg) can further reduce similarity.

2.4 Performance on Larger Models

To demonstrate PECEL's generalization ability on larger models, in Table 3, we report the recognition accuracy of LIFT [10] and our proposed PECEL when using larger pretrained foundation models, *i.e* CLIP with ViT-Large as the backbone (CLIP-ViT-L) [9]. Table 3 shows that the proposed PECEL can also outperform LIFT with a larger pretrained foundation model and achieve higher performance on many-shot, medium-shot and few-shot classes.

3 RESULTS ON PLACES-LT

We present the results on Places-LT in Table 4. As reported in Table 4, due to the complementary expert learning, the proposed PECEL can achieve 52.5% classification accuracy on Places-LT, outperforming the previous state-of-the-art method LIFT [10]. PECEL can also achieve a more balanced performance in different class groups. When decreasing the bottleneck dimension to 4 (denoted with †), due to the proposed parameter sharing strategy, PECEL can achieve comparable performance with LIFT with about 77% fewer parameters.

It's noted that compared with other datasets, the improvement of PECEL on Places-LT is somewhat marginal. To investigate this, as shown in Figure 4, we analyze the classification results of validation

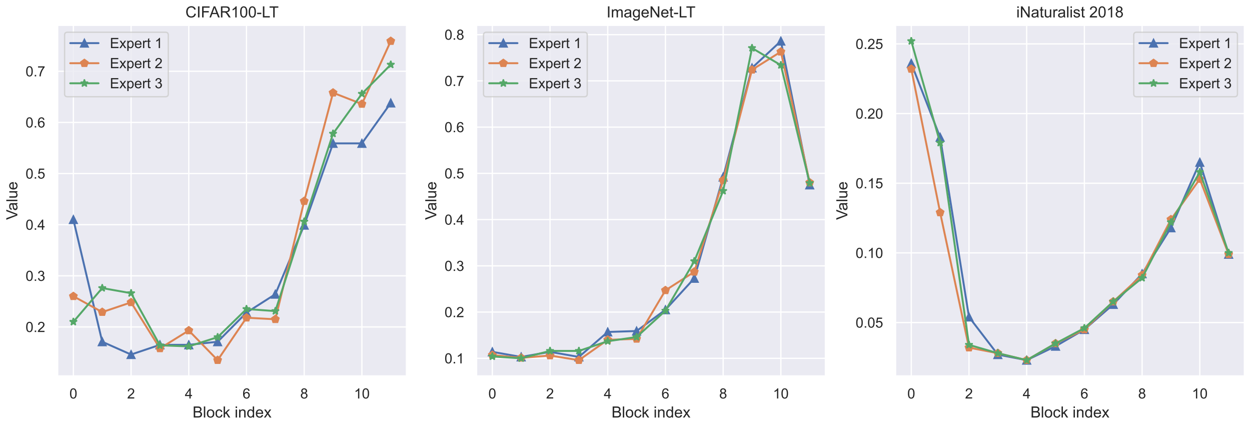

Figure 1: The learned scale factors of different PEFT blocks in PECEL.

Table 3: The recognition accuracy on CIFAR100-LT (left) and ImageNet-LT (right).

Method	Backbone	Many	Med	Few	All
LIFT [10]	CLIP-ViT-B	85.2	82.4	76.8	81.7
PECEL	CLIP-ViT-B	87.5	83.3	81.8	84.3
LIFT [10]	CLIP-ViT-L	89.7	87.6	85.9	87.8
PECEL	CLIP-ViT-L	91.4	87.9	87.7	89.0

Method	Backbone	Many	Med	Few	All
LIFT [10]	CLIP-ViT-B	81.3	77.4	73.4	78.3
PECEL	CLIP-ViT-B	82.1	77.8	76.3	79.2
LIFT [10]	CLIP-ViT-L	85.3	81.7	78.3	82.7
PECEL	CLIP-ViT-L	86.3	82.5	79.6	83.6

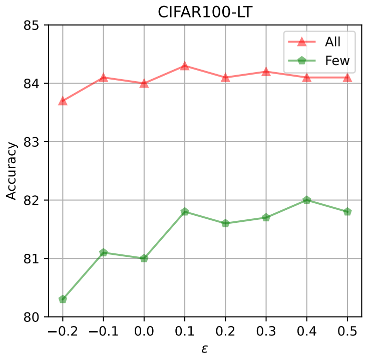

Figure 2: The impact of bias value ϵ in the sample-aware logit adjustment on CIFAR100-LT.

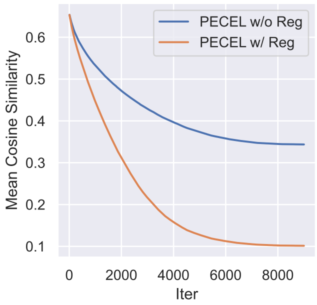

Figure 3: The cosine similarity of the shared parameters in PECEL with and without regularization loss.

samples in Places-LT. We visualize the samples in Top-3 classes with the lowest accuracy, *i.e Building facade*, *River* and *Mountain*. For each class, we present the number of training samples, the class accuracy on the validation set, the Top-5 wrongly predicted classes and some corresponding samples. For example, class *Building facade* includes 841 training samples and 100 validation samples, with validation accuracy of 2%, *i.e* only 2 samples in the validation set are correctly predicted. There are 9, 8, 6, 5 and 5 samples in the validation set that are wrongly predicted as *Synagogue-outdoor*, *Embassy*, *Parking garage-outdoor*, *Courthouse* and *Hospital*, respectively. These misclassified images are highlighted in red. The correctly classified images are highlighted in green. The right part in Figure 4 showcases some reference samples of these misclassified classes for comparison.

In Figure 4, it's noted that the semantics of ground-truth classes and the wrongly predicted classes are often overlapped. For example, the *Synagogue-outdoor* often contains *Building facade*. The misclassified samples are also more similar to the corresponding classes. Moreover, an image can contain multiple categories. For example, the images of *River* that are *misclassified* as *Bridges* do contain both rivers and bridges, which cannot demonstrate the predictions are incorrect. Therefore, the evaluation on Places-LT may not fully demonstrate the efficacy of the proposed PECEL.

Table 4: Accuracy on Places-LT. Param: number of trainable parameters in backbone; †: lower bottleneck dimension. The best and second-best results are highlighted in bold and underline, respectively.

	Places-LT					
	Backbone	Param	Many	Med	Few	All
Training from scratch.						
SADE [15] NeurIPS'22	ResNet152	169.3M	-	-	-	40.9
NCL [5] CVPR'22	ResNet152	169.3M	-	-	-	41.8
LiVT [14] CVPR'23	ViT-B	85.0M	48.1	40.6	27.5	40.8
MDCS [16] ICCV'23	ResNet152	169.3M	43.1	42.9	36.3	42.4
LGLA [11] ICCV'23	ResNet152	169.3M	-	-	-	42.0
Fine-tuning pretrained models.						
Zero-Shot	ViT-B	-	38.3	39.2	45.9	40.2
Full Fine-Tuning	ViT-B	85.0M	51.6	48.5	36.2	47.2
BALLAD [8] arXiv'21	ViT-B	149.6M	49.3	50.2	48.4	49.5
VL-LTR [12] ECCV'22	ViT-B	149.6M	**54.2**	48.5	42.0	50.1
RAC [6] CVPR'22	ViT-B	85.0M	48.7	48.3	41.8	47.2
UDCPG [3] ACM MM'23	ResNet50	23.5M	50.8	48.8	44.6	48.7
LPT [2] ICLR'23	ViT-B	1.01M	51.3	52.2	50.5	51.5
LIFT [10] ICML'24	ViT-B	0.18M	51.7	**53.1**	50.9	52.2
PECEL†	ViT-B	**0.04M**	51.6	52.0	**53.9**	52.2
PECEL	ViT-B	0.17M	52.2	52.4	53.5	**52.5**

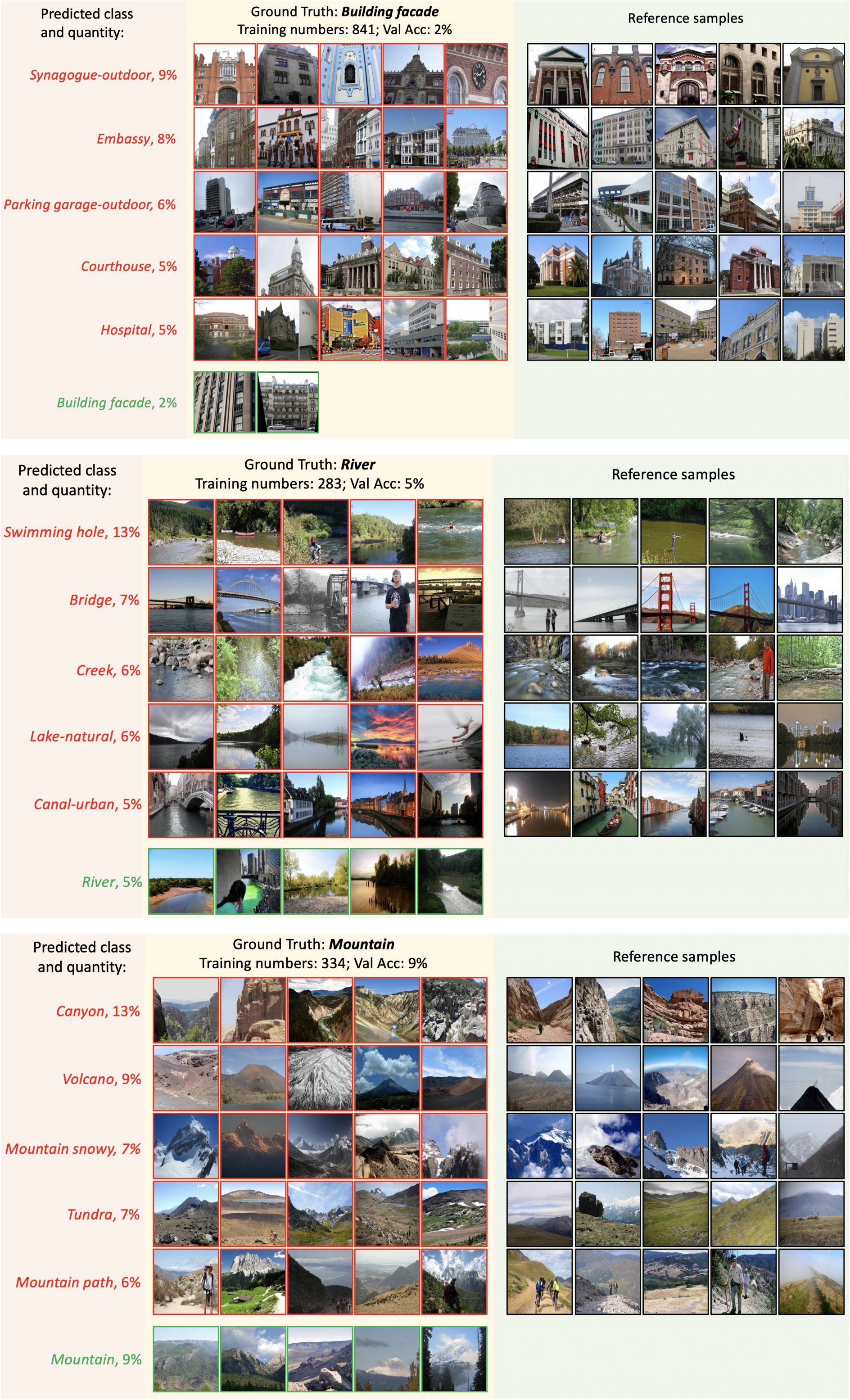

Figure 4: Visualization of the validation samples in Places-LT. We visualize the Top-3 categories with the lowest accuracy, *i.e* *Building facade, River* and *Mountain*. In the left part, the correctly and wrongly predicted samples are highlighted in green and red, respectively. In the right part, we also present some true samples of the wrongly predicted classes for reference and comparison.

REFERENCES

[1] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. Ieee, 248–255.

[2] Bowen Dong, Pan Zhou, Shuicheng Yan, and Wangmeng Zuo. 2023. LPT: Long-tailed Prompt Tuning for Image Classification. In *International Conference on Learning Representations*.

[3] Xiaoxuan He, Siming Fu, Xinpeng Ding, Yuchen Cao, and Hualiang Wang. 2023. Uniformly Distributed Category Prototype-Guided Vision-Language Framework for Long-Tail Recognition. In *Proceedings of the 31st ACM International Conference on Multimedia*. 5027–5037.

[4] Alex Krizhevsky et al. 2009. Learning multiple layers of features from tiny images. (2009).

[5] Jun Li, Zichang Tan, Jun Wan, Zhen Lei, and Guodong Guo. 2022. Nested collaborative learning for long-tailed visual recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 6949–6958.

[6] Alexander Long, Wei Yin, Thalaiyasingam Ajanthan, Vu Nguyen, Pulak Purkait, Ravi Garg, Alan Blair, Chunhua Shen, and Anton van den Hengel. 2022. Retrieval augmented classification for long-tail visual recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 6959–6969.

[7] Ilya Loshchilov and Frank Hutter. 2018. Decoupled Weight Decay Regularization. In *International Conference on Learning Representations*.

[8] Teli Ma, Shijie Geng, Mengmeng Wang, Jing Shao, Jiasen Lu, Hongsheng Li, Peng Gao, and Yu Qiao. 2021. A simple long-tailed recognition baseline via vision-language model. *arXiv preprint arXiv:2111.14745* (2021).

[9] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*. PMLR, 8748–8763.

[10] Jiang-Xin Shi, Tong Wei, Zhi Zhou, Jie-Jing Shao, Xin-Yan Han, and Yu-Feng Li. 2024. Long-Tail Learning with Foundation Model: Heavy Fine-Tuning Hurts. In *International Conference on Machine Learning*.

[11] Yingfan Tao, Jingna Sun, Hao Yang, Li Chen, Xu Wang, Wenming Yang, Daniel Du, and Min Zheng. 2023. Local and Global Logit Adjustments for Long-Tailed Learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 11783–11792.

[12] Changyao Tian, Wenhai Wang, Xizhou Zhu, Jifeng Dai, and Yu Qiao. 2022. Vl-ltr: Learning class-wise visual-linguistic representation for long-tailed visual recognition. In *European Conference on Computer Vision*. Springer, 73–91.

[13] Grant Van Horn, Oisin Mac Aodha, Yang Song, Yin Cui, Chen Sun, Alex Shepard, Hartwig Adam, Pietro Perona, and Serge Belongie. 2018. The inaturalist species classification and detection dataset. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 8769–8778.

[14] Zhengzhuo Xu, Ruikang Liu, Shuo Yang, Zenghao Chai, and Chun Yuan. 2023. Learning Imbalanced Data with Vision Transformers. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 15793–15803.

[15] Yifan Zhang, Bryan Hooi, Lanqing Hong, and Jiashi Feng. 2022. Self-supervised aggregation of diverse experts for test-agnostic long-tailed recognition. *Advances in Neural Information Processing Systems* 35 (2022), 34077–34090.

[16] Qihao Zhao, Chen Jiang, Wei Hu, Fan Zhang, and Jun Liu. 2023. MDCS: More Diverse Experts with Consistency Self-distillation for Long-tailed Recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 11597–11608.

[17] Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. 2017. Places: A 10 million Image Database for Scene Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2017).