

Supplementary Materials for "Visual-Language Collaborative Representation Network for Broad-Domain Few-Shot Image Classification"

Anonymous Author(s)
Submission Id: 1441

Due to the length constraints of the main text, we provide additional supplementary materials related to this paper. These materials primarily include 1) supplementary experiments in this PDF and 2) supplementary code in the zip file. We commit to making the code and the constructed dataset of this paper publicly available on GitHub upon final acceptance.

1 SUPPLEMENTARY COMPARISON

The comparison experiment results in the main text present a comparison between the state-of-the-art visual few-shot learning (FSL) models, the visual-language models (VLMs) based on CLIP, and our MCRNet. Due to space constraints, the complete results of the few-shot learning models were not listed, hence the need for this supplementary information. As shown in Tab. 1, meta-learning-based models like MAML, and metric-learning-based models like RelationNet, DeepDBC, DeepEMD, and RankDNN all yield lower results compared to our MCRNet. This further demonstrates the superiority of MCRNet. Our analysis reveals that the performance of FSL models generally improves with 5-shot compared to 1-shot, whereas the performance of VLMs, especially the baseline CLIP, shows little difference between 1-shot and 5-shot. This is because FSL models typically train classifiers online based on the prototype features of five support images, rather than using a matrix to determine the query category as VLMs do. In contrast, our MCRNet incorporates a category-adaptive fine-tuning mechanism, which reuses the features of the five support images and their category information during representation, enabling the learned features to acquire new category knowledge. As a result, MCRNet not only utilizes support information more effectively than VLMs but also delves deeper than FSL models by adjusting distributions beyond just training classifiers.

In addition, in FSL research, the debate over the merits and drawbacks of meta-learning, metric learning, and data augmentation methods has always been a focal point. In the supplementary experiments, we observed that in the extensive validation process across diverse domain datasets, methods based on metric learning such as RankDNN or DeepEMD perform better than other methods, especially in the 1-shot setting. However, as the core focus of this paper is not on evaluation experiments and analysis, we cannot simply conclude. In future work, we aim to expand our benchmark dataset tasks and domains to provide a detailed analysis of these FSL methods and VLMs.

2 VISUALIZATION

MCRNet is built upon CLIP, and in the main text, we assert that the key focus of MCRNet lies in its ability to re-represent the feature prototypes generated by CLIP. This is achieved through MCRNet's innovative mechanism that integrates fusion and representation,

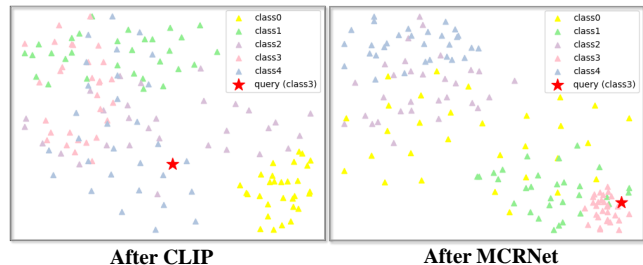


Figure 1: The feature visualization after CLIP and the feature visualization of our MCRNet are compared. It can be observed that MCRNet can correct the misrepresentation of query data by CLIP, leading to correct classification.

utilizing contrastive learning loss to supervise the category and distance relationships between the image-text features of queries and supports. Essentially, MCRNet brings similar supports and queries closer in multi-modal information and pushes dissimilar ones apart. To demonstrate this effect, we conducted feature visualizations as shown in Fig. 1. This comparison of feature visualizations was carried out on five classes in the LeafVirus dataset. Due to the high similarity between classes in LeafVirus, where the differences between categories are minimal, and these plant virus categories are unfamiliar to CLIP, CLIP erroneously classified a query belonging to class 3 into class 4. However, after processing through MCRNet, the query was correctly classified. This validates the effectiveness of MCRNet's re-representation.

REFERENCES

- [1] Chelsea Finn, Pieter Abbeel, and Sergey Levine. 2017. Model-Agnostic Meta-Learning for Fast Adaptation of Deep Networks. In *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017 (Proceedings of Machine Learning Research, Vol. 70)*, Doina Precup and Yee Whye Teh (Eds.). PMLR, 1126–1135. <http://proceedings.mlr.press/v70/finn17a.html>
- [2] Peng Gao, Shijie Geng, Renrui Zhang, Teli Ma, Rongyao Fang, Yongfeng Zhang, Hongsheng Li, and Yu Qiao. 2024. CLIP-Adapter: Better Vision-Language Models with Feature Adapters. *Int. J. Comput. Vis.* 132, 2 (2024), 581–595.
- [3] Qianyu Guo, Haotong Gong, Xujun Wei, Yanwei Fu, Yizhou Yu, Wenqiang Zhang, and Weifeng Ge. 2023. RankDNN: Learning to Rank for Few-Shot Learning. In *Thirty-Seventh AAAI Conference on Artificial Intelligence, AAAI 2023, Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence, IAAI 2023, Thirteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2023, Washington, DC, USA, February 7-14, 2023*, Brian Williams, Yiling Chen, and Jennifer Neville (Eds.). AAAI Press, 728–736. <https://doi.org/10.1609/AAAI.V37I1.25150>
- [4] Fusheng Hao, Fengxiang He, Liu Liu, Fuxiang Wu, Dacheng Tao, and Jun Cheng. 2023. Class-Aware Patch Embedding Adaptation for Few-Shot Image Classification. In *IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023*. IEEE, 18859–18869.
- [5] Yangji He, Weihang Liang, Dongyang Zhao, Hong-Yu Zhou, Weifeng Ge, Yizhou Yu, and Wenqiang Zhang. 2022. Attribute Surrogates Learning and Spectral Tokens Pooling in Transformers for Few-shot Learning. In *IEEE/CVF Conference*

Table 1: Experimental comparison results of MCRNet and SOTA models in the biological domain (Animal, Insect, and Mushroom) as well as in the agricultural domain (LeafVirus) on 5-way-1-shot and 5-way-5-shot settings. The numbers in bold indicate the best performance, while the underlined ones denote the second best. All the backbone of the following models is ViT.

Method	Animal		Insect		Mushroom		LeafVirus	
	1-shot	5-shot	1-shot	5-shot	1-shot	5-shot	1-shot	5-shot
<i>Unimodality Few-Shot Learning Models</i>								
MAML [ICML2017] [1]	28.06 \pm 0.55	29.01 \pm 0.53	27.83 \pm 0.51	29.88 \pm 0.47	25.65 \pm 0.54	26.09 \pm 0.51	46.42 \pm 0.95	48.05 \pm 0.84
ProtoNet [NIPS2017] [7]	30.67 \pm 0.58	39.92 \pm 0.81	29.15 \pm 0.50	44.28 \pm 0.71	25.72 \pm 0.49	38.31 \pm 0.71	51.92 \pm 0.85	71.48 \pm 0.87
RelationNet [CVPR2018] [8]	30.72 \pm 0.57	38.33 \pm 0.70	29.82 \pm 0.50	41.77 \pm 0.63	27.52 \pm 0.57	33.75 \pm 0.66	37.64 \pm 0.87	59.73 \pm 0.93
DeepDBC [CVPR2022] [9]	32.78 \pm 0.62	41.37 \pm 0.86	28.54 \pm 0.47	42.24 \pm 0.78	28.72 \pm 0.55	40.63 \pm 0.76	49.95 \pm 1.04	72.26 \pm 0.83
DeepEMD [TPAMI2023] [11]	34.54 \pm 0.63	42.24 \pm 0.82	35.78 \pm 0.63	49.52 \pm 0.75	31.91 \pm 0.63	44.22 \pm 0.77	59.19 \pm 1.04	79.36 \pm 1.00
RanKDDN [AAAI2023] [3]	35.03 \pm 0.50	40.92 \pm 0.25	37.99 \pm 0.60	46.53 \pm 0.53	32.50 \pm 0.55	45.35 \pm 0.67	62.44 \pm 0.80	77.42 \pm 0.70
FewTURE [CVPR2020] [10]	34.28 \pm 0.26	44.44 \pm 0.53	32.59 \pm 0.40	44.13 \pm 0.55	29.29 \pm 0.34	43.89 \pm 0.58	53.94 \pm 0.48	74.99 \pm 0.51
HTCTrans [CVPR2022] [5]	42.15 \pm 0.60	47.82 \pm 0.75	47.47 \pm 0.42	<u>59.03\pm0.63</u>	32.71 \pm 0.29	34.17 \pm 0.52	64.87 \pm 0.55	<u>82.92\pm0.48</u>
CPEA [ICCV2023] [4]	42.46 \pm 0.63	52.07 \pm 0.49	44.54 \pm 0.53	60.67 \pm 0.87	33.21 \pm 0.42	48.24 \pm 0.57	<u>65.94\pm0.39</u>	81.54 \pm 0.55
<i>Vision-Language Models</i>								
CLIP [ICML2021] [6]	73.61 \pm 0.25	74.40 \pm 0.32	20.67 \pm 0.37	20.79 \pm 0.33	45.58 \pm 0.35	46.23 \pm 0.35	35.59 \pm 0.40	34.64 \pm 0.34
Tip-Adapter [ECCV2022] [12]	74.06 \pm 0.47	75.38 \pm 0.49	23.10 \pm 0.56	36.68 \pm 0.53	44.25 \pm 0.46	47.99 \pm 0.41	39.91 \pm 0.44	47.24 \pm 0.55
CoOP [CVPR2022] [13]	<u>75.19\pm0.62</u>	75.23 \pm 0.69	20.02 \pm 0.71	19.98 \pm 0.89	46.24 \pm 0.72	45.30 \pm 0.70	33.32 \pm 0.62	35.29 \pm 0.58
APE-T [ICCV2023] [14]	74.80 \pm 0.47	<u>79.97\pm0.58</u>	21.33 \pm 0.62	21.02 \pm 0.58	48.60 \pm 0.30	48.97 \pm 0.34	39.75 \pm 0.57	41.00 \pm 0.59
CLIP-Adapter [IJCV2024] [2]	74.20 \pm 0.28	75.80 \pm 0.33	22.57 \pm 0.36	22.99 \pm 0.41	<u>49.85\pm0.51</u>	<u>52.17\pm0.69</u>	36.48 \pm 0.66	37.24 \pm 0.47
MCRNet (Ours)	75.86\pm0.54	84.33\pm0.72	70.27\pm0.76	81.09\pm0.40	51.25\pm0.35	64.97\pm0.88	70.79\pm0.68	88.87\pm0.67

on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18–24, 2022. IEEE, 9109–9119.

- [6] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning Transferable Visual Models From Natural Language Supervision. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18–24 July 2021, Virtual Event (Proceedings of Machine Learning Research, Vol. 139)*, Marina Meila and Tong Zhang (Eds.). PMLR, 8748–8763.
- [7] Jake Snell, Kevin Swersky, and Richard S. Zemel. 2017. Prototypical Networks for Few-shot Learning. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4–9, 2017, Long Beach, CA, USA*, Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett (Eds.), 4077–4087. <https://proceedings.neurips.cc/paper/2017/hash/cb8da6767461f2812ae4290eac7cbc42-Abstract.html>
- [8] Flood Sung, Yongxin Yang, Li Zhang, Tao Xiang, Philip H. S. Torr, and Timothy M. Hospedales. 2018. Learning to Compare: Relation Network for Few-Shot Learning. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18–22, 2018*. Computer Vision Foundation / IEEE Computer Society, 1199–1208. <https://doi.org/10.1109/CVPR.2018.00131>
- [9] Jiangtao Xie, Fei Long, Jiaming Lv, Qilong Wang, and Peihua Li. 2022. Joint Distribution Matters: Deep Brownian Distance Covariance for Few-Shot Classification. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18–24, 2022*. IEEE, 7962–7971. <https://doi.org/10.1109/CVPR52688.2022.00781>
- [10] Han-Jia Ye, Hexiang Hu, De-Chuan Zhan, and Fei Sha. 2020. Few-Shot Learning via Embedding Adaptation With Set-to-Set Functions. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13–19, 2020*. Computer Vision Foundation / IEEE, 8805–8814.
- [11] Chi Zhang, Yujun Cai, Guosheng Lin, and Chunhua Shen. 2023. DeepEMD: Differentiable Earth Mover’s Distance for Few-Shot Learning. *IEEE Trans. Pattern Anal. Mach. Intell.* 45, 5 (2023), 5632–5648. <https://doi.org/10.1109/TPAMI.2022.3217373>
- [12] Renrui Zhang, Wei Zhang, Rongyao Fang, Peng Gao, Kunchang Li, Jifeng Dai, Yu Qiao, and Hongsheng Li. 2022. Tip-Adapter: Training-Free Adaption of CLIP for Few-Shot Classification. In *Computer Vision - ECCV 2022 - 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXV (Lecture Notes in Computer Science, Vol. 13695)*, Shai Avidan, Gabriel J. Brostow, Moustapha Cissé, Giovanni Maria Farinella, and Tal Hassner (Eds.). Springer, 493–510.
- [13] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. 2022. Conditional Prompt Learning for Vision-Language Models. In *IEEE/CVF Conference*

on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18–24, 2022. IEEE, 16795–16804.

- [14] Xiangyang Zhu, Renrui Zhang, Bowei He, Aojun Zhou, Dong Wang, Bin Zhao, and Peng Gao. 2023. Not All Features Matter: Enhancing Few-shot CLIP with Adaptive Prior Refinement. In *IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1–6, 2023*. IEEE, 2605–2615. <https://doi.org/10.1109/ICCV51070.2023.00246>