Efficient Test-Time Adaptation of Vision-Language Models (Appendix)

Adilbek Karmanov[1*] Dayan Guan[2*] Shijian Lu[1,2†] Abdulmotaleb El Saddik[1,3] Eric Xing[1,4]
[1]Mohamed bin Zayed University of Artificial Intelligence [2]Nanyang Technological University
[3]University of Ottawa [4]Carnegie Mellon University

1. Benchmark Details

This section provides detailed information on the two benchmarks used in our work. **OOD Benchmark** is used to evaluate the model robustness against natural distribution shifts using the traditional ImageNet and its out-of-distribution (OOD) versions containing images with varying styles and corruptions. Herein below, we provide a concise overview of each of the OOD datasets.

- **ImageNet-V2** [11] consists of 10,000 images and 1,000 ImageNet classes, and was collected by applying an updated natural data collection pipeline to the original ImageNet data.
- **ImageNet-A** [6] is a subset of 7,500 visually similar but naturally perturbed ImageNet images of 200 classes.
- **ImageNet-R** [5] includes 30,000 images belonging to 200 categories of the ImageNet dataset, but with diverse artistic styles.
- **ImageNet-S** [13] consists of 50,000 sketches of 1000 class objects from the ImageNet dataset, and represents a domain shift from natural images to sketches.

Cross-Domain Benchmark consists of 10 image classification datasets to evaluate the effectiveness of the method on different domains. This benchmark incorporates the following datasets: Caltech101 [3] for general image classification, OxfordPets (Pets) [10], StanfordCars (Cars) [7], Flowers102 [9], Food101 [1], and FGVCAircraft (Aircraft) [8] for fine-grained image classification, EuroSAT [4] for satellite image classification, UCF101 [12] for action classification, DTD [2] for texture classification, and SUN397 [14] for scene classification.

The detailed statistics of all the datasets are shown in Table 1.

Dataset	Classes	Test Size
ImageNet	1,000	50,000
ImageNet-V2	1,000	10,000
ImageNet-S	1,000	50,000
ImageNet-A	200	7,500
ImageNet-R	200	30,000
Aircraft	100	3,333
Caltech101	100	2,465
Cars	196	8,041
DTD	47	1,692
EuroSAT	10	8,100
Flowers102	102	2,463
Food101	101	30,300
Pets	37	3,669
SUN397	397	19,850
UCF101	101	3,783

Table 1. Datasets statistics.

2. Parameter Studies on Thresholds

This section provides more parameter studies on three thresholds defined in our work. Our experiments are conducted on the ImageNet validation set using the default settings.

The threshold for negative pseudo-labeling. In Eq 4 of our manuscript, the threshold p_l is used to select negative pseudo labels by applying the negative mask. We conduct parameter studies on p_l and the results are illustrated in Figure 1. The best performance is achieved when p_l is set to 0.03 and subsequent increases in p_l do not yield notable improvement or degradation in the results, indicating the stability of this parameter. It can be noticed that the performance deteriorates when p_l is less than 0.03 because the confident classes with low probabilities should not be included in negative pseudo labels.

The threshold range for testing feature selection in the negative cache. In Eq 5 of our manuscript, the thresholds $[\tau_l, \tau_h]$ are used to check whether the testing feature

*Equal Contribution.
†Corresponding Author.

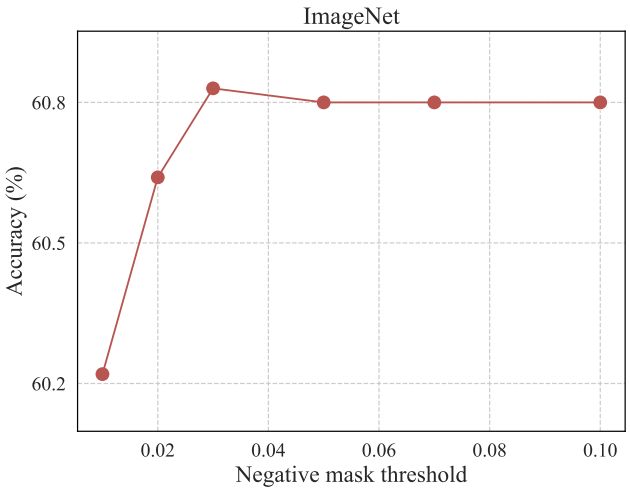

Figure 1. Parameter studies on the *Negative Mask Threshold* p_l for the negative pseudo-labeling in *Negative Cache*. The results are reported on ImageNet top-1 accuracy using only the *Negative Cache* to produce an adapted prediction. The experiments are conducted with CLIP-ResNet50.

will be considered to be included in *Negative Cache* if the entropy of the prediction is in the specified interval. Table 2 presents the results of an ablation study focusing on the impact of adjusting the threshold range for testing feature selection in the *Negative Cache*. This investigation delves into the testing feature selection of uncertain samples using two distinct approaches: one involving values closer to the minimum range threshold (τ_l) and the other to the maximum range threshold (τ_h), achieved by reversing the second condition: $\mathrm{H}(f_{\text{test}}\mathbf{W}_c^T) < \mathrm{H}(\tilde{\mathbf{q}}^{ent}\mathbf{W}c^T)$ or $\mathrm{H}(f_{\text{test}}\mathbf{W}_c^T) > \mathrm{H}(\tilde{\mathbf{q}}^{ent}\mathbf{W}_c^T)$. By collecting the lowest entropy features within the range [0.2, 0.5], the highest result is attained 60.83%. Opting values below 0.2 indicates the selection of confident samples for the *Negative Cache*, resulting in a reduction in the confidence of CLIP's prediction. Furthermore, a shift by 0.1 from [0.2, 0.5] to [0.3, 0.6] in the thresholds leads to an inclusion of more noisy samples during the early collection phase of the negative cache, resulting in a 0.48% decrease in performance. Selecting maximum entropy features with the same threshold range displays a slight decline in performance compared to the minimum entropy feature selection within the specified range. Hence, the most valuable uncertain samples fall within the [0.2, 0.5] range, whose entropy is closer to 0.2. The reported results exclusively utilize the *Negative Cache* to generate an adapted prediction.

The residual and sharpness ratios. The experiments in Table 3 show that the optimal residual ratio is 2.0 for TDA (instead of 1.0 in Tip-Adapter), indicating a higher significance of adapted features compared with CLIP features in test-time adaptation. The optimal sharpness ratio for TDA is 5.0, which is close to the 5.5 in Tip-Adapter.

Minimum entropy features			Maximum entropy features		
τ_l	τ_h	Accuracy	τ_l	τ_h	Accuracy
0.0	1.0	60.69	0.2	0.4	60.51
0.0	0.2	60.67	**0.2**	**0.5**	**60.53**
0.0	0.3	60.69	0.2	0.6	60.51
0.1	0.3	60.76	0.3	0.5	60.30
0.1	0.4	60.77	0.3	0.7	60.30
0.2	0.4	60.81	0.4	0.6	60.16
0.2	**0.5**	**60.83**	0.4	0.7	60.16
0.2	0.6	60.81	0.5	0.7	60.34
0.3	0.5	60.34	0.5	0.8	60.34
0.3	0.6	60.35	0.6	0.8	60.35

Table 2. Ablation study of the impact of varying *Threshold Range* $[\tau_l, \tau_h]$ for testing feature selection in the *Negative Cache*. The study investigates the testing feature selection of the uncertain samples in two ways: choosing the minimum and maximum entropy features in the given range. The results are reported on ImageNet top-1 accuracy using only the *Negative Cache* to produce an adapted prediction. The experiments are conducted with CLIP-ResNet50.

Residual Ratio	0.5	1.0	**2.0**	3.0	4.0	5.0
TDA	61.07	61.20	**61.35**	61.29	60.90	60.63
Sharpness Ratio	0.5	1.0	3.0	**5.0**	7.0	9.0
TDA	60.98	61.20	61.29	**61.35**	61.20	61.19

Table 3. Analysis on the residual and sharpness ratios of TDA.

3. More Experimental Analysis

Caches built for inference. The caches are built on the fly during inference, starting empty and progressively accumulating samples. At the start of the testing phase on ImageNet, where only 1% of the data was used, we observed a slight accuracy drop of 0.06%. We also noted that bypassing cache usage at the early testing phase leads to a marginal accuracy improvement of 0.1%. We didn't adopt this approach as it increases complexity by introducing an extra hyperparameter for determining when to use caches.

Class imbalance under high shot capacity. Our analysis with a 6-shot positive cache reveals minimal class imbalance (only 4 out of 1000 ImageNet classes have less than 6 samples) but identifies a significant cache accuracy drop

from 90.3% to 86.6% when shot capacity increases from 3 to 6. Such accuracy drop happens because larger cache capacities tend to accumulate noise, thereby reducing the reliability of cached labels and negatively affecting the adapted predictions. Hence, the performance decline with larger caches is mainly due to noise accumulation rather than class imbalance.

4. Broader Impact

The broader impact of test-time adaptation of vision-language models lies in its potential to enhance real-world applicability, improve accessibility and inclusivity, address bias and fairness concerns, and advance research and development. By allowing models to adapt to new, unseen data during inference, these models can be more versatile and adaptable, benefiting various domains such as healthcare and assistive technologies. Test-time adaptation also offers opportunities to mitigate biases, personalize user experiences, and push the boundaries of what vision-language models can achieve. However, ethical considerations must be taken into account to ensure responsible development and deployment, ensuring transparency, fairness, and accountability in the adaptation process.

References

[1] Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. 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 *Proceedings of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 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 *CVPR Workshops*, 2004. 1

[4] Patrick Helber, Benjamin Bischke, Andreas Dengel, and Damian Borth. Eurosat: A novel dataset and deep learning benchmark for land use and land cover classification. *IEEE J. Sel. Top. Appl. Earth Obs. Remote. Sens.*, 2019. 1

[5] Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, Dawn Song, Jacob Steinhardt, and Justin Gilmer. The many faces of robustness: A critical analysis of out-of-distribution generalization. In *ICCV*, 2021. 1

[6] Dan Hendrycks, Kevin Zhao, Steven Basart, Jacob Steinhardt, and Dawn Song. Natural adversarial examples. In *CVPR*, pages 15262–15271, 2021. 1

[7] Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d object representations for fine-grained categorization. In *ICCV Workshops*, 2013. 1

[8] S. Maji, J. Kannala, E. Rahtu, M. Blaschko, and A. Vedaldi. Fine-grained visual classification of aircraft. Technical report, 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 *CVPR*, 2012. 1

[11] Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do imagenet classifiers generalize to imagenet? In *ICML*, 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, Zachary Lipton, and Eric P Xing. Learning robust global representations by penalizing local predictive power. In *NeurIPS*, 2019. 1

[14] J. Xiao, J. Hays, K. A. Ehinger, A. Oliva, and A. Torralba. Sun database: Large-scale scene recognition from abbey to zoo. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2010. 1