Causal Disentanglement and Cross-Modal Alignment for Enhanced Few-Shot Learning
Supplementary Material

Tianjiao Jiang, Zhen Zhang, Yuhang Liu*, Javen Qinfeng Shi
Australian Institute for Machine Learning, The University of Adelaide, Australia
{tianjiao.jiang, zhen.zhang02, yuhang.liu01, javen.shi}@adelaide.edu.au

6. Comparison with Other Methods

To allow a more comprehensive comparison with recent methods, we evaluated the performance of AMU-Tuning [3] and Transductive-CLIP [2] alongside our own approach. As shown in Tab. 6, our method consistently outperforms AMU-Tuning. Compared to Transductive-CLIP, our method achieves better performance in the 1-shot and 2-shot settings, but falls short in the 4-, 8-, and 16-shot settings.

It is important to note that Transductive-CLIP is highly sensitive to the number of classes in the query batch (denoted as k_{eff}); its performance can drop significantly with even a slight increase in k_{eff}. Moreover, as discussed in Related Work, AMU-Tuning incorporates an additional pre-trained model, MoCo v3 [1], alongside CLIP, rendering direct comparisons somewhat unfair. Similarly, Transductive-CLIP operates under a *transductive setting*, where predictions are made jointly for a batch of samples, leveraging inter-sample dependencies. In contrast, our method uses a more general *inductive setting*, performing inference independently for each sample. This difference in inference paradigm makes direct comparison inherently inequitable.

Table 6. Average few-shot classification accuracy (%) across 11 datasets for different methods.

Method	1-shot	2-shot	4-shot	8-shot	16-shot
AMU-Tuning	64.50	66.95	70.57	72.87	74.71
Transductive-CLIP ($k_{eff} = 5$)	65.25	68.58	73.77	77.98	81.25
Transductive-CLIP ($k_{eff} = 7$)	61.60	64.08	68.99	73.58	76.97
CCA-FT (Ours)	66.00	68.62	72.10	74.84	77.60

7. Ablation Study: Different ICA Dimensions

To assess the robustness of CCA-FT with respect to varying ICA dimensions M, we perform few-shot experiments for $M \in \{128, 256, 512, 1024\}$ and compare the results with Tip-Adapter-F. As illustrated in Fig. 6, CCA-FT consistently surpasses Tip-Adapter-F, even when the ICA dimension is as low as 128.

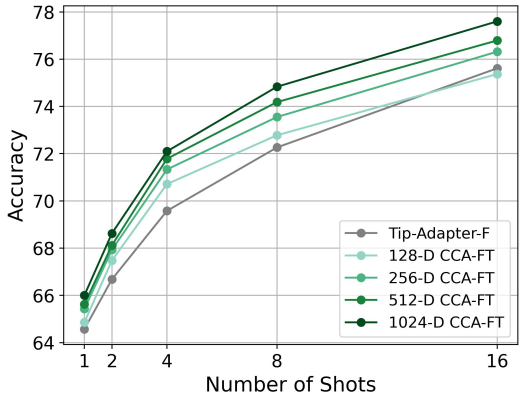

Figure 6. Average classification accuracy (%) of CCA-FT with varying feature dimensions M across 11 datasets, compared with Tip-Adapter-F.

References

[1] Xinlei Chen, Saining Xie, and Kaiming He. An empirical study of training self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9640–9649, 2021.

[2] Ségolène Martin, Yunshi Huang, Fereshteh Shakeri, Jean-Christophe Pesquet, and Ismail Ben Ayed. Transductive zero-shot and few-shot clip. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 28816–28826, 2024.

[3] Yuwei Tang, Zhenyi Lin, Qilong Wang, Pengfei Zhu, and Qinghua Hu. Amu-tuning: Effective logit bias for clip-based few-shot learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23323–23333, 2024.

*Corresponding author. Email: yuhang.liu01@adelaide.edu.au