

Supplementary Materials: Category-Prompt Refined Feature Learning for Long-Tailed Multi-Label Image Classification

Anonymous Authors

This supplementary material contains complete experimental setup and additional experimental results in our paper “Category-Prompt Refined Feature Learning for Long-Tailed Multi-Label Image Classification”, including the comparison results with CLIP-based methods, further analysis on components of CPRFL, the effect of expansion coefficient τ , and more visualization examples.

1 EXPERIMENTAL SETUP

1.1 More Experimental Details

Without otherwise stating, we set visual features and category-prompts, both with a dimensionality of 2048. As for the pretrained CLIP models, we adopt CLIP ResNet-50 or ViT-Base/16 [4] and use the corresponding CLIP Transformer as the text encoder. During training, the parameters of the text encoder are kept frozen, and only learnable modules are optimized. For the Transformer encoder in the VSI network, the number of attention heads is 8, the dimension of each head is set to 512, and the hidden dimension of the FFN is set to 2048. In both two sub-networks, we adopt GELU as the non-linear activation function, which leads to faster convergence in experiments. We use ResNet-101 [1] as the backbone, which is a widely used feature extractor in MLC. During training, the input images are randomly cropped and resized into 448×448 with RandAugment [7] for data augmentation. For network optimization, we use the Adam optimizer [2] with a weight decay of $1e-4$ and the training epochs are set to 30. The batch size is 32, and the learning rates for COCO-LT, and VOC-LT are empirically initialized with $1e-5$, $5e-5$. All experiments are performed on one Nvidia RTX-3090 GPU, and our model is implemented in PyTorch 1.12.0.

1.2 Evaluation Metrics

Consistent with the evaluation metrics in [3], the classes in both datasets are categorized into head, medium, and tail groups based on the number of training samples. Specifically, head classes comprise over 100 samples, medium classes consist of 20 to 100 samples each, and tail classes contain fewer than 20 samples each. We use mean average precision (**mAP**) as the evaluation metric to assess the performance of long-tailed multi-label visual recognition across all classes.

2 ADDITIONAL EXPERIMENTAL RESULTS

2.1 Comparison Results with CLIP-Based Methods

To further demonstrate the effectiveness of our CPRFL, we compare it with several existing CLIP-based methods on two long-tailed multi-label datasets. These methods include zero-shot CLIP [4] and popular prompt tuning methods, i.e., CoOp [9], CoCoOp [8], Dual-CoOp [5] and LMPT [6]. We present the results using CLIP ResNet-50 and ViT-Base/16. Table 1 illustrates the mAP performance of different methods. Notably, these CLIP-based methods all utilize

the pretrained CLIP’s image encoder and we use “*” to indicate this. Experimental results indicate that our proposed method outperforms previous CLIP-based methods, achieving a remarkable 7.72% increase in total mAP with CLIP-RN50 on the COCO-LT dataset. Although the CLIP-based methods tend to outperform our method on tail classes, this is primarily due to the CLIP image encoder’s pretraining on a vast dataset containing many tail samples, inadvertently leveraging prior visual data exposure and potentially creating an unfair comparison. Furthermore, these CLIP-based methods tend to compromise the performance of the head classes while improving the tail classes. In contrast, our CPRFL can achieve synchronous improvements in head-to-tail recognition performance for LTMLC, addressing the imbalance class distribution effectively.

2.2 Further Analysis on Components of CPRFL

To further evaluate the contributions of various components to our method for long-tailed multi-label classification, we present the mAP improvements over baseline by integrating VSI, PI, and RW components of the proposed CPRFL in Figure 1. As illustrated in the figure, our method yields the most significant performance boost for tail classes compared to the baseline, with the tail mAP surpassing the baseline by 14.87% on VOC-LT and 19.19% on COCO-LT. These results underscore the effectiveness of our CPRFL approach in enhancing recognition performance for tail classes in LTMLC tasks. The improvements achieved through the integration of VSI, PI, and RW components highlight the comprehensive strengths of our method in addressing challenges posed by the imbalanced distribution of class frequencies and its ability to recognize a broader range of categories accurately.

2.3 Ablation Study on Expansion Coefficient

The expansion coefficient τ plays a pivotal role in determining the dimensionality of the hidden layers within the Prompt Initialization (PI) network. To assess the impact of varying values of τ on category-prompt initialization, we present the results in Figure 2. A key observation is that a value of $\tau = 0$ represents the use of a single linear layer for projection. The results indicate that optimal performance tends to emerge around $\tau = 0.5$. Notably, using a single linear layer ($\tau = 0$) for projection does not achieve the same level of performance as using a nonlinear extraction structure. Lower values of τ may facilitate easier learning and generalization for the model; however, the single linear layer’s ease of use cannot fully offset its limitations when projecting from semantic space to the visual-semantic joint space.

2.4 Additional Qualitative Analysis

To better understand how our method handles long-tailed multi-label data, we conduct additional qualitative experiments using ResNet-50, CLIP, and our proposed CPRFL. Figure 3 presents more

Table 1: The mAP (%) performance of the proposed CPRFL and CLIP-based methods on two long-tailed multi-label datasets. We present the mAP results on overall, head, medium, and tail classes under CLIP ResNet-50 and ViT-Base/16. * indicates that the method uses CLIP’s image encoder. Bold indicates the best scores.

Datasets	VOC-LT				COCO-LT			
	total	head	medium	tail	total	head	medium	tail
<i>CLIP:RN50</i>								
Zero-Shot CLIP*	84.30	63.60	88.03	97.03	56.19	35.73	60.52	68.45
CoOp* [9]	81.34	65.10	81.54	93.37	54.94	38.06	56.67	67.51
CoCoOp* [8]	78.63	64.33	80.51	87.94	46.02	36.02	50.57	48.82
DualCoOp* [5]	81.03	66.45	80.53	92.33	53.11	40.48	55.20	62.11
LMPT* [6]	85.44	66.45	88.11	97.86	58.97	41.87	61.60	69.60
CPRFL-CLIP(ours)	86.28	81.84	90.51	86.43	66.69	66.35	70.99	61.33
<i>CLIP:ViT16</i>								
Zero-Shot CLIP*	85.77	66.52	88.93	97.83	60.17	38.52	65.06	72.28
CoOp* [9]	86.02	67.71	88.79	97.67	60.68	41.97	63.18	73.85
CoCoOp* [8]	84.47	64.58	87.82	96.88	61.49	39.81	64.63	76.42
LMPT* [6]	87.88	72.10	89.26	98.49	66.19	44.89	69.80	79.08
CPRFL-CLIP(ours)	85.84	82.65	89.49	85.49	66.84	66.55	71.45	61.01

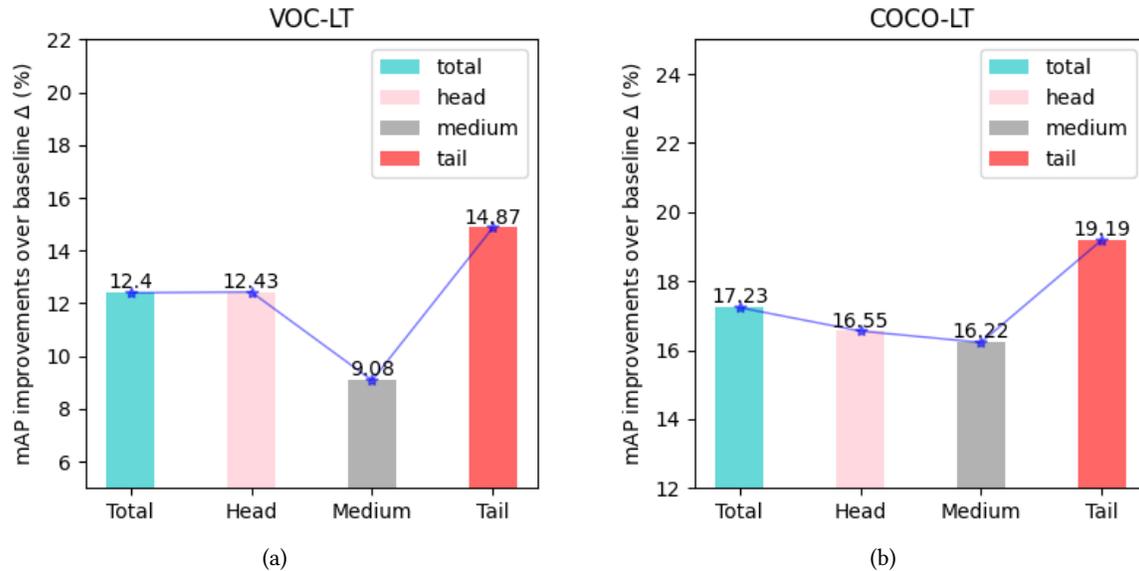


Figure 1: The mAP (%) improvements over baseline by integrating three components of the proposed CPRFL on VOC-LT and COCO-LT datasets.

visualization examples from these different models, with CPRFL demonstrating superior performance, particularly for the tail classes.

REFERENCES

- [1] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 770–778.
- [2] Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).
- [3] Ziwei Liu, Zhongqi Miao, Xiaohang Zhan, Jiayun Wang, Boqing Gong, and Stella X Yu. 2019. Large-scale long-tailed recognition in an open world. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2537–2546.
- [4] 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.
- [5] Ximeng Sun, Ping Hu, and Kate Saenko. 2022. Dualcoop: Fast adaptation to multi-label recognition with limited annotations. *Advances in Neural Information Processing Systems* 35 (2022), 30569–30582.
- [6] Peng Xia, Di Xu, Lie Ju, Ming Hu, Jun Chen, and Zongyuan Ge. 2023. LMPT: Prompt Tuning with Class-Specific Embedding Loss for Long-tailed Multi-Label Visual Recognition. *arXiv preprint arXiv:2305.04536* (2023).
- [7] Qizhe Xie, Zihang Dai, Eduard Hovy, Thang Luong, and Quoc Le. 2020. Unsupervised data augmentation for consistency training. *Advances in neural information processing systems* 33 (2020), 6256–6268.
- [8] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. 2022. Conditional prompt learning for vision-language models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 16816–16825.
- [9] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. 2022. Learning to prompt for vision-language models. *International Journal of Computer Vision* 130, 9 (2022), 2337–2348.

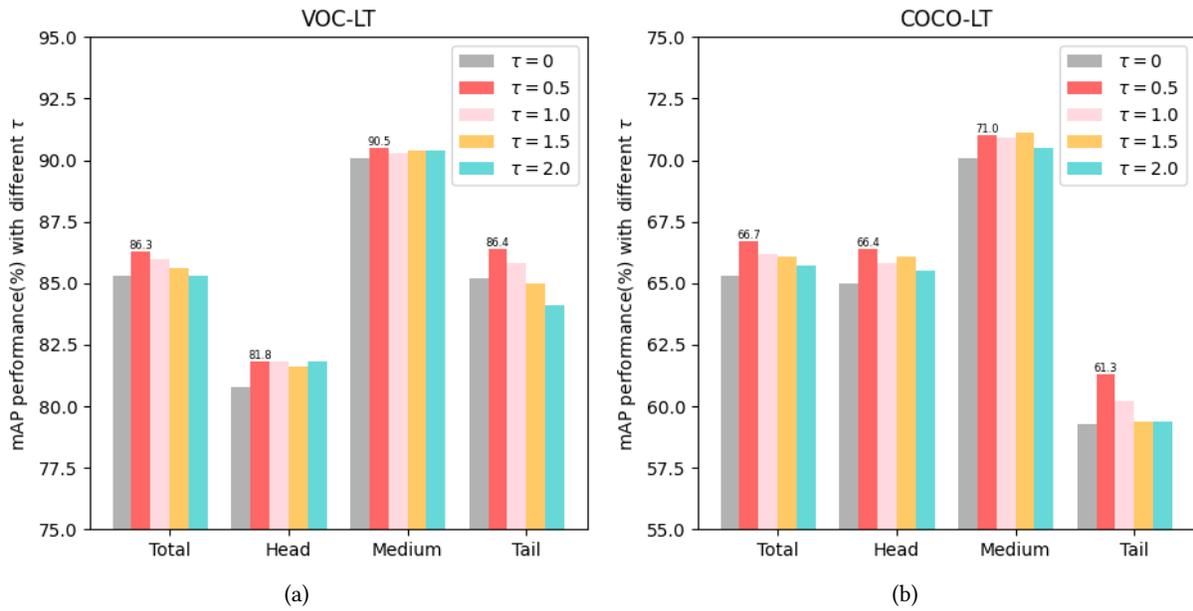


Figure 2: The mAP (%) performance with different expansion coefficient τ of on VOC-LT and COCO-LT datasets.

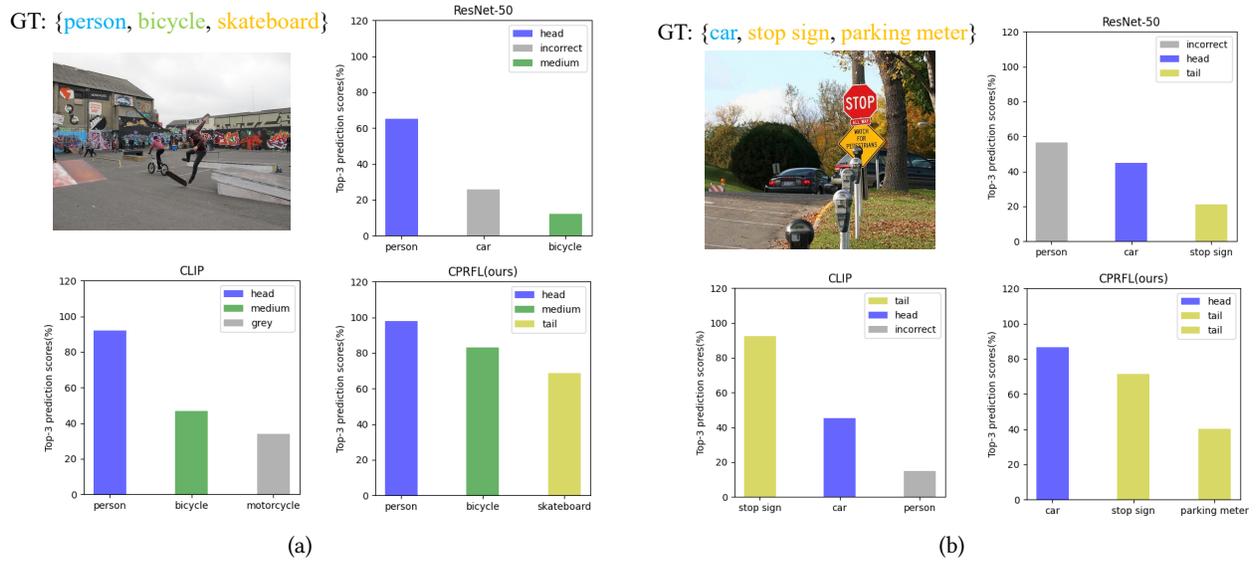


Figure 3: More visualization examples of Top-3 predicted categories by ResNet-50, CLIP and our CPRFL.