

Supplementary Materials: KNN Transformer with Pyramid Prompts for Few-Shot Learning

Anonymous Authors

A APPENDIX

A.1 Comparison with the SOTA Methods in Fine-Grained Scenarios

Table 1: Results (%) on CUB-200-2011. Bold font indicates the best results. Blue font indicates the suboptimal results.

Method	Venue	CUB-200-2011 5-way	
		1-shot	5-shot
ProtoNet [14]	NeurIPS’2017	71.88 ± 0.91	87.42 ± 0.48
AM3 [19]	NeurIPS’2019	74.1	79.7
AFHN [10]	CVPR’2020	70.53 ± 1.01	83.95 ± 0.63
MixtFSL [1]	ICCV’2021	73.94 ± 1.10	86.01 ± 0.50
CCG+HAN [4]	ICCV’2021	74.66 ± 0.21	88.37 ± 0.12
RENet [8]	ICCV’2021	79.49 ± 0.44	91.11 ± 0.24
DC [20]	ICLR’2021	79.56 ± 0.87	90.67 ± 0.35
APP2S [13]	AAAI’2022	77.64 ± 0.19	90.43 ± 0.18
RankDNN [6]	AAAI’2023	82.93	91.47
BMI [12]	MM’2023	85.79 ± 0.33	90.91 ± 0.22
LastShot [21]	TPAMI’2024	80.20 ± 0.21	91.49 ± 0.12
KTPP	ours	87.30 ± 0.34	91.85 ± 0.24

To further evaluate the performance of our proposed KTPP method, experiments are conducted on the fine-grained dataset: CUB-200-2011 (CUB) [17], which contains 11,788 images across 200 bird subcategories, divided into 100 training, 50 validation, and 50 test categories.

Table 1 presents the fine-grained classification results on CUB. It can be observed that KTPP significantly surpasses the state-of-the-art methods on 1-shot and 5-shot tasks. Specifically, KTPP improves 1-shot accuracy by 1.51% and 5-shot accuracy by 0.32% compared to the state-of-the-art methods. This superiority can be attributed to the integration of semantically enhanced class-specific features, which play a crucial role in distinguishing subtle inter-class differences and capturing shared intra-class representations. Moreover, compared to semantic-based methods (AM3 [19], BMI [12]), KTPP also demonstrates substantial improvements, ranging from 0.94% to 13.20% on 1-shot and 5-shot tasks, due to full leverage of semantic information by deep and pyramid cross-modal interactions.

A.2 Comparison with the SOTA Methods in Cross-Domain Scenarios

To further validate the effectiveness of the proposed KTPP method across novel tasks, KTPP is evaluated in the challenging cross-domain miniImageNet → CUB scenario. Following the standard protocols [11, 18, 23], we utilize the training set comprising 64 classes from miniImageNet [16] as the source domain for training. Subsequently, we evaluate the generalization performance of KTPP on the novel CUB dataset, which serves as the target domain.

Table 2: Results (%) on the cross-domain miniImageNet → CUB. Bold font indicates the best results. Blue font indicates the suboptimal results.

Method	Venue	MiniImageNet → CUB	
		1-shot	5-shot
GNN [5]	ICLR’2018	45.69 ± 0.68	62.25 ± 0.65
FT [15]	ICLR’2020	47.47 ± 0.75	66.98 ± 0.68
ATA [18]	IJCAI’2021	45.00 ± 0.50	66.22 ± 0.50
T3S [22]	AAAI’2022	45.92	69.16
AFA [7]	ECCV’2022	46.86 ± 0.50	68.25 ± 0.50
RDC [11]	CVPR’2022	51.20 ± 0.50	67.77 ± 0.40
StyleAdv [3]	CVPR’2023	48.49 ± 0.72	68.72 ± 0.67
LDP-net [23]	CVPR’2023	49.82	70.39
ALFA [2]	TPAMI’2024	-	70.22 ± 0.14
KTPP	ours	61.75 ± 0.47	78.97 ± 0.37

The cross-domain classification results on miniImageNet → CUB are shown in Table 2. KTPP shows substantial superiority over its counterparts. Specifically, KTPP outperforms the state-of-the-art results by a significant margin, with improvements of 11.93% in 1-shot accuracy and 8.58% in 5-shot accuracy. These improvements demonstrate the excellent transferability and domain-agnostic capabilities of KTPP, which is attributed to the rapid adaptation from training to novel datasets by effectively capturing discriminative features based on semantic priors. Consequently, our model can generalize well to novel classes with limited labeled samples, even in the presence of domain shift.

A.3 Ablation Study on More Benchmarks

Table 3: Ablation study of KTPP on CIFAR-FS. “KCA” means K-NN Context Attention. “PCP” denotes Pyramid Cross-modal Prompts.

Training phase		Module		1-shot	5-shot
Pre	Meta	KCA	PCP		
✓				67.18 ± 0.55	83.57 ± 0.33
✓	✓			70.22 ± 0.54	84.75 ± 0.29
✓		✓		74.70 ± 0.57	87.82 ± 0.31
✓	✓	✓		78.74 ± 0.57	88.83 ± 0.31
✓	✓		✓	82.61 ± 0.50	88.67 ± 0.30
✓	✓	✓	✓	83.63 ± 0.57	90.19 ± 0.30

To comprehensively evaluate the effectiveness of KCA and PCP in the proposed KTPP method, we investigate their impact on another dataset: CIFAR-FS [9].

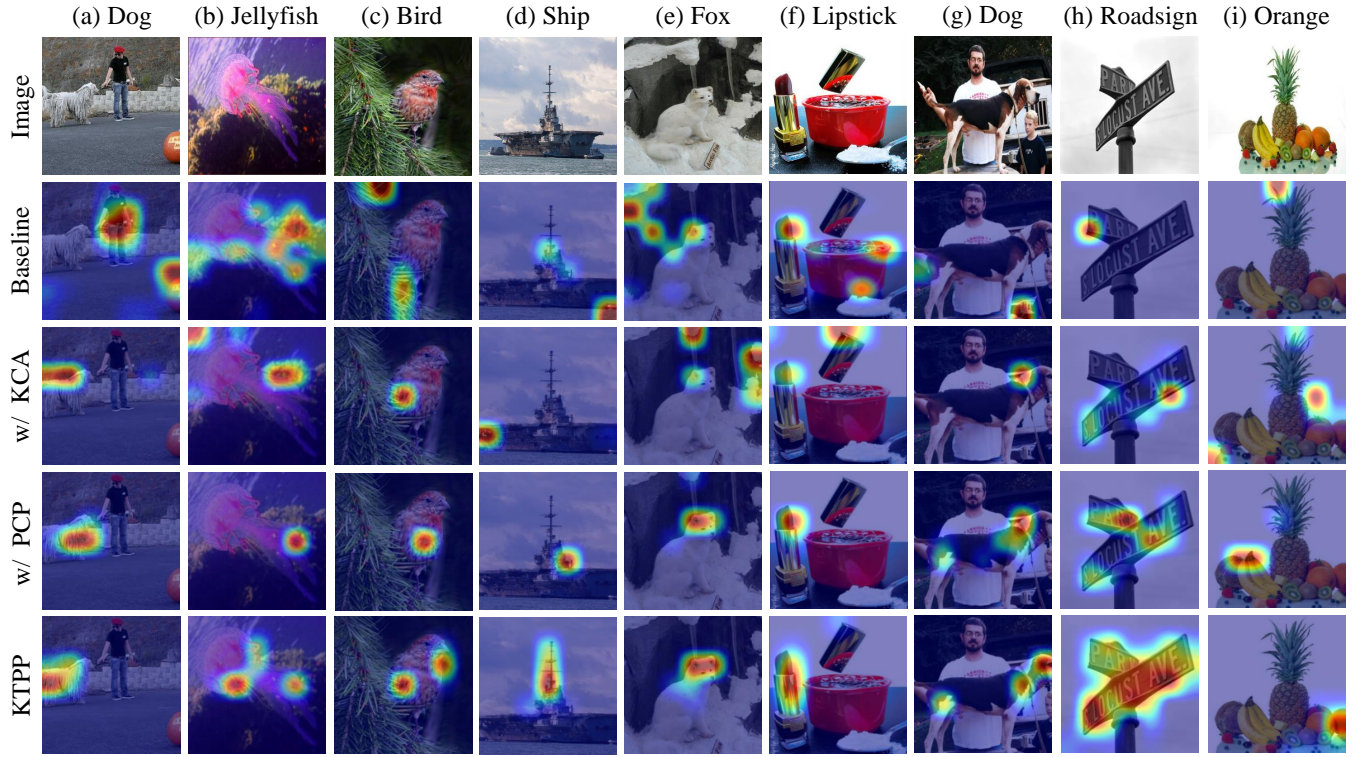


Figure 1: The visualization of attention maps obtained by different approaches. Our KTPP w/ KCA and PCP can effectively select class-specific regions and filter out noisy regions, compared with the baseline.

As shown in Table 3, the experimental results exhibit a similar tendency to the results on miniImageNet presented in Table 3 (manuscript). Firstly, KCA demonstrates significant improvements over the baseline, improving 1-shot accuracy by 8.52% and 5-shot accuracy by 4.08% on the meta phase, benefiting from the extraction of noise-free visual features via a coarse-fine manner. Secondly, PCP also surpasses the baseline by a large margin, with improvements of 12.39% in 1-shot accuracy and 3.92% in 5-shot accuracy. These improvements show the excellent capability of adaptively modulating visual features via deep and pyramid cross-modal interactions in PCP. Finally, by combining KCA and PCP, our model further achieves the most favorable outcomes due to semantically enhanced noisy-free visual representations, improving 1-shot accuracy by 13.41% and 5-shot accuracy by 5.44%.

A.4 Visualization of Attention Maps

To intuitively illustrate the effectiveness of the proposed KTPP method, we visualize attention maps generated by different approaches.

As shown in Fig. 1, we visualize four different approaches: baseline, w/ KCA, w/ PCP, and KTPP (i.e. w/ KCA and PCP). Firstly, in cluttered backgrounds and occlusions, the baseline easily leads to incorrect classifications such as (a), (c), (i), and (g). Moreover, the baseline is susceptible to spurious correlations, falsely associating other objects with the label, as observed in (b), (d), (e), and

(f). Secondly, in contrast to the baseline, the utilization of either KCA or PCP effectively mitigates the influence of class-irrelevant entities such as (a), (c), (d), (f), (g), and (h). This is because KCA can progressively filter out irrelevant features and PCP can adaptively adjust visual features based on semantic information. However, they still struggle with more complex scenarios when used individually, as observed in (b), (e), and (i). Finally, through the integration of KCA and PCP, i.e., KTPP, our model further enhances noise-free visual representations via deep cross-modal interactions, thereby effectively selecting discriminative regions and ignoring noisy background regions in the more complex scenarios.

REFERENCES

- [1] Arman Afrasiyabi, Jean-François Lalonde, and Christian Gagné. 2021. Mixture-Based Feature Space Learning for Few-Shot Image Classification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 9041–9051.
- [2] Sungyong Baik, Myungsub Choi, Janghoon Choi, Heewon Kim, and Kyoung Mu Lee. 2024. Learning to Learn Task-Adaptive Hyperparameters for Few-Shot Learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 46, 3 (2024), 1441–1454. <https://doi.org/10.1109/TPAMI.2023.3261387>
- [3] Yuqian Fu, Yu Xie, Yanwei Fu, and Yu-Gang Jiang. 2023. Styleadv: Meta style adversarial training for cross-domain few-shot learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 24575–24584.
- [4] Zhi Gao, Yuwei Wu, Yunde Jia, and Mehrtash Harandi. 2021. Curvature generation in curved spaces for few-shot learning. In *Proceedings of the IEEE/CVF international conference on computer vision*. 8691–8700.
- [5] Victor Garcia and Joan Bruna. 2018. Few-shot learning with graph neural networks. In *6th International Conference on Learning Representations, ICLR 2018*.
- [6] Qianyu Guo, Gong Haotong, Xujun Wei, Yanwei Fu, Yizhou Yu, Wenqiang Zhang, and Weifeng Ge. 2023. Rankdnn: Learning to rank for few-shot learning. In

- Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 37. 728–736.
- [7] Yanxu Hu and Andy J Ma. 2022. Adversarial feature augmentation for cross-domain few-shot classification. In *European Conference on Computer Vision*. Springer, 20–37.
- [8] Dahyun Kang, Heeseung Kwon, Juhong Min, and Minsu Cho. 2021. Relational embedding for few-shot classification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 8822–8833.
- [9] Kwonjoon Lee, Subhansu Maji, Avinash Ravichandran, and Stefano Soatto. 2019. Meta-learning with differentiable convex optimization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 10657–10665.
- [10] Kai Li, Yulun Zhang, Kunpeng Li, and Yun Fu. 2020. Adversarial feature hallucination networks for few-shot learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 13470–13479.
- [11] Pan Li, Shaogang Gong, Chengjie Wang, and Yanwei Fu. 2022. Ranking Distance Calibration for Cross-Domain Few-Shot Learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 9099–9108.
- [12] Zhuoling Li and Yong Wang. 2023. Better Integrating Vision and Semantics for Improving Few-shot Classification. In *Proceedings of the 31st ACM International Conference on Multimedia*. 4737–4746.
- [13] Rongkai Ma, Pengfei Fang, Tom Drummond, and Mehrtash Harandi. 2022. Adaptive poincaré point to set distance for few-shot classification. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 36. 1926–1934.
- [14] Jake Snell, Kevin Swersky, and Richard Zemel. 2017. Prototypical networks for few-shot learning. *NeurIPS* 30 (2017).
- [15] Hung-Yu Tseng, Hsin-Ying Lee, Jia-Bin Huang, and Ming-Hsuan Yang. 2020. Cross-Domain Few-Shot Classification via Learned Feature-Wise Transformation. In *International Conference on Learning Representations*.
- [16] Oriol Vinyals, Charles Blundell, Timothy Lillicrap, Daan Wierstra, et al. 2016. Matching networks for one shot learning. *Advances in neural information processing systems* 29 (2016).
- [17] Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. 2011. The caltech-ucsd birds-200-2011 dataset. (2011).
- [18] Haoqing Wang and Zhi-Hong Deng. 2021. Cross-Domain Few-Shot Classification via Adversarial Task Augmentation. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*, Zhi-Hua Zhou (Ed.). International Joint Conferences on Artificial Intelligence Organization, 1075–1081. <https://doi.org/10.24963/ijcai.2021/149> Main Track.
- [19] Chen Xing, Negar Rostamzadeh, Boris Oreshkin, and Pedro O O Pinheiro. 2019. Adaptive cross-modal few-shot learning. *Advances in Neural Information Processing Systems* 32 (2019).
- [20] Shuo Yang, Lu Liu, and Min Xu. 2021. Free Lunch for Few-shot Learning: Distribution Calibration. In *International Conference on Learning Representations*.
- [21] Han-Jia Ye, Lu Ming, De-Chuan Zhan, and Wei-Lun Chao. 2024. Few-Shot Learning With a Strong Teacher. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 46, 3 (2024), 1425–1440. <https://doi.org/10.1109/TPAMI.2022.3160362>
- [22] Wang Yuan, Zhizhong Zhang, Cong Wang, Haichuan Song, Yuan Xie, and Lizhuang Ma. 2022. Task-level self-supervision for cross-domain few-shot learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 36. 3215–3223.
- [23] Fei Zhou, Peng Wang, Lei Zhang, Wei Wei, and Yanning Zhang. 2023. Revisiting Prototypical Network for Cross Domain Few-Shot Learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 20061–20070.