COSMIC: Clique-Oriented Semantic Multi-space Integration for Robust CLIP Test-Time Adaptation

Supplementary Material

1. Overview

This supplementary material provides additional experiments, visualizations, and implementation details to support our main paper. The content is organized as follows:

- **Extra Ablation Experiments** (Sec. 2). We analyze the impact of DINOv2 cache capacity, image augmentation views, and AFV class center calculation methods on our model's performance.
- **Extra Overhead Discussion** (Sec. 3). We analyze the storage and time complexity of our method, showing its efficiency through reduced dual graph update frequency and detailed complexity approximations.
- **Extra Visualization** (Sec. 4). We present t-SNE visualizations of class and hyper-class distributions, cached features from CLIP and DINOv2, and examples of queried classes to illustrate our method's effectiveness.
- **Extra Implementation Details** (Sec. 5). We provide comprehensive dataset statistics and textual prompts used for various recognition tasks.

These materials offer a deeper understanding of our method's overhead, robustness, visual performance, and experimental setup.

2. Extra Ablation Experiments

2.1. Ablation of the Capacity of DINOv2 Cache

We present the performance of COSMIC with different capacities (number of examples stored) in DINOv2's cache in Tab. 1. It shows that increasing the number of stored examples leads to better prediction, but COSMIC achieves reasonable performance even with a smaller cache capacity.

Table 1. Performance comparison using different cache capacities (number of examples stored) in DINOv2's cache on ImageNet-Val [3]. For each test, we use CLIP-RN-50 and DINOv2 ViT-S/14 as our visual encoders.

Method	# of Examples Stored	Top-1 Accuracy (%)	Gain (%)
CLIP-RN-50	-	66.99	-
Ours	1	67.10	0.11
	3	68.59	1.60
	6	68.90	1.91
	8	68.92	1.93
	10	68.86	1.87

2.2. Ablation of the Augment Views of Images

We evaluated the performance of COSMIC with different numbers of augmented views of images, as shown in Tab. 2. The results indicate that increasing the number of views enhances prediction. Specifically, COSMIC achieves its highest Top-1 accuracy gain (2.02%) with 16 views. However, even with fewer augmented views, COSMIC still performs well and has the advantage of faster inference times.

Table 2. Performance comparison using different augment views of images on ImageNet-Val [3]. We use CLIP-RN-50 and DINOv2 ViT-S/14 for each test as our visual encoders.

Method	# of Augment Views	Top-1 Accuracy (%)	Gain (%)
CLIP-RN-50	-	66.99	-
Ours	1	68.37	1.47
	2	68.59	1.69
	4	68.79	1.89
	8	68.87	1.97
	16	68.92	2.02
	32	68.90	2.00

2.3. Ablation of Calculation of AFV Class Center

To investigate the impact of various class center calculation methods in the Auxiliary Fine-grained Visual space on performance, we conducted a comparative analysis. Tab. 3 shows our method significantly improves upon the CLIP-RN-50 baseline using both average and attention-weighted AFV class centers. The average method achieves the highest Top-1 accuracy gain (1.91%), slightly surpassing the attention-weighted method (1.62%) and the EMA method (1.60%). This suggests that equal consideration of all cached features may better capture class-level representations. The slight performance decrease (-0.03%) of the EMA method without entropy-based selection emphasizes the importance of careful feature selection. These results highlight the critical role of AFV class center calculation in leveraging cached features, with the simple averaging method emerging as the preferred choice due to its effectiveness and simplicity.

3. Extra Overhead Discussion

As shown in Tab. 4, time cost can be reduced in real applications by decreasing the frequency of dual graph updates—such as every 50 steps—while still achieving SOTA. **[Storage]**: Constructing additional graph structures only requires $\mathcal{O}(K^2)$ space to store the adjacency matrix. Additionally, storing extra visual features only requires (class_num (K) + clique_num) $\times$ cache_size (n) $\times$ feat_dim (d_i) space, which is highly efficient with pytorch tensor. The approximated total storage/sample: $\mathcal{O}((d_1 +$

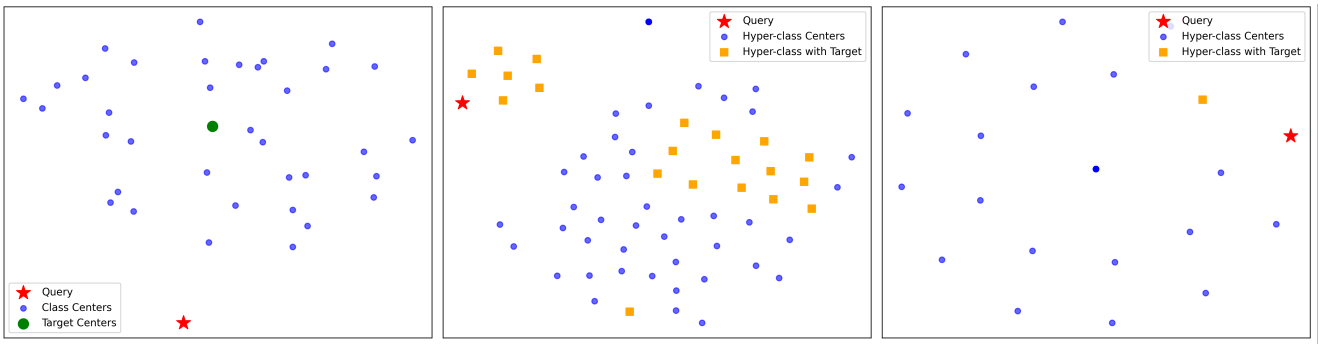


Figure 1. With a sample from Pets dataset [11], we implement t-SNE visualization of test features querying Textual Class Centers (**left**), CLIP Shared Semantics Hyper-class Centers (**middle**), and Auxiliary Fine-grained Visual Hyper-class Centers (**right**). "Target" denotes the ground-truth label. CLIP-ViT-B/16 and DINOv2 ViT-L/14 serve as visual encoders.

Table 3. Performance comparison using different AFV class center calculations on ImageNet-Val [3]. CLIP-RN-50 and DINOv2 ViT-S/14 are used as visual encoders. "Average" means the centroid of cached features for each class. "Attn weighted" means the weighted average of cached features for each class, with weights being the attention scores between the test feature and cached features. "EMA" means the exponential moving average of historical test features. "EN" means the prediction entropy-based selection of features.

Method	AFV Center	Top-1 Accuracy (%)	Gain (%)
CLIP-RN-50	-	66.99	-
Ours	Average	68.90	1.91
	Attn weighted	68.61	1.62
	EMA	68.59	1.60
	EMA w/o EN	66.96	-0.03

$d_2)nK + K^2 + \text{clique\_num} \times (d_1 + d_2))$. **[Time]**: Time complexity $b(K-b)3^{(b/3)}$ of maximal clique search is presented in main text. The approximated total time/sample: $\max(\mathcal{O}(\text{CLIP}), \mathcal{O}(\text{DINOv2})) + \mathcal{O}(d_1(2K)^2 + d_2K^2 + b(K-b)3^{(b/3)} + nK \times \text{clique\_num})$ where b is graph degeneracy.

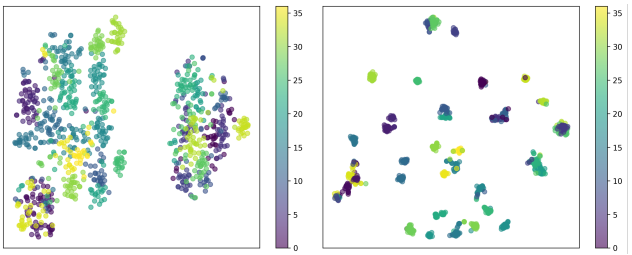
4. Extra Visualization

4.1. Distribution of Classes and Hyper-classes

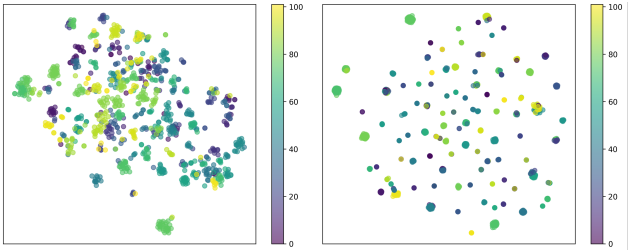
To showcase the effectiveness of our method during the cold-start phase, we visualize the distribution of randomly selected test samples from the first 100 tests in the Pets dataset [11] across three query spaces: Textual Class Centers, CLIP Shared Semantics Hyper-class Centers, and Auxiliary Fine-grained Visual Hyper-class Centers. We employ t-SNE to reduce the dimensionality of the high-dimensional features. As illustrated in Fig. 1, hyper-classes exhibit a more uniform distribution in the feature space. Notably, when ground truth (GT) feature centers are obscured by neighboring points, the Hyper-class Centers containing the GT target are more readily queried by test samples, resulting in improved prediction accuracy.

4.2. T-SNE of Cached Features from CLIP & DINOv2

In Fig. 2, we visualize the cached visual features from CLIP and DINOv2 caches after testing on various subsets of data using t-SNE. We observe that features of the same class (same color) in the DINOv2 cache are more clustered, especially during the cold-start phase, where it exhibits more distinctive class clustering and effectively mitigates overlap between similar categories, thereby facilitating fine-grained visual feature retrieval.



(a) Pets [11]



(b) Flower102 [10]

Figure 2. Distribution of cached visual features from CLIP (**left**) and DINOv2 (**right**) caches. The capacity of both caches are set to 50 and we capture the distribution in 1000 test iterations. CLIP-ViT-B/16 and DINOv2 ViT-L/14 serve as visual encoders.

Table 4. We use CLIP ViT-B-16 and DINOv2 ViT-S/14 as the backbone, updating the dual graph every 50 steps to show the average time & storage overhead per test sample.

Test	Type	CLIP Inference	TDA Overhead	COSMIC Overhead		
				DINOv2 Inference	CLIP Graph	DINOv2 Graph
Flower102 [10]	Time$_{(ms)}$	12.52	8.42	10.65	5.37	3.75
	Storage$_{(mb)}$	147.87	40.93	42.84	0.43	0.15
	Top-1 Acc$_{(\%)}$	72.76	75.11	-	77.10	80.92
Ucf101 [13]	Time$_{(ms)}$	17.94	10.99	12.20	6.31	4.32
	Storage$_{(mb)}$	147.91	40.61	74.09	1.62	1.46
	Top-1 Acc$_{(\%)}$	94.36	94.40	-	94.77	95.33

4.3. Samples of Queried Classes

Fig. 3 illustrates the enhanced performance achieved by querying hyper-classes within the CLIP Shared Semantics and Auxiliary Fine-grained Visual graphs, as opposed to the conventional approach of querying classes in the naive CLIP cache. Both graphs leverage the structured relationships and hierarchical organization of hyper-classes, enabling more precise and contextually relevant retrieval of semantic information.

5. Extra Implementation Details

5.1. Dataset Details

In Tab. 5, we present detailed statistics for each dataset used in our experiments, including the number of classes, test set sizes, and their respective target tasks.

5.2. Textual Prompts Details

Tab. 6 outlines the prompt formats for various visual recognition datasets. These prompts guide the model in identifying specific objects or scenes within each class, with tailored designs for optimal performance. This variation enhances the model's generalization and accuracy.

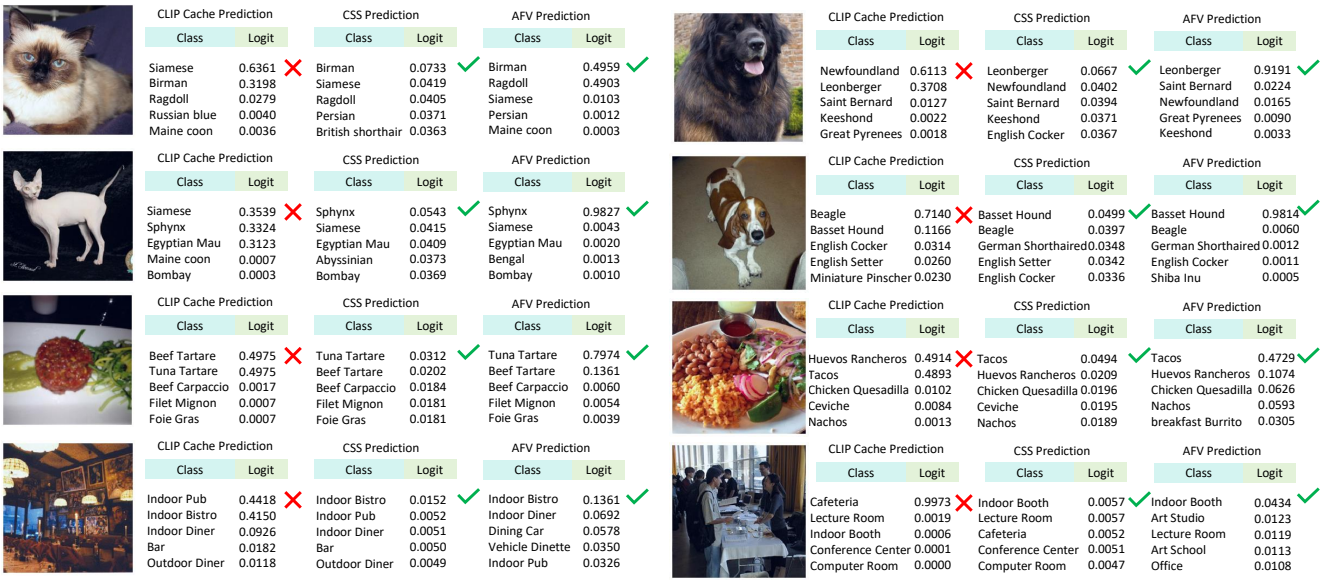


Figure 3. Samples of queried classes with clip feature cache, CSS Graph, and AFV Graph respectively. For each test, CLIP-ViT-B/16 and DINOv2 ViT-L/14 are used as visual encoders.

Table 5. Dataset Summary for Various Recognition Tasks. Note that we evaluate test datasets for all benchmarks.

Dataset	Classes	Train size	Validation size	Test size	Target Task
Caltech101 [4]	100	4,128	1,649	2,465	Object recognition
DTD [2]	47	2,820	1,128	1,692	Texture recognition
EuroSAT [5]	10	13,500	5,400	8,100	Satellite image recognition
FGVCAircraft [9]	100	3,334	3,333	3,333	Fine-grained aircraft recognition
Flowers102 [10]	102	4,093	1,633	2,463	Fine-grained flowers recognition
Food101 [1]	101	50,500	20,200	30,300	Fine-grained food recognition
OxfordPets [11]	37	2,944	736	3,669	Fine-grained pets recognition
StanfordCars [8]	196	6,509	1,635	8,041	Fine-grained car recognition
SUN397 [15]	397	15,880	3,970	19,850	Scene recognition
UCF101 [13]	101	7,639	1,898	3,783	Action recognition
ImageNet [3]	1,000	1.28M	-	50,000	Object recognition
ImageNet-A [7]	200	-	-	7,500	Robustness of adversarial attack
ImageNet-V2 [12]	1,000	-	-	10,000	Robustness of collocation
ImageNet-R [6]	200	-	-	30,000	Robustness of multi-domains
ImageNet-Sketch [14]	1,000	-	-	50,889	Robustness of sketch domain

Table 6. Textual Prompts for Various Recognition Tasks. The left column lists the dataset names, while the right column provides the prompt templates for each dataset, with empty curly braces {} representing the class placeholder.

Dataset	Prompts
Caltech101 [4]	"a photo of a {}."
DTD [2]	"{} texture."
EuroSAT [5]	"a centered satellite photo of {}."
FGVCAircraft [9]	"a photo of a {}, a type of aircraft."
Flowers102 [10]	"a photo of a {}, a type of flower."
Food101 [1]	"a photo of {}, a type of food."
OxfordPets [11]	"a photo of a {}, a type of pet."
StanfordCars [8]	"a photo of a {}, a type of car."
SUN397 [15]	"a bad photo of the {}.", "a {} in a video game.", "a origami {}.", "a photo of the small {}.", "art of the {}.", "a photo of the large {}.", "itap of a {}."
UCF101 [13]	"a photo of a person doing {}."
ImageNet [3]	"a bad photo of the {}.", "a {} in a video game.", "a origami {}.", "a photo of the small {}.", "art of the {}.", "a photo of the large {}.", "itap of a {}."
ImageNet-A [7]	"a bad photo of the {}.", "a {} in a video game.", "a origami {}.", "a photo of the small {}.", "art of the {}.", "a photo of the large {}.", "itap of a {}."
ImageNet-V2 [12]	"a bad photo of the {}.", "a {} in a video game.", "a origami {}.", "a photo of the small {}.", "art of the {}.", "a photo of the large {}.", "itap of a {}."
ImageNet-R [6]	"a bad photo of the {}.", "a {} in a video game.", "a origami {}.", "a photo of the small {}.", "art of the {}.", "a photo of the large {}.", "itap of a {}."
ImageNet-Sketch [14]	"a bad photo of the {}.", "a {} in a video game.", "a origami {}.", "a photo of the small {}.", "art of the {}.", "a photo of the large {}.", "itap of a {}."

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. 4, 5

[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. 4, 5

[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, 5

[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. 4, 5

[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. 4, 5

[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. 4, 5

[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. 4, 5

[8] 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. 4, 5

[9] Subhransu Maji, Esa Rahtu, Juho Kannala, Matthew Blaschko, and Andrea Vedaldi. Fine-grained visual classification of aircraft. *arXiv preprint arXiv:1306.5151*, 2013. 4, 5

[10] 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. 2, 3, 4, 5

[11] 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. 2, 4, 5

[12] 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. 4, 5

[13] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. 2012. 3, 4, 5

[14] 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. 4, 5

[15] 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. 4, 5