Test-Time Retrieval-Augmented Adaptation for Vision-Language Models

Supplementary Material

1. Dataset Details

1.1. Cross-domain Benchamrk

The cross-domain (CD) benchmark is used to evaluate the method's transferability by testing on domains that diverge from the original training distribution. It contains 10 datasets, the details of which are shown below. Aircraft [8] consists of 3,333 images across 100 aircraft variants, with each variant characterized by specific manufacturer, model, and variant designations. Caltech101 [3] contains 2,465 images spanning 100 distinct object categories. Stanford Cars [7] comprises 8,041 images of 196 car classes, categorized by Make, Model, and Year. Describable textures dataset (DTD) [2] features 1,692 images organized into 47 texture categories, each described by human-selected adjectives. EuroSAT [4] contains 8,100 labeled satellite images across 10 land use and land cover classes. Oxford Flower102 [9] includes 2,463 images of 102 flower categories common in the United Kingdom. Food101 [1] consists of 30,300 images of food. Oxford-IIIT Pets [10] contains 3,669 images of 37 pet categories (12 cat breeds and 25 dog breeds). SUN397 [14] encompasses 397 scene categories with approximately 19,850 images total. UCF101 [12] is an action recognition dataset comprising 3,783 images across 101 action categories. The number of classes and the size of the test set are shown in the upper part of Table 1.

1.2. Out-of-distribution Benchamrk

The out-of-distribution (OOD) benchmark is used to evaluate robustness to distribution shifts within the same general domain. It contains 4 ImageNet-varients. ImageNet-A [6] contains 7,500 images from 200 classes. It features naturally perturbed ImageNet images that are visually similar but challenging to classify. ImageNet-V2 [11] is a collection of 10,000 images across 1,000 ImageNet classes. It was created by implementing an enhanced natural data collection pipeline on the original ImageNet dataset. ImageNet-R [5] encompasses 30,000 images from 200 ImageNet categories. It's distinguished by its diverse artistic renditions of the original ImageNet content. ImageNet-S [13] is composed of 50,000 sketches representing 1000 class objects from the ImageNet dataset. It exemplifies a domain shift from natural photographic images to hand-drawn sketches. The number of classes and the size of the test set are shown at the bottom part in Table 1.

Table 1. Summary of the statistics of all datasets of the cross-domain and out-of-distribution benchmarks.

Dataset	**Classes**	**Test size**
Aircraft	100	3333
Caltech101	100	2465
Cars	196	8041
DTD	47	1692
EuroSAT	10	8100
Flowers102	102	2463
Food101	101	30300
Pets	37	3669
SUN397	397	19850
UCF101	101	3783
ImageNet-A	200	7,500
ImageNet-V2	1,000	10,000
ImageNet-R	200	30,000
ImageNet-S	1,000	50,000

2. Test-time Computational Cost

The test-time computational costs, including speed and peak GPU memory usage on the Aircraft dataset, are shown in Table 2. Compared with the training-based method TPT, our TT-RAA is much faster and uses much less GPU memory. Compared with the training-free method TDA, our TT-RAA has similar speed and limited GPU usage overhead. Although our method introduces new components compared to TDA, they are fully vectorized and benefit from GPU acceleration. The additional GPU memory usage primarily comes from the storage of covariance matrices.

Table 2. Test-time computational cost. Our TT-RAA is much faster and uses much less GPU memory than the training-based method, TPT. Although we introduced new components, our method is fully vectorized and thus has similar speed compared with TDA.

Method	**Speed** (ms/sample)	**GPU Usage** (MB)
TPT	103	2213.53
TDA	12.76	348.09
TT-RAA	12.93	535.10

3. Experiments on other Vision-Language Models

To validate the generalizability of our TT-RAA, we conduct experiments on other VLMs, SigLiP, ALIGN, and FLAVA, on the CD benchmark. Figure 1 shows our method TT-RAA consistently shows improvements on these models.

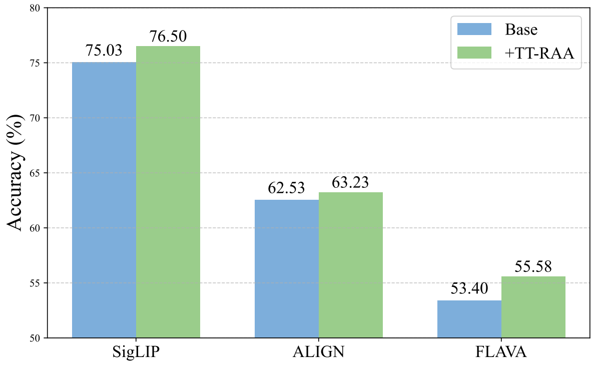

Figure 1. Results of SigLip, ALIGN, and FLAVA on the CD benchmark with and without TT-RAA. Our TT-RAA consistently shows improvements on these models. Base refers to their original models without TT-RAA.

4. Detailed Component Analysis of Multimodal Retrieval Augmentation

Detailed component analysis on vision space similarity retrieval (VSSR), vision space discriminant analysis (VSDA), and multimodal space retrieval augmentation (MSRA) inside the MRA on the CD benchmark is shown in Table 3. If SMGD is not used, TDA with one-shot capacity is employed as a dynamic database for the retrieval needs of MRA. VSDA cannot function without the covariance data provided by SMGD. The results show the effectiveness of each component and the synergy when combined.

5. Limitations

Despite the promising results demonstrated by TT-RAA, several limitations and challenges warrant discussion. The adaptation stability can vary significantly depending on target domain characteristics, particularly in scenarios with extreme distribution shifts or limited visual diversity within categories. This instability is especially evident in fine-grained classification tasks such as the Aircraft and Cars datasets, where subtle visual differences are crucial for accurate categorization. Furthermore, the method's performance is sensitive to several hyperparameters, including the update coefficient which may require domain-specific tuning. This hyperparameter dependency potentially limits the application of TT-RAA in real-world scenarios where optimal parameter selection might not be feasible.

Table 3. Detailed component analysis of the multimodal retrieval augmentation module. It contains vision space similarity retrieval (VSSR), vision space discriminant analysis (VSDA), and multimodal space retrieval augmentation (MSRA). The results show the effectiveness of each component and the synergy when combined.

SMGD	VSSR	VSDA	MSRA	CD Average
-	-	-	-	64.59
-	✓	-	-	67.04
-	-	-	✓	65.85
-	✓	-	✓	67.49
✓	✓	-	-	68.46
✓	-	-	✓	65.84
✓	-	✓	-	66.23
✓	✓	✓	-	68.53
✓	✓	✓	✓	68.76

References

[1] Lukas Bossard, Matthieu Guillaumin, et al. Food-101: Mining discriminative components with random forests. In *European Conference on Computer Vision*, 2014. 1

[2] M. Cimpoi, S. Maji, I. Kokkinos, S. Mohamed, , and A. Vedaldi. Describing textures in the wild. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2014. 1

[3] Li Fei-Fei, Rob Fergus, and Pietro Perona. Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2004. 1

[4] Patrick Helber, Benjamin Bischke, et al. EuroSAT: A novel dataset and deep learning benchmark for land use and land cover classification. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 2019. 1

[5] Dan Hendrycks, Steven Basart, Norman Mu, et al. The many faces of robustness: A critical analysis of out-of-distribution generalization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021. 1

[6] Dan Hendrycks, Kevin Zhao, Steven Basart, Jacob Steinhardt, and Dawn Song. Natural adversarial examples. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15262–15271, 2021. 1

[7] Jonathan Krause, Michael Stark, et al. 3D object representations for fine-grained categorization. In *IEEE International Conference on Computer Vision Workshops*, 2013. 1

[8] Subhransu Maji, Esa Rahtu, Juho Kannala, Matthew Blaschko, et al. Fine-grained visual classification of aircraft. *arXiv preprint arXiv:1306.5151*, 2013. 1

[9] Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. In *Indian Conference on Computer Vision, Graphics and Image Processing*, 2008. 1

[10] Omkar M. Parkhi, Andrea Vedaldi, Andrew Zisserman, and C. V. Jawahar. Cats and dogs. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2012. 1

[11] Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, et al. Do imagenet classifiers generalize to imagenet? In *International Conference on Machine Learning*, 2019. 1

[12] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. UCF101: A dataset of 101 human actions classes from videos in the wild. *CoRR*, abs/1212.0402, 2012. 1

[13] Haohan Wang, Songwei Ge, et al. Learning robust global representations by penalizing local predictive power. In *Advances in Neural Information Processing Systems*, 2019. 1

[14] Jianxiong Xiao, James Hays, Krista A Ehinger, Aude Oliva, and Antonio Torralba. SUN database: Large-scale scene recognition from abbey to zoo. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2010. 1