Online Gaussian Test-Time Adaptation of Vision-Language Models

Supplementary Material

A. Prompts

We show the handcrafted prompts in Table 9a, while the ensemble of prompts is shown in Table 9b.

Table 9. Prompt templates.

(a) Handcrafted prompts.

Dataset	Prompt template
ImageNet	"a photo of a []."
SUN397	"a photo of a []."
Aircraft	"a photo of a [], a type of aircraft.",
EuroSAT	"a centered satellite photo of [].",
Cars	"a photo of a [].",
Food101	"a photo of [], a type of food.",
Pets	"a photo of [], a type of pet.",
Flower102	"a photo of a [], a type of flower.",
Caltech101	"a photo of a [].",
DTD	"[] texture.",
UCF101	"a photo of a person doing [].",

(b) Ensemble prompts.

"itap of a []."
"a bad photo of the []."
"a origami []."
"a photo of the large []."
"a [] in a video game."
"art of the []."
"a photo of the small []."

B. Results with different architectures

In the main paper, all experiments are done using the ViT-B/16 version of CLIP. Here, we show that results with other backbones (ViT-L/14, ViT-B/32, ResNet50 and ResNet101), presented in Tables 10, 11, 12, 13 and 14, are coherent with the observations made previously. Note that we use the same *fixed hyper-parameters* across all datasets and architectures. For each dataset, the methods are tested using the same 100 runs.

B.1. Results with other ViT architectures

Table 10. We show results obtained with other ViT-based architectures and the handcrafted prompts of Table 9a.

(a) With ViT-B/32.

	ImageNet	SUN397	Aircraft	EuroSAT	StanfordCars	Food101	Pets	Flower102	Caltech101	DTD	UCF101	Average
Zero-Shot	62.03	62.11	19.14	45.38	60.17	80.40	87.33	66.67	91.44	42.61	63.52	61.9
TDA (CVPR '24)	62.8 ±0.07	63.7 ±0.12	18.4 ±0.33	46.3 ±0.93	60.3 ±0.26	80.0 ±0.06	86.7 ±0.27	67.6 ±0.31	**91.0** ±0.41	43.2 ±0.46	**65.3** ±0.32	62.3
DMN (CVPR '24)	61.5 ±0.12	63.4 ±0.19	18.4 ±0.29	47.5 ±1.17	60.0 ±0.31	77.5 ±0.12	86.6 ±0.35	**68.1** ±0.32	89.2 ±0.63	42.9 ±0.68	64.8 ±0.49	61.8
OGA (ours)	**63.0** ±0.11	**64.5** ±0.16	**18.7** ±0.32	**49.3** ±1.09	**61.6** ±0.20	**80.1** ±0.10	**88.2** ±0.29	67.8 ±0.31	89.5 ±0.56	**44.2** ±0.54	65.2 ±0.37	**62.9**

(b) With ViT-L/14.

	ImageNet	SUN397	Aircraft	EuroSAT	StanfordCars	Food101	Pets	Flower102	Caltech101	DTD	UCF101	Average
Zero-Shot	73.44	67.66	32.52	60.27	76.89	90.92	93.49	79.58	95.21	53.43	75.05	72.6
TDA (CVPR '24)	74.4 ±0.05	69.3 ±0.10	32.8 ±0.41	63.9 ±0.81	77.0 ±0.25	**90.8** ±0.05	93.5 ±0.15	80.3 ±0.32	94.5 ±0.33	55.0 ±0.37	76.7 ±0.27	73.5
DMN (CVPR '24)	74.4 ±0.10	70.0 ±0.16	32.3 ±0.46	**64.1** ±0.79	78.1 ±0.31	89.8 ±0.09	93.1 ±0.23	**81.6** ±0.31	94.4 ±0.43	54.5 ±0.62	78.1 ±0.40	73.7
OGA (ours)	**75.2** ±0.12	**70.7** ±0.19	**33.2** ±0.57	63.9 ±0.93	**79.2** ±0.29	90.7 ±0.08	**93.9** ±0.18	81.3 ±0.34	**94.9** ±0.37	**56.1** ±0.60	**78.4** ±0.36	**74.3**

Table 11. We show results obtained with other ViT-based architectures and the ensemble of prompts of Table 9b.

(a) With ViT-B/32.

	ImageNet	SUN397	Aircraft	EuroSAT	StanfordCars	Food101	Pets	Flower102	Caltech101	DTD	UCF101	Average
Zero-Shot	63.74	63.99	18.39	43.00	60.14	79.78	84.96	63.62	92.17	43.20	62.09	61.4
TDA (CVPR '24)	**64.1** ±0.07	65.3 ±0.11	17.6 ±0.35	49.5 ±1.15	60.6 ±0.26	79.4 ±0.08	84.0 ±0.29	64.1 ±0.29	**91.8** ±0.34	44.8 ±0.48	64.2 ±0.28	62.3
DMN (CVPR '24)	62.4 ±0.12	64.6 ±0.18	17.4 ±0.35	46.2 ±1.41	60.3 ±0.33	77.0 ±0.11	83.7 ±0.35	**65.4** ±0.35	90.0 ±0.55	45.6 ±0.60	65.5 ±0.42	61.6
OGA (ours)	63.7 ±0.10	**65.4** ±0.18	**18.3** ±0.31	**49.5** ±1.15	**61.5** ±0.23	**79.4** ±0.10	**85.9** ±0.27	64.4 ±0.42	90.2 ±0.56	**46.5** ±0.58	**65.6** ±0.34	**62.8**

(b) With ViT-L/14.

	ImageNet	SUN397	Aircraft	EuroSAT	StanfordCars	Food101	Pets	Flower102	Caltech101	DTD	UCF101	Average
Zero-Shot	75.90	70.44	31.23	50.16	77.68	91.29	92.83	77.47	95.58	55.73	76.29	72.2
TDA (CVPR '24)	76.3 ±0.05	71.5 ±0.11	31.3 ±0.40	63.5 ±0.47	77.9 ±0.23	**90.9** ±0.05	93.0 ±0.19	78.5 ±0.36	95.3 ±0.31	56.6 ±0.29	78.1 ±0.25	73.9
DMN (CVPR '24)	75.8 ±0.09	71.5 ±0.17	31.9 ±0.38	**64.6** ±0.95	78.8 ±0.29	90.0 ±0.08	93.4 ±0.23	**80.9** ±0.27	95.6 ±0.35	56.1 ±0.60	**79.3** ±0.39	74.4
OGA (ours)	**76.3** ±0.11	**72.2** ±0.19	**32.4** ±0.40	64.3 ±1.02	**79.5** ±0.22	90.8 ±0.08	**93.9** ±0.24	79.9 ±0.43	**95.7** ±0.33	**57.5** ±0.55	79.2 ±0.41	**74.7**

B.2. Results with CNNs architectures

Table 12. We show results obtained with CNNs-based architectures and the handcrafted prompts of Table 9a.

(a) With ResNet50.

	ImageNet	SUN397	Aircraft	EuroSAT	StanfordCars	Food101	Pets	Flower102	Caltech101	DTD	UCF101	Average
Zero-Shot	58.18	58.84	16.95	36.10	55.80	77.36	85.72	65.98	85.92	42.79	61.86	58.7
TDA (CVPR '24)	**59.1** ±0.07	60.3 ±0.13	16.2 ±0.37	39.1 ±1.84	56.5 ±0.22	**77.0** ±0.08	85.1 ±0.30	**67.1** ±0.34	**86.9** ±0.44	42.8 ±0.35	62.7 ±0.33	59.3
DMN (CVPR '24)	57.2 ±0.10	59.2 ±0.18	15.8 ±0.33	**44.8** ±1.91	55.3 ±0.34	73.6 ±0.13	83.3 ±0.43	66.5 ±0.37	85.3 ±0.59	42.2 ±0.60	61.9 ±0.46	58.6
OGA (ours)	58.8 ±0.12	**61.3** ±0.14	**16.3** ±0.34	43.8 ±1.97	**57.7** ±0.22	76.1 ±0.13	**85.5** ±0.37	66.1 ±0.44	85.4 ±0.58	**43.9** ±0.55	**62.9** ±0.44	**59.8**

(b) With ResNet101.

	ImageNet	SUN397	Aircraft	EuroSAT	StanfordCars	Food101	Pets	Flower102	Caltech101	DTD	UCF101	Average
Zero-Shot	61.26	59.04	18.12	32.80	63.15	80.67	86.89	64.35	90.02	37.06	61.01	59.5
TDA (CVPR '24)	62.4 ±0.07	60.7 ±0.14	17.8 ±0.34	41.2 ±0.70	63.5 ±0.22	80.4 ±0.08	86.2 ±0.25	64.4 ±0.41	**89.5** ±0.47	38.1 ±0.42	62.6 ±0.31	60.6
DMN (CVPR '24)	62.2 ±0.10	61.4 ±0.18	17.5 ±0.34	41.5 ±1.03	64.2 ±0.30	79.3 ±0.11	87.0 ±0.32	**66.4** ±0.34	89.1 ±0.50	38.4 ±0.58	64.0 ±0.49	61.0
OGA (ours)	**62.6** ±0.10	**61.9** ±0.15	**17.9** ±0.31	**44.4** ±1.22	**64.5** ±0.21	**80.6** ±0.11	**87.6** ±0.26	65.4 ±0.34	89.2 ±0.48	**39.3** ±0.59	**64.7** ±0.38	**61.6**

Table 13. We show results obtained with CNNs-based architectures and the ensemble of prompts of Table 9b.

(a) With ResNet50.

	ImageNet	SUN397	Aircraft	EuroSAT	StanfordCars	Food101	Pets	Flower102	Caltech101	DTD	UCF101	Average
Zero-Shot	60.25	60.96	16.41	27.09	56.31	76.42	82.77	62.65	87.79	40.48	60.16	57.4
TDA (CVPR '24)	**60.7** ±0.06	62.0 ±0.11	15.6 ±0.33	31.8 ±1.52	56.9 ±0.24	**75.9** ±0.09	82.7 ±0.28	64.1 ±0.41	**88.2** ±0.43	39.8 ±0.42	61.6 ±0.29	58.1
DMN (CVPR '24)	58.4 ±0.10	60.5 ±0.17	15.2 ±0.30	33.1 ±1.43	55.9 ±0.35	72.8 ±0.15	81.3 ±0.37	**64.5** ±0.39	87.3 ±0.55	39.3 ±0.64	61.0 ±0.51	57.2
OGA (ours)	59.7 ±0.12	**63.0** ±0.13	**15.8** ±0.33	**34.3** ±1.56	**58.2** ±0.24	75.3 ±0.15	**83.3** ±0.34	63.5 ±0.44	87.0 ±0.55	**40.3** ±0.58	**61.8** ±0.49	**58.4**

(b) With ResNet101.

	ImageNet	SUN397	Aircraft	EuroSAT	StanfordCars	Food101	Pets	Flower102	Caltech101	DTD	UCF101	Average
Zero-Shot	62.46	61.06	17.61	25.09	62.88	80.68	84.79	61.88	90.83	41.49	60.69	59.0
TDA (CVPR '24)	63.0 ±0.08	62.1 ±0.12	16.9 ±0.30	28.6 ±1.02	63.5 ±0.24	80.1 ±0.08	84.5 ±0.26	61.8 ±0.45	**90.0** ±0.47	40.9 ±0.45	62.1 ±0.34	59.4
DMN (CVPR '24)	62.9 ±0.10	62.6 ±0.16	16.8 ±0.32	33.7 ±1.51	**64.8** ±0.29	79.3 ±0.10	85.1 ±0.28	**64.3** ±0.44	89.4 ±0.52	41.1 ±0.54	63.4 ±0.43	60.3
OGA (ours)	**63.0** ±0.11	**62.6** ±0.18	**17.2** ±0.34	**34.4** ±1.38	64.5 ±0.20	**80.5** ±0.10	**86.3** ±0.27	63.0 ±0.31	89.6 ±0.46	**41.7** ±0.47	**63.5** ±0.38	**60.6**

B.3. Summary

Table 14. We show results averaged over the 11 datasets. Handcrafted refer to Table 9a while Ensemble corresponds to Table 9b.

		ViT-B/16	ViT-B/32	ViT-L/14	ResNet50	ResNet101
Handcrafted	**Zero-Shot**	65.3	61.9	72.6	58.7	59.5
	TDA	67.7	62.3	73.5	59.3	60.6
	DMN	67.5	61.8	73.7	58.6	61.0
	OGA (ours)	**68.5**	**62.9**	**74.3**	**59.8**	**61.6**
Ensemble	**Zero-Shot**	65.6	61.4	72.2	57.4	59.0
	TDA	66.9	62.3	73.9	58.1	59.4
	DMN	66.4	61.6	74.4	57.2	60.3
	OGA (ours)	**67.3**	**62.8**	**74.7**	**58.4**	**60.6**

C. Detailed results with batch size 1

Table 15. We report the average accuracy over 100 runs with batch size 1 with the ViT-B/16 backbone.

	ImageNet	SUN397	Aircraft	EuroSAT	StanfordCars	Food101	Pets	Flower102	Caltech101	DTD	UCF101	Average
Zero-Shot	66.74	62.58	24.87	48.28	65.64	85.87	89.10	70.81	93.31	43.44	67.49	65.3
OGA (ours)	**68.5** ±0.12	**66.2** ±0.19	**25.1** ±0.34	**64.3** ±0.76	**67.9** ±0.21	**86.0** ±0.07	**91.6** ±0.30	**72.8** ±0.36	**93.1** ±0.26	**45.6** ±0.57	**71.5** ±0.37	**68.4**

D. Implementation details

D.1. Hyper-parameters

Both comparative methods use per-dataset hyper-parameters in their benchmarks. Since we do not have access to ground truth labels to tune those hyper-parameters in a TTA scenario, for a more rigorous comparison we use the same *fixed hyper-parameters* for all datasets, i.e. the ones they tuned for ImageNet. For TDA, this means the positive logits mixing coefficients is set to 2, while the negative logits mixing coefficient is set to 0.117. For DMN, since we only consider zero-shot scenarios, we only need to set the coefficient relative to the dynamic memory, which is therefore kept fixed at 1. As highlighted in the main paper, the hyper-parameter ν of OGA is always fixed at 0.05.

D.2. Covariance and centroids initialization

The centroids are initialized to the text embeddings, while the covariance is initialized to $\frac{1}{d} I_d$. For each slot in the memory, the entropy is initialized to the maximum possible value, i.e. $-\log(\frac{1}{K})$.

E. Gaussian modeling motivation

Since normalized visual features lie on the unit hypersphere, their distribution obviously cannot be multivariate Gaussian. There are, however, a number of arguments for adopting such a Gaussian model. First, estimating the classification error resulting from approximating hyperspherical distributions with a Gaussian has been analyzed in [6] for 3 families of hyperspherical distributions. Notably, [6] characterizes a set of various hyperspherical distributions for which adopting a Gaussian model does not increase the classification error, compared to using the true distributions. Second, adopting a Gaussian model is further motivated by the fact that estimating high-dimensional distribution parameters from limited samples is challenging. Since the Gaussian is fully characterized by its mean and covariance matrix, it makes it particularly attractive, leveraging extensive literature on high-dimensional covariance estimation. Third, the experimental investigation of the error sources mitigates the importance of a more accurate modeling of the distributions. Specifically, classification errors arise from 3 sources: (a) **Modeling error** (the true distribution lies outside the parametric family), (b) **Parameter estimation error** (incorrect parameter estimation), and (c) **Irreducible error** (overlapping true distributions). To assess whether **Modeling error**

Table 16. Accuracy on the complete test sets for OGA at the end of a run (OGA - endpoint) compared to a model where the cache is filled with samples randomly drawn using ground truth labels (OGA - oracle).

	ViT-B/16	ViT-B/32	ViT-L/14	ResNet50	ResNet101
Zero-Shot	65.3	61.9	72.6	58.7	59.5
OGA - oracle	79.5	75.7	84.7	73.3	75.5
OGA - endpoint	69.1	64.2	75.9	62.0	62.8

(a) dominates other sources for OGA, we run an additional experiment. We compare the accuracy of our OGA model at the end of a run, to a model **with oracle** where the cache is filled with samples randomly drawn using ground truth labels. The accuracy is computed on the entire test set, and results are presented in Table 16. The large gap of accuracy provides strong empirical evidence suggesting that classification errors are dominated by **Parameter estimation error** arising from zero-shot miss classifications as well as the sampling bias induced by the minimal entropy selection rule.