

Enhancing Pre-trained ViTs for Downstream Task Adaptation: A Locality-Aware Prompt Learning Method

Anonymous Authors

Table 1: Classification datasets statistics.

Dataset	Description	Classes	Train	Test
Flowers102	Fine-grained	102	5726	2463
Stanford Cars		196	8144	8041
Aircraft		100	6667	3333
Stanford Dogs		120	12000	8580
CIFAR-10	Natural	10	50000	10000
CIFAR-100		100	50000	10000
DTD		47	3760	1880
ImageNet		1000	1281166	50000
EuroSAT	Specialized	10	18900	8100
Resisc45		45	6300	25200
Pattern		38	24320	6080
UCF		101	9537	3783

1 DATASETS STATISTICS

The comprehensive statistics of the classification datasets are presented in Table 1. For StanfordDogs, CIFAR-10, CIFAR-100, DTD, ImageNet, Resisc45, and Pattern, we followed the official dataset split strategy. For Flowers102, StanfordCars, Aircraft, EuroSAT, and UCF, we followed the split strategy used in CoOp [4].

For image retrieval datasets, ROxford5k and RParis6k [3] contain 4,993 and 6,322 high-resolution (1024×768) images, respectively, and each dataset has 70 queries from 11 landmarks.

For point correspondences dataset, SPair-71k [1] comprises 70,958 image pairs from 18 classes with diverse variations in viewpoint and scale, of which 53340 pairs serve as the training set, 5384 pairs serve as the validation set, and 12234 pairs serve as the test set.

For video object segmentation dataset, DAVIS [2] consists of 50 video sequences with 3455 densely annotated frames in pixel level. 30 videos with 2079 frames are for training, and 20 videos with 1376 frames are for validation.

REFERENCES

- [1] Juhong Min, Jongmin Lee, Jean Ponce, and Minsu Cho. 2019. Spair-71k: A large-scale benchmark for semantic correspondence. *arXiv preprint arXiv:1908.10543* (2019).
- [2] Jordi Pont-Tuset, Federico Perazzi, Sergi Caelles, Pablo Arbeláez, Alex Sorkine-Hornung, and Luc Van Gool. 2017. The 2017 davis challenge on video object segmentation. *CoRR* abs/1704.00675 (2017). arXiv:1704.00675
- [3] Filip Radenović, Ahmet Iscen, Giorgos Tolias, Yannis Avrithis, and Ondřej Chum. 2018. Revisiting oxford and paris: Large-scale image retrieval benchmarking. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 5706–5715.
- [4] 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.