SCAP: Transductive Test-Time Adaptation via Supportive Clique-based Attribute Prompting

Supplementary Material

1. Benchmark Details

Here, we provide the sub-dataset details of the out-of-distribution (OOD) and cross-domain benchmarks exploited in this paper.

1.1. Details on Datasets in the OOD Benchmark

Our OOD benchmark consists of four out-of-distribution datasets derived from ImageNet [1].

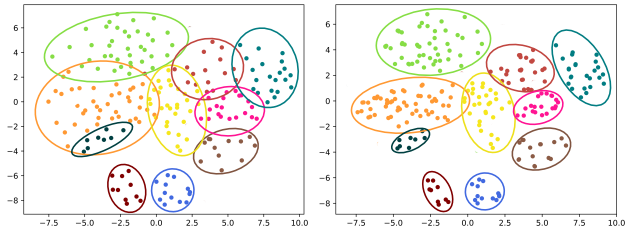
- **ImageNet-A** [5] contains 7,500 images of 200 classes that are naturally perturbed and misclassified by ResNet-50 [3] in ImageNet.
- **ImageNet-R** [4]includes 30,000 images covering 200 classes across 16 artistic and stylistic domains, such as cartoons, graffiti, and sketches, posing significant challenges due to their diverse visual transformations.
- **ImageNet-Sketch** [11] contains 50,000 images of 1,000 categories. The distribution of ImageNet-Sketch differs greatly from the pre-training data of CLIP since it only contains black-and-white sketches. Thus, it has been a challenging dataset for TTA.
- **ImageNet-V2** [10] comprises 10,000 images across 1,000 classes, sampled a decade after the original ImageNet dataset. It presents a naturally evolved distribution shift, making it a realistic benchmark for evaluating generalization performance..

OOD Datasets	Size	Number of Classes
ImageNet-A	7,500	200
ImageNet-R	30,000	200
ImageNet-Sketch	50,000	1,000
ImageNet-V2	10,000	1,000

Table 1. Overview of datasets in OOD benchmark

Cross-Domain	Size	Number of Classes
Aircraft	3,333	100
Caltech101	2,465	100
Flower102	2,463	102
Pets	3,669	37

Table 2. Overview of datasets in Cross-domain benchmark.



(a) Without Attribute-Prompts (b) With Attribute-Prompts

Figure 1. Visualization of the distribution of image features with and without the learned Attribute Prompts.

1.2. Details on Datasets in Cross-domain Benchmark

The cross-domain benchmark comprises four datasets from distinct visual domains, providing a diverse evaluation setting for assessing model generalization.

- **Aircraft** [7] consists of 10,200 images representing 102 different aircraft model variants, each with 100 images. The dataset predominantly features airplanes and poses challenges in fine-grained visual classification.
- **Caltech101** [2] includes 2,465 images spanning 101 object categories along with a background class. Each category contains 40 to 800 images, with most classes averaging around 50 samples, making it a benchmark for object recognition tasks.
- **Flower102** [8] comprises 2,463 images of 102 flower species. Due to the significant domain gap between flowers and standard pretraining datasets, this dataset serves as a robust test for domain generalization.
- **Pets** [9] contains images of cats and dogs across 37 different breeds. The dataset exhibits high intra-class variations in scale, pose, and lighting, making it challenging for fine-grained classification tasks.

2. The Effectiveness of Attribute Prompts

To demonstrate the effectiveness of our supportive clique-based attribute prompting, we visualize the embedding distributions of images from different classes in Figure 1, where with and without the corresponding Attribute Prompts setting are conducted. As illustrated in Figure 1, incorporating Attribute Prompts improves intra-class compactness and inter-class discrimination. Notably, several previously ambiguous samples near class boundaries shift closer to their respective class centers, thereby enhancing the overall quality of the learned image representations.

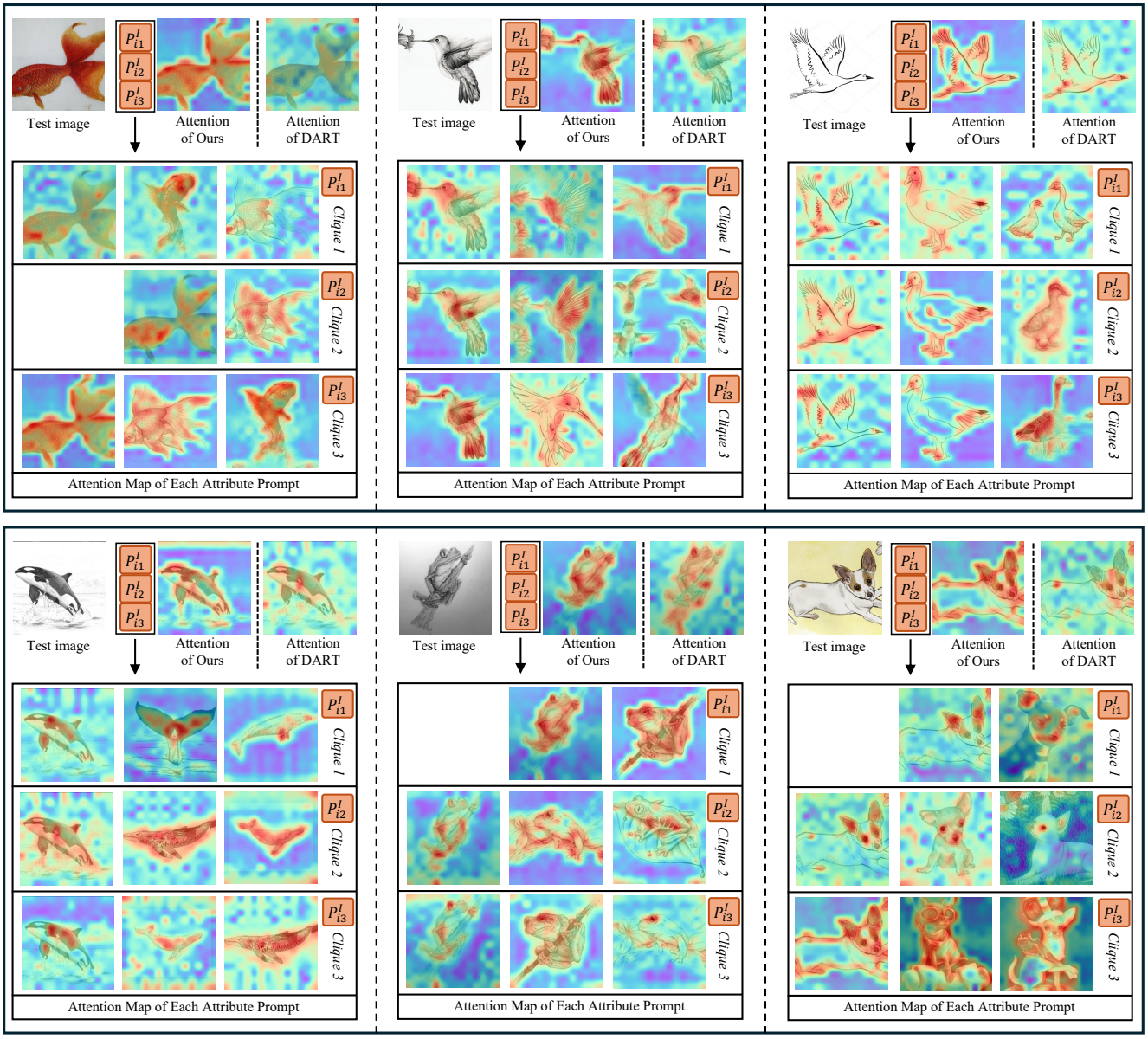


Figure 2. Additional visualization results of the attention maps.

Method	Publication	Cars	SUN397	Aircraft	EuroSAT	Food101	Pets	Flower102	Caltech101	DTD	UCF101	Average
TDA	CVPR 2024	67.28	67.62	23.91	58.00	86.14	88.63	71.42	94.24	47.40	70.66	67.53
SCAP	**This Paper**	**69.25**	66.41	**25.44**	**58.62**	**86.58**	**90.27**	**71.65**	**94.42**	**47.79**	68.36	**67.88**

Table 3. **Additional Results on the domain-shift datasets.** SCAP compared with the second-highest TDA.

This improvement is attributed to our proposed *Concentration Loss*, which, together with *Entropy Loss*, encourages samples within the same supportive clique to leverage shared attributes, leading to more tightly clustered and semantically consistent features.

3. Visualization of Attention Maps

In Fig. 2, we provide additional visualizations of the attention maps in comparison with DART [6]. The results demonstrate that each of our *visual attribute prompts* guides CLIP to attend to the specific attributes shared among

images within a supportive clique, thereby enabling fine-grained attribute-based prompt learning. By aggregating all relevant *visual attribute prompts* associated with a given test image, our approach effectively directs attention toward the most salient attributes, leading to enhanced feature extraction and improved classification performance. In contrast, the attention maps generated by DART often exhibit dispersed or incomplete attention, failing to sufficiently capture critical object attributes. These findings highlight the superior prompt-learning capability of SCAP in accurately and comprehensively leveraging visual information.

4. Additional Results on the domain-shift datasets

In Tab. 3, we report results on six additional datasets alongside the four datasets in our main paper from the cross-domain benchmark. SCAP consistently outperforms the second-best method, TDA, on eight out of ten datasets, achieving an average accuracy improvement of **0.35%**. These results further demonstrate SCAP's effectiveness in adapting to diverse domain shifts, highlighting its robustness in transductive TTA scenarios.

References

[1] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, pages 248–255. Ieee, 2009. 1

[2] 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 *CVPRW*, pages 178–178. IEEE, 2004. 1

[3] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016. 1

[4] 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 *ICCV*, pages 8340–8349, 2021. 1

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

[6] Zichen Liu, Hongbo Sun, Yuxin Peng, and Jiahuan Zhou. Dart: dual-modal adaptive online prompting and knowledge retention for test-time adaptation. In *AAAI*, pages 14106–14114, 2024. 2

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

[8] 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

[9] Omkar M Parkhi, Andrea Vedaldi, Andrew Zisserman, and CV Jawahar. Cats and dogs. In *CVPR*, pages 3498–3505. IEEE, 2012. 1

[10] Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do imagenet classifiers generalize to imagenet? In *ICML*, pages 5389–5400. PMLR, 2019. 1

[11] Haohan Wang, Songwei Ge, Zachary Lipton, and Eric P Xing. Learning robust global representations by penalizing local predictive power. *NeurIPS*, 32, 2019. 1