# Free on the Fly: Enhancing Flexibility in Test-Time Adaptation with Online EM

## Supplementary Material

This section presents a thorough overview covering dataset specifications, implementation details, and additional experimental results. The content is organized as follows:



## 1. Dataset Details

| Dataset | # Classes | # Images |
|---|---|---|
| FGVCAircraft [11] | 100 | 3,333 |
| Caltech101 [4] | 100 | 2,465 |
| StanfordCars [10] | 196 | 8,041 |
| DTD [2] | 47 | 1,880 |
| EuroSAT [5] | 10 | 5,400 |
| Flowers102 [12] | 102 | 2,463 |
| Food101 [1] | 101 | 25,250 |
| OxfordPets [13] | 37 | 3,680 |
| SUN397 [16] | 397 | 19,850 |
| UCF101 [9] | 101 | 3,783 |
| ImageNet [3] | 1,000 | 50,000 |
| ImageNet-A [7] | 200 | 7,500 |
| ImageNetV2 [14] | 1,000 | 10,000 |
| ImageNet-R [6] | 200 | 30,000 |
| ImageNet-Sketch [15] | 1,000 | 50,889 |

Table 1. Dataset information for Cross-Domain benchmark and Out-of-Distribution benchmark.

This section provides an overview of the datasets used in our experiments, detailing the number of classes and the size of the test sets for each dataset. Table 1 summarizes these statistics, including widely-used benchmarks such as ImageNet, Caltech101, and StanfordCars, alongside specialized datasets like ImageNet-A and ImageNet-Sketch.

## 2. Implementation Details

To compute the inverse covariance matrix in a numerically stable manner, we adopt a regularization-based approach. Given the sample covariance matrix estimate $\hat{\Sigma}_{\text{reg}}$, directly inverting it can lead to instability, particularly in cases where the number of samples $N$ is smaller than the feature dimension $D$, or when the covariance matrix is ill-conditioned. To address these issues, we employ a regularized formulation. The regularized covariance matrix is defined as:

$$\hat{\Sigma}_{\text{reg}} = (N-1)\hat{\Sigma} + \text{tr}(\hat{\Sigma})I_D,$$

where $N$ is the number of samples, $\hat{\Sigma}_{\text{reg}}$ is the sample covariance matrix, $tr(\hat{\Sigma}_{\text{reg}})$ represents the trace of the matrix $\hat{\Sigma}_{\text{reg}}$, and $I_D$ is the $D$-dimension identity matrix. This formulation incorporates two components:(1) the scaled covariance matrix $(N-1)\hat{\Sigma}_{\text{reg}}$, which emphasizes the sample covariance structure; (2) the trace-based regularization term $tr(\hat{\Sigma}_{\text{reg}})I_D$, which enhances numerical stability by increasing the diagonal dominance of $\hat{\Sigma}_{\text{reg}}$. The inverse of the covariance matrix is then computed as:

$$\hat{\Sigma}^{-1} = D\left(\hat{\Sigma}_{\text{reg}}\right)^{-1},$$

Then we have:

$$\hat{\Sigma}^{-1} = D\left((N-1)\hat{\Sigma} + \text{tr}(\hat{\Sigma})I_D\right)^{-1}.$$

This formulation ensures that the inverse covariance matrix remains well-defined and numerically stable even in high-dimensional, small-sample regimes. The trace-based regularization effectively adjusts the off-diagonal elements, mitigating the impact of small eigenvalues that could otherwise lead to instability. By applying this regularization scheme, we achieve a robust estimation of $\Sigma^{-1}$ that is suitable for our Gaussian discriminant analysis for TTA.

## 3. Qualitative Analysis

In Figure 1, we present the t-SNE visualizations comparing the classifier weights derived from the zero-shot CLIP text encoder with those generated by our FreeTTA method. Each point represents the classifier weight for a specific class. Figure 1a and Figure 1b illustrate the results on the Food101 dataset, where (a) represents the zero-shot CLIP classifier weights and (b) shows the classifier weights after applying our FreeTTA approach using 80% of the test set data for estimation. Similarly, Figure 1c and Figure 1d provide the t-SNE results for the UCF101 dataset, with (c) representing zero-shot CLIP and (d) showing FreeTTA. Notably, the regions highlighted by the red boxes in Figure 1a and Figure 1c show classifier weights in zero-shot CLIP that are closely clustered, indicating the presence of indistinguishable classes with highly similar decision boundaries. In contrast, the dynamically optimized weights generated

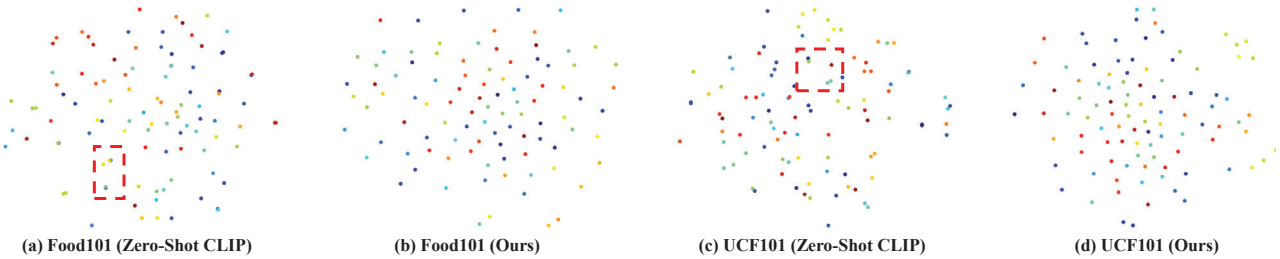


Figure 1. t-SNE visualizations of classifier weights for two datasets. (a) and (c) represent the zero-shot CLIP text encoder weights for the Food101 and UCF101 datasets, respectively. (b) and (d) depict our FreeTTA classifier weights estimated after processing 80% of the test data, and demonstrate the improved discriminative ability and robustness of the weights obtained through FreeTTA. The red boxes highlight regions in (a) and (c) where zero-shot CLIP shows several indistinguishable classes with highly similar decision boundaries, emphasizing the difficulty in separation.

by our FreeTTA method, as shown in Figure 1b and Figure 1d, exhibit better separation and improved discriminative ability. This highlights the effectiveness of our approach in modeling the target domain to dynamically refine classifier weights, thereby producing more robust and accurate decision boundaries.

Our method demonstrates significantly better discrimination in both datasets, especially in semantically challenging categories. The comparison highlights the effectiveness of FreeTTA in refining classifier weights, leading to more accurate and robust decision boundaries, and ultimately enhancing performance in test-time adaptation.

## 4. Computational Efficiency

| Method | Testing Time | Accuracy |
|---|---|---|
| TPT | ∼10h | 68.98 |
| TDA | ∼13min | 69.51 |
| Ours FreeTTA | ∼**30min** | **70.21** |

Table 2. Comparison of computational efficiency.

Table 2 compares the computational efficiency of our proposed FreeTTA method with the classic TPT on the ImageNet [3] dataset. As shown, our FreeTTA significantly reduces the testing time from approximately 10 hours to 2 hours while achieving superior accuracy (70.21% compared to 68.98%). This demonstrates the capability of FreeTTA to balance computational cost and performance, making it a practical choice for real-world applications with time constraints. However, FreeTTA remains twice as slow as optimization-free TDA [8], mainly due to matrix inversion operations. Enhanced parallel processing may improve this efficiency.

## 5. Limitations and Broader Impacts

A limitation of our method lies in its reduced impact under minimal domain shifts, where dynamic adaptation may not provide substantial accuracy improvements over other approaches. Additionally, test samples with high ambiguity or excessive noise can challenge our entropy-based uncertainty weighting mechanism, potentially leading to unstable parameter updates. The lack of labeled data in test domains further complicates the evaluation of adaptation accuracy, making it difficult to fully assess performance in certain scenarios.

While the proposed method offers an efficient, training-free solution for test-time adaptation, it remains reliant on pre-trained VLMs like CLIP. As a result, any inherent biases in the pre-trained model could inadvertently influence the adaptation process and propagate to downstream tasks. To address these concerns, future work may explore advanced uncertainty modeling and bias mitigation techniques to enhance robustness and fairness across diverse applications.

## References

[1] Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Food-101–mining discriminative components with random forests. In *Computer vision–ECCV 2014: 13th European conference, zurich, Switzerland, September 6-12, 2014, proceedings, part VI 13*, pages 446–461. Springer, 2014. 1

[2] Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. Describing textures in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3606–3613, 2014. 1

[3] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 1, 2

[4] 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

*2004 conference on computer vision and pattern recognition workshop*, pages 178–178. IEEE, 2004. 1

[5] 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 Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 12(7):2217–2226, 2019. 1

[6] Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, 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*, pages 8340–8349, 2021. 1

[7] 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

[8] Adilbek Karmanov, Dayan Guan, Shijian Lu, Abdulmotaleb El Saddik, and Eric Xing. Efficient test-time adaptation of vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14162–14171, 2024. 2

[9] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, et al. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*, 2017. 1

[10] Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d object representations for fine-grained categorization. In *Proceedings of the IEEE international conference on computer vision workshops*, pages 554–561, 2013. 1

[11] Subhransu Maji, Esa Rahtu, Juho Kannala, Matthew Blaschko, and Andrea Vedaldi. Fine-grained visual classification of aircraft. *arXiv preprint arXiv:1306.5151*, 2013. 1

[12] Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. In *2008 Sixth Indian conference on computer vision, graphics & image processing*, pages 722–729. IEEE, 2008. 1

[13] Omkar M Parkhi, Andrea Vedaldi, Andrew Zisserman, and CV Jawahar. Cats and dogs. In *2012 IEEE conference on computer vision and pattern recognition*, pages 3498–3505. IEEE, 2012. 1

[14] Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do imagenet classifiers generalize to imagenet? In *International conference on machine learning*, pages 5389–5400. PMLR, 2019. 1

[15] Haohan Wang, Songwei Ge, Zachary Lipton, and Eric P Xing. Learning robust global representations by penalizing local predictive power. *Advances in Neural Information Processing Systems*, 32, 2019. 1

[16] Jianxiong Xiao, James Hays, Krista A Ehinger, Aude Oliva, and Antonio Torralba. Sun database: Large-scale scene recognition from abbey to zoo. In *2010 IEEE computer society conference on computer vision and pattern recognition*, pages 3485–3492. IEEE, 2010. 1