

Training-Free Test-Time Adaptation via Shape and Style Guidance for Vision-Language Models

Appendix

1 A Detailed Theoretical Analysis about Generalizable Factors.

2 Following the previous works [1, 2], we give the theoretical analysis of the generalizable factors’
 3 important roles under the training-free test-time adaptation in vision-language models. There is a
 4 disentangled latent vector $\mathbf{v}(x) = (\mathbf{v}_1(x), \dots, \mathbf{v}_{d_v}(x))$ corresponding to the input image x_{test} and
 5 feature f_{test} . For convenience, we assume $\mathbf{v}_i \in [0, 1]$ and focus on binary classification, where label
 6 $y \in \{-1, 1\}$. Then we can define $\text{corr}_i^{\text{train}} = \text{corr}(y^{\text{train}}, \mathbf{v}_i^{\text{train}})$ is the correlation between the train
 7 label y^{train} and the i -th factor $\mathbf{v}_i^{\text{train}}$ corresponding to x^{train} , we can define $\text{corr}_i^{\text{test}} = \text{corr}(y^{\text{test}}, \mathbf{v}_i^{\text{test}})$
 8 as the same way. After that, we can obtain different partitions from \mathbf{v} based on above correlations:

$$\begin{aligned}\mathbf{v}_{pp} &= \{\mathbf{v}_i \mid \text{corr}_i^{\text{train}} > 0, \text{corr}_i^{\text{test}} > 0\}, \\ \mathbf{v}_{pn} &= \{\mathbf{v}_i \mid \text{corr}_i^{\text{train}} > 0, \text{corr}_i^{\text{test}} \leq 0\}.\end{aligned}\quad (1)$$

9 We regard \mathbf{v}_{pp} as the generalizable factors which are commonly positively-correlated with the
 10 label, and \mathbf{v}_{pn} as the spurious factors which are only positively-correlated with the label during the
 11 training-time. As the same way, we can obtain the rest two partitions from \mathbf{v} as:

$$\begin{aligned}\mathbf{v}_{np} &= \{\mathbf{v}_i \mid \text{corr}_i^{\text{train}} \leq 0, \text{corr}_i^{\text{test}} > 0\}, \\ \mathbf{v}_{nn} &= \{\mathbf{v}_i \mid \text{corr}_i^{\text{train}} \leq 0, \text{corr}_i^{\text{test}} \leq 0\}.\end{aligned}\quad (2)$$

12 Meanwhile, the classifier \mathbf{W} also has the corresponding partitions $\{\mathbf{w}_{pp}, \mathbf{w}_{pn}, \mathbf{w}_{np}, \mathbf{w}_{nn}\}$. Then we
 13 can obtain the prediction P as follows:

$$\begin{aligned}P_{\mathbf{w}}(\mathbf{x}) &= \mathbf{w} \cdot \mathbf{v}(\mathbf{x}) \\ &= \mathbf{w}_{pp} \cdot \mathbf{v}_{pp} + \mathbf{w}_{pn} \cdot \mathbf{v}_{pn} + \mathbf{w}_{np} \cdot \mathbf{v}_{np} + \mathbf{w}_{nn} \cdot \mathbf{v}_{nn},\end{aligned}\quad (3)$$

14 and we can obtain the pseudo label \hat{y} based on $P_{\mathbf{w}}(\mathbf{x})$ as:

$$\hat{y} = \begin{cases} 1 & p_{\mathbf{w}}(\mathbf{x}) > 0 \\ -1 & \text{otherwise} \end{cases}.\quad (4)$$

15 Then we can define the harmful sample as one that reduces the difference in the mean logits
 16 (prediction) between classes during the test-time procedure. And the change in logits of \mathbf{x}^{test} due to
 17 the model’s update by \mathbf{x} is as follows:

$$\Delta(w \cdot \mathbf{v}(\mathbf{x}^{\text{test}})) = \Delta w(\mathbf{x}) \cdot \mathbf{v}(\mathbf{x}^{\text{test}}).\quad (5)$$

18 Under the training-free test-time adaptation scenario in VLMs, we utilise the properties of the cache-
 19 based method, the change of classifier $\Delta w(\mathbf{x})$ can be approximated as the stored visual feature, due
 20 to the stored visual feature can be regarded as the residual update of the raw classifier, thus Eq. 5 can
 21 be approximated as:

$$\Delta(w \cdot \mathbf{v}(\mathbf{x}^{\text{test}})) = \Delta w(\mathbf{x}) \cdot \mathbf{v}(\mathbf{x}^{\text{test}}) \approx \mathbf{v}(\mathbf{x}) \cdot \mathbf{v}(\mathbf{x}^{\text{test}})\quad (6)$$

22 As a result, the change in the gap between the mean logits of samples belonging to the two classes is
 23 as follows:

$$\begin{aligned}&\Delta \left(\mathbb{E}_{\mathbf{x}^{\text{test}} \sim \mathcal{X}_{+1}^{\text{test}}} [w \cdot \mathbf{v}(\mathbf{x}^{\text{test}})] - \mathbb{E}_{\mathbf{x}^{\text{test}} \sim \mathcal{X}_{-1}^{\text{test}}} [w \cdot \mathbf{v}(\mathbf{x}^{\text{test}})] \right) \\ &= \mathbb{E}_{\mathbf{x}^{\text{test}} \sim \mathcal{X}_{+1}^{\text{test}}} [\Delta w \cdot \mathbf{v}(\mathbf{x}^{\text{test}})] - \mathbb{E}_{\mathbf{x}^{\text{test}} \sim \mathcal{X}_{-1}^{\text{test}}} [\Delta w \cdot \mathbf{v}(\mathbf{x}^{\text{test}})] \\ &= \mathbf{v}(\mathbf{x}) \cdot \left(\mathbb{E}_{\mathbf{x}^{\text{test}} \sim \mathcal{X}_{+1}^{\text{test}}} [\mathbf{v}(\mathbf{x}^{\text{test}})] - \mathbb{E}_{\mathbf{x}^{\text{test}} \sim \mathcal{X}_{-1}^{\text{test}}} [\mathbf{v}(\mathbf{x}^{\text{test}})] \right).\end{aligned}\quad (7)$$

24 It is straightforward that if the change of Eq. 7 is negative, it means that the logit (prediction) space
 25 of the two classes is getting closer, which will reduce the class discriminability.

26 Considering a test sample x with a high-confidence pseudo-label of $\hat{y} = 1$, its prediction in Eq. 3
 27 should satisfy as

$$\begin{aligned} p_{\mathbf{w}}(\mathbf{x}) &= \mathbf{w}_{pp} \cdot \mathbf{v}_{pp} + \mathbf{w}_{pn} \cdot \mathbf{v}_{pn} + \mathbf{w}_{np} \cdot \mathbf{v}_{np} + \mathbf{w}_{nn} \cdot \mathbf{v}_{nn} \gg 0, \\ |\mathbf{w}_{pp} \cdot \mathbf{v}_{pp} + \mathbf{w}_{pn} \cdot \mathbf{v}_{pn}| &\gg |\mathbf{w}_{np} \cdot \mathbf{v}_{np} + \mathbf{w}_{nn} \cdot \mathbf{v}_{nn}|. \end{aligned} \quad (8)$$

28 From Eq. 8, it is obvious that \mathbf{v}_{pp} and \mathbf{v}_{pn} become the dominant, and \mathbf{v}_{np} and \mathbf{v}_{nn} tend to become
 29 zero. Therefore, we can give the relationship of \mathbf{v}_{pp} and \mathbf{v}_{pn} to compose the desired value of $(\mathbf{x}^{\text{test}})$
 30 as

$$\begin{aligned} \mathbb{E}_{\mathbf{x}^{\text{test}} \sim \mathcal{X}_{+1}^{\text{test}}} [\mathbf{v}_{pp}(\mathbf{x}^{\text{test}})] &> \mathbb{E}_{\mathbf{x}^{\text{test}} \sim \mathcal{X}_{-1}^{\text{test}}} [\mathbf{v}_{pp}(\mathbf{x}^{\text{test}})], \\ \mathbb{E}_{\mathbf{x}^{\text{test}} \sim \mathcal{X}_{+1}^{\text{test}}} [\mathbf{v}_{pn}(\mathbf{x}^{\text{test}})] &\leq \mathbb{E}_{\mathbf{x}^{\text{test}} \sim \mathcal{X}_{-1}^{\text{test}}} [\mathbf{v}_{pn}(\mathbf{x}^{\text{test}})]. \end{aligned} \quad (9)$$

31 And Eq. 7 can be approximated as follows:

$$\begin{aligned} \mathbf{v}(\mathbf{x}) \cdot &\left(\mathbb{E}_{\mathbf{x}^{\text{test}} \sim \mathcal{X}_{+1}^{\text{test}}} [\mathbf{v}(\mathbf{x}^{\text{test}})] - \mathbb{E}_{\mathbf{x}^{\text{test}} \sim \mathcal{X}_{-1}^{\text{test}}} [\mathbf{v}(\mathbf{x}^{\text{test}})] \right) \\ \approx &\mathbf{v}_{pp}(\mathbf{x}) \cdot \left(\mathbb{E}_{\mathbf{x}^{\text{test}} \sim \mathcal{X}_{+1}^{\text{test}}} [\mathbf{v}_{pp}(\mathbf{x}^{\text{test}})] - \mathbb{E}_{\mathbf{x}^{\text{test}} \sim \mathcal{X}_{-1}^{\text{test}}} [\mathbf{v}_{pp}(\mathbf{x}^{\text{test}})] \right) \\ &+ \mathbf{v}_{pn}(\mathbf{x}) \cdot \left(\mathbb{E}_{\mathbf{x}^{\text{test}} \sim \mathcal{X}_{+1}^{\text{test}}} [\mathbf{v}_{pn}(\mathbf{x}^{\text{test}})] - \mathbb{E}_{\mathbf{x}^{\text{test}} \sim \mathcal{X}_{-1}^{\text{test}}} [\mathbf{v}_{pn}(\mathbf{x}^{\text{test}})] \right) \end{aligned} \quad (10)$$

32 In fact, Eq. 10 is consistent with Eq. 4 in the main manuscript with simplified notation, and if it is
 33 negative, the sample is harmful even if it has high confidence. Considering the relation in Eq. 9,
 34 the part before the plus sign in Eq. 10 is positive while the rest part is negative. Therefore, to make
 35 the whole Eq. 10 positive, generalizable factors \mathbf{v}_{pp} should be highlighted and spurious factors
 36 \mathbf{v}_{pn} should be avoided, which approve the claims about the benefits of generalizable factors for
 37 training-free test-time adaptation in VLMs.

38 In summary, our SSG introduces the theoretical analysis about generalizable factors for the training-
 39 free VLM TTA methods for the first time, which can not be intuitively supported by theory in [2]
 40 without any improvement. We are fully inspired by the insight and theoretical contribution from
 41 previous methods [1, 2], and we extend the derivation from entropy minimization to the property of
 42 cache-based methods.

43 B More Experiments

44 B.1 Other Shape Perturbations

45 In the shape-sensitive factors measurement, we choose the patch-shuffle to permute the raw image. In
 46 fact, there are other shape perturbations that can be applied, thus we choose the other two commonly
 47 utilized perturbations, pixel-shuffle and center occlusion, to replace the default patch-shuffle in our
 48 SSG, aiming to explore the performance difference. Through the pixel-shuffle, the mean color of the
 49 image is maintained, but it becomes difficult to clarify both the object and the background. At the
 50 same time, Center occlusion is able to preserve the background, but when the object is not centered
 51 or large, the object’s shape is not fully disrupted.

Table 1: Comparison results in terms of different shape permutations.

Method	ImageNet-A	OOD Average
Pixel-Shuffle	57.32	62.09
Center Occlusion	58.79	63.79
Patch-Shuffle	62.02	65.20

52 We conduct experiments with the ViT-B/16 backbone on the out-of-distribution benchmark, and
 53 the results are shown in Table 1. From the results, it is obvious that patch-shuffle obtains the best
 54 performance on both ImageNet-A and average OOD benchmark performance. The reason is that
 55 patch-shuffle can corrupt the shape of the object but preserve local information. Center occlusion
 56 and pixel-shuffle show an obvious decrease in accuracy because they cannot corrupt the shape of an
 57 object solely and preserve other details.

58 B.2 The Effect of PPD

59 In the design of SSG, PPD can be applied with entropy to cast as the comparison criteria during the
60 update of the visual cache, such as $H(f_{\text{test}} \mathbf{W}_C^T) - \lambda * \text{PPD}$. This means that not only high-confidence
61 is considered, but also high shape-sensitive and style-insensitive factors are considered. In default,
62 we set λ as 1 in our manuscript. In this section, we modify the weight λ , to ablate the effect of PPD
63 in our SSG.

64 We conduct experiments with ViT-B/16 on the out-of-distribution benchmark, and the average results
65 are shown in Table 2. From the results, we can find that SSG with $\lambda \in \{0.5, 1.0, 2.0\}$ still performs
66 promising OOD average performance, which indicates that our SSG is robust to the key hyper-
67 parameters. Meanwhile, we also find when SSG is equipped with larger λ such as 5.0 or 10.0, the
68 performance decreases obviously. This phenomenon shows that entropy is important to compare and
69 store the visual keys along with the visual values, too large PPD influences the effect of entropy, thus
70 resulting in decreased performance.

Table 2: Ablation results in terms of different weights λ of PPD.

λ	0.5	1.0	2.0	5.0	10.0
SSG	64.91	65.20	64.73	64.59	64.02

71 B.3 Significance Testing

72 We supplement significance testing using the Wilcoxon Signed-Rank Test, aiming to show the statisti-
73 cal significance. Specifically, we evaluate whether the accuracy of our SSG models is significantly
74 higher than TDA on different datasets, and whether our SSG^+ is significantly higher than Boost-
75 Adapter. Results are shown in Table 3. The notation used in the tables is as follows: $\checkmark\checkmark$ indicates
76 highly significant results ($p < 0.001$), \checkmark denotes significant results ($0.001 \leq p < 0.05$).

Table 3: The statistical results.

Datasets	IN-A	IN-V2	IN-R	IN-S	Datasets	IN-A	IN-V2	IN-R	IN-S
RN-50	$\checkmark\checkmark$	$\checkmark\checkmark$	$\checkmark\checkmark$	\checkmark	RN-50	$\checkmark\checkmark$	$\checkmark\checkmark$	$\checkmark\checkmark$	\checkmark
ViT-B/16	\checkmark	\checkmark	$\checkmark\checkmark$	$\checkmark\checkmark$	ViT-B/16	$\checkmark\checkmark$	\checkmark	\checkmark	$\checkmark\checkmark$

(a) SSG vs TDA

(b) SSG^+ vs BoostAdapter

77 B.4 Efficiency Comparison

78 We supplement the efficiency comparison in ImageNet-V, containing 10000 images in the following
79 table. From the table, even for the 10k image scale, our SSG only increases 1 minute compared with
80 TDA while improving the OOD accuracy by about 1.3%. Our SSG only demands 56% inference
81 time compared with DEP while achieving better performance.

Table 4: Efficiency Comparison.

Methods	Testing Time	OOD Accuracy
TDA	22 min	63.89
DPE	41 min	64.43
SSG (Ours)	23 min 10 s	65.20

82 B.5 Augmentation-only Ablation

83 We supplement the augmentation-only ablation results below. Model-(a) means our SSG applying
84 shape and style perturbation without PPD-based reweighting or cache updates, and model-(b) means

our full SSG. From the following table, we can find that our full model shows an obviously better performance than model (a).

Table 5: Augmentation-only Ablation.

OOD Accuracy	(a)	(b)
SSG (CLIP-RN50)	46.38	47.78
SSG (CLIP-ViT-B/16)	63.58	65.20

B.6 Ablation of k^*

Following TPT [3] and TDA [4], we set the k^* as the top-10%. We supplement the ablation results about the k^* value, as shown below. As shown in the table, we can find that setting k^* as the top-10% achieves the best performance, which is consistent with the conclusion from TPT.

Table 6: Ablation of k^* .

OOD Accuracy	5%	10%	15%	20%
SSG (CLIP-RN50)	47.45	47.78	47.59	47.55
SSG (CLIP-ViT-B/16)	64.89	65.20	65.04	65.02

B.7 Comparison results on ImageNet

We provide the SSG performance on the ImageNet below. As shown in the table, SSG/SSG+ gives superior performance compared with training-free methods TDA/BoostAdapter.

Table 7: Comparison results on the ImageNet.

CLIP-RN50	ImageNet	CLIP-ViT-B/16	ImageNet
TPT	60.74	TPT	68.98
TDA	61.35	TDA	69.51
DPE	63.41	DPE	71.91
BoostAdapter	61.54	BoostAdapter	69.92
SSG	62.54	SSG	71.21
SSG+	62.87	SSG+	71.66

B.8 Comparison results against simple baseline augmentations

We provide the comparison experiment between (a) PPD with Gaussian noise and rotations, and (b) PPD with shape patch-shuffle and style colour transformation with hue adjustment.

Table 8: Comparisons against simple baseline augmentations.

OOD Accuracy	(a)	(b)
SSG (CLIP-RN50)	46.02	47.78
SSG (CLIP-ViT-B/16)	63.08	65.20

From the above table, we can find that model (b) gives a much better performance than model (a). In fact, to obtain the PPD_{st} and PPD_{sh} , we need to corrupt/perturb the image’s shape/style information, and patch-shuffle and colour transformation are suitable to reach this goal, because the augmented images obviously lose or change the shape/style information. But for the rotation and Gaussian noise augmentation, the shape information is perturbed slightly and more than the style information is destroyed, which even gives a negative impact on the performance.

B.9 Additional ablation results for CLIP-RN50

First, we provide the ablation results about shape-sensitive and style-insensitive factors. Furthermore, we supplement the baseline method performance, and the baseline method only contains one positive cache and can be considered as the cache-based method TDA without a negative cache. From the table, both shape-sensitive and style-insensitive factors give the better performance, while the combination gives the best performance.

Table 9: Ablation results about shape-sensitive and style-insensitive factors on CLIP-RN50.

OOD Accuracy	Baseline	PPD _{st}	PPD _{sh}	PPD
SSG (CLIP-RN50)	46.33	46.85	47.11	47.78

Second, we report the ablation results about hyperparameters (patch number) on the shape perturbation (patch shuffle) below. From the table, the same conclusion is show as patch number as {4, 6} gives the better performance, and we set patch number as 4 for all experiments.

Table 10: Ablation results about hyperparameters of patch number on CLIP-RN50.

OOD Accuracy	2	4	6	8	16
SSG (CLIP-RN50)	47.03	47.78	47.51	46.93	45.23

C Additional Implementation Details

C.1 Dataset Details

The first benchmark is the out-of-distribution datasets, which are used to evaluate the model’s robustness to natural distribution shifts. There are 4 out-of-distribution versions of ImageNet, which contains images with varying styles and corruptions. We give a concise overview of each dataset below.

- **ImageNet-A**[5] is a subset of 7,500 naturally perturbed ImageNet images of 200 classes.
- **ImageNet-V2**[6] contains 10,000 images and 1,000 ImageNet classes, and was collected by an updated natural data collection pipeline to the original ImageNet data.
- **ImageNet-R**[7] contains 30,000 images belonging to 200 categories of the ImageNet dataset, but with diverse artistic styles.
- **ImageNet-S**[8] contains 50,000 sketches of 1000 class objects from the ImageNet dataset, and represents a domain shift from natural images to sketches.

The second benchmark is the cross-domain datasets, which aims to evaluate the model’s transfer performance on 10 diverse recognition datasets. We give a concise overview of each dataset below.

- **FGVCAircraft**[9] has 3,333 images with 100 classes.
- **Caltech101**[10] contains 2,465 images with 100 classes.
- **StanfordCars**[11] has 8,041 images with 196 classes.
- **DTD**[12] contains 1,692 images with 47 classes.
- **EuroSAT**[13] contains 8,100 images with 10 classes.
- **Flowers102**[14] contains 2,4636 images with 102 classes.
- **Food101**[15] contains 30,300 images with 101 classes.
- **OxfordPets**[16] contains 3,669 images with 37 classes.
- **SUN397**[17] contains 19,850 images with 397 classes.
- **UCF101**[18] contains 3,783 images with 101 classes.

C.2 Textual Prompts Used in Experiments

As shown in Table 11, we detail the specific hand-crafted prompts for each dataset. Following DPE [19], we also employ [20] prompts to further enhance performance.

Table 11: Textual prompts used in experiments.

Dataset	Prompts
	“itap of a {CLASS}.”
	“a bad photo of the {CLASS}.”
ImageNet-A	“a origami {CLASS}.”
ImageNet-V2	“a photo of the large {CLASS}.”
ImageNet-R	“a {CLASS} in a video game.”
ImageNet-S	“art of the {CLASS}.”
	“a photo of the small {CLASS}.”
Caltech101	“a photo of a {CLASS}.”
DTD	“{CLASS} texture.”
EuroSAT	““a centered satellite photo of {CLASS}.”
FGVCAircraft	“a photo of a {CLASS}, a type of aircraft.”
Flowers102	“a photo of a {CLASS}, a type of flower.”
Food101	“a photo of a {CLASS}, a type of food.”
OxfordPets	“a photo of a {CLASS}, a type of pet.”
StanfordCars	“a photo of a {CLASS}.”
SUN397	“a photo of a {CLASS}.”
UCF101	“a photo of a person doing {CLASS}.”

D Broader Impact

In this work, we aim to enhance the reliability of machine learning models by exploiting generalizable factors in training-free test-time adaptation (TTA) scenarios, with CLIP [45] as our exemplar. We design SSG method to strengthen CLIP’s cross-domain and out-of-distribution robustness during test-time process, thereby addressing critical challenges in real-world domain shifts. We hope this work inspire future works to further study the robustness and generalization of large-scale vision-language models, especially for the generalizable factors and training-free methods.

References

- [1] O. Wiles, S. Goyal, F. Stimberg, S.-A. Rebuffi, I. Ktena, K. D. Dvijotham, and A. T. Cemgil, “A fine-grained analysis on distribution shift,” in *International Conference on Learning Representations*, 2022.
- [2] J. Lee, D. Jung, S. Lee, J. Park, J. Shin, U. Hwang, and S. Yoon, “Entropy is not enough for test-time adaptation: From the perspective of disentangled factors,” in *The Twelfth International Conference on Learning Representations*, 2024.
- [3] M. Shu, W. Nie, D.-A. Huang, Z. Yu, T. Goldstein, A. Anandkumar, and C. Xiao, “Test-time prompt tuning for zero-shot generalization in vision-language models,” *Advances in Neural Information Processing Systems*, vol. 35, pp. 14 274–14 289, 2022.
- [4] A. Karmanov, D. Guan, S. Lu, A. El Saddik, and E. Xing, “Efficient test-time adaptation of vision-language models,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 14 162–14 171.
- [5] D. Hendrycks, K. Zhao, S. Basart, J. Steinhardt, and D. Song, “Natural adversarial examples,” in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 15 262–15 271.
- [6] B. Recht, R. Roelofs, L. Schmidt, and V. Shankar, “Do imagenet classifiers generalize to imagenet?” in *International conference on machine learning*. PMLR, 2019, pp. 5389–5400.
- [7] D. Hendrycks, S. Basart, N. Mu, S. Kadavath, F. Wang, E. Dorundo, R. Desai, T. Zhu, S. Parajuli, M. 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*, 2021, pp. 8340–8349.

- 166 [8] H. Wang, S. Ge, Z. Lipton, and E. P. Xing, “Learning robust global representations by penalizing local
167 predictive power,” *Advances in Neural Information Processing Systems*, vol. 32, 2019.
- 168 [9] S. Maji, E. Rahtu, J. Kannala, M. Blaschko, and A. Vedaldi, “Fine-grained visual classification of aircraft,”
169 *arXiv preprint arXiv:1306.5151*, 2013.
- 170 [10] L. Fei-Fei, R. Fergus, and P. Perona, “Learning generative visual models from few training examples: An
171 incremental bayesian approach tested on 101 object categories,” in *2004 conference on computer vision
172 and pattern recognition workshop*. IEEE, 2004, pp. 178–178.
- 173 [11] J. Krause, M. Stark, J. Deng, and L. Fei-Fei, “3d object representations for fine-grained categorization,” in
174 *Proceedings of the IEEE international conference on computer vision workshops*, 2013, pp. 554–561.
- 175 [12] M. Cimpoi, S. Maji, I. Kokkinos, S. Mohamed, and A. Vedaldi, “Describing textures in the wild,” in
176 *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014, pp. 3606–3613.
- 177 [13] P. Helber, B. Bischke, A. Dengel, and D. Borth, “Eurosat: A novel dataset and deep learning benchmark
178 for land use and land cover classification,” *IEEE Journal of Selected Topics in Applied Earth Observations
179 and Remote Sensing*, vol. 12, no. 7, pp. 2217–2226, 2019.
- 180 [14] M.-E. Nilsback and A. Zisserman, “Automated flower classification over a large number of classes,” in *2008
181 Sixth Indian conference on computer vision, graphics & image processing*. IEEE, 2008, pp. 722–729.
- 182 [15] L. Bossard, M. Guillaumin, and L. Van Gool, “Food-101—mining discriminative components with random
183 forests,” in *Computer vision—ECCV 2014: 13th European conference, zurich, Switzerland, September 6–12,
184 2014, proceedings, part VI 13*. Springer, 2014, pp. 446–461.
- 185 [16] O. M. Parkhi, A. Vedaldi, A. Zisserman, and C. Jawahar, “Cats and dogs,” in *2012 IEEE conference on
186 computer vision and pattern recognition*. IEEE, 2012, pp. 3498–3505.
- 187 [17] J. Xiao, J. Hays, K. A. Ehinger, A. Oliva, and A. Torralba, “Sun database: Large-scale scene recognition
188 from abbey to zoo,” in *2010 IEEE computer society conference on computer vision and pattern recognition*.
189 IEEE, 2010, pp. 3485–3492.
- 190 [18] K. Soomro, “Ucf101: A dataset of 101 human actions classes from videos in the wild,” *arXiv preprint
191 arXiv:1212.0402*, 2012.
- 192 [19] C. Zhang, S. Stepputtis, K. Sycara, and Y. Xie, “Dual prototype evolving for test-time generalization of
193 vision-language models,” *Advances in Neural Information Processing Systems*, vol. 37, pp. 32 111–32 136,
194 2024.
- 195 [20] S. Pratt, I. Covert, R. Liu, and A. Farhadi, “What does a platypus look like? generating customized prompts
196 for zero-shot image classification,” in *Proceedings of the IEEE/CVF International Conference on Computer
197 Vision*, 2023, pp. 15 691–15 701.