

Online Gaussian Test-Time Adaptation of Vision-Language Models

Clément Fuchs^{*1} Maxime Zanella^{*1,2} Christophe De Vleeschouwer¹
¹UCLouvain, Belgium ²UMons, Belgium

Abstract

Online test-time adaptation (OTTA) of vision-language models (VLMs) has recently garnered increased attention to take advantage of data observed along a stream to improve future predictions. Unfortunately, existing methods rely on dataset-specific hyper-parameters and incomplete evaluation protocols, limiting their generalization to new tasks. Thus, we propose Online Gaussian Adaptation (OGA), a novel method that models the likelihoods of visual features using Gaussian distributions and incorporates zero-shot priors into a concise Maximum A Posteriori (MAP) estimation framework with fixed hyper-parameters across all datasets. To further extend OTTA methods deployment capabilities, we show that combining OTTA with popular few-shot techniques—a practical yet overlooked setting in prior research—is highly beneficial. Besides, our experimental study reveals that common OTTA evaluation protocols, which average performance over at most three runs per dataset, are inadequate due to the substantial variability observed across runs. Hence, we advocate for more rigorous evaluation practices, including increasing the number of runs and considering additional quantitative metrics, such as our proposed Expected Tail Accuracy (ETA), calculated as the average accuracy in the worst 10% of runs. We hope these contributions will encourage more rigorous evaluation practices in the OTTA community, which we believe to be an essential step for real-world deployment in multimodal applications. Code is available at <https://github.com/cfuchs2023/OGA>.

1. Introduction

Vision-Language alignment has emerged as a powerful paradigm for pretraining models capable of handling a wide variety of downstream tasks with little or no labeled data. Contrastive methods such as CLIP [23] learn transferable visual representations by jointly optimizing a visual encoder and a textual encoder to align the representations of paired images and captions. This enables the creation of

^{*} Equal contributions and corresponding authors.
 {clement.fuchs,maxime.zanella}@uclouvain.be

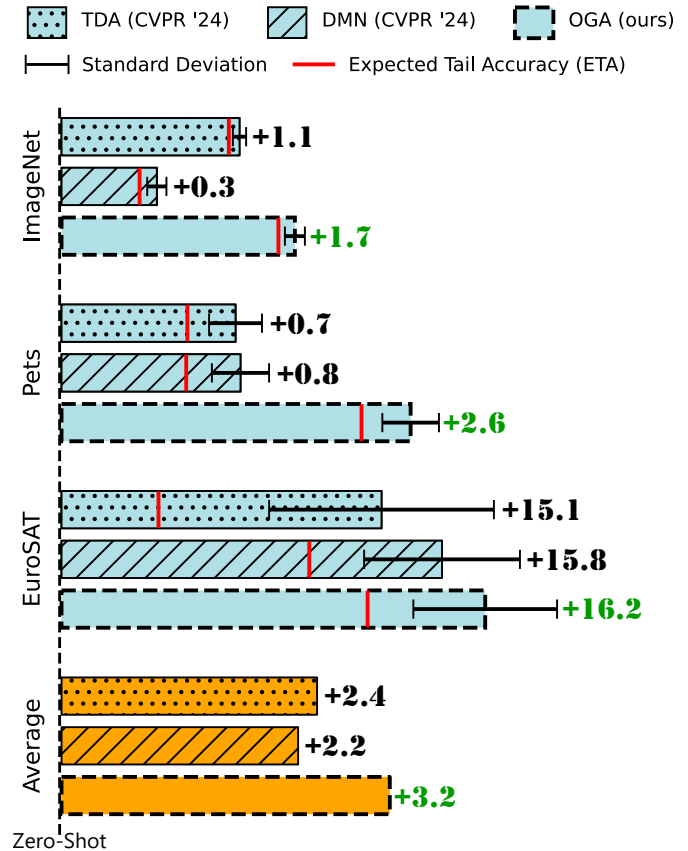


Figure 1. The presented results are averaged over 100 runs. We propose the *Expected Tail Accuracy* (ETA), i.e., the average over the 10% worst runs, in solid red line. Our method named OGA not only significantly outperforms competitors on average but also has an ETA exceeding their average accuracy on several datasets (e.g., ImageNet and Pets). See Table 1 for more detailed results.

an image classifier without retraining the model, using textual descriptions of the classes. The classification procedure then relies simply on measuring the similarities between the textual features and those of the images, enabling zero-shot predictions. This has resulted in impressive zero-shot performance, as demonstrated on widely recognized supervised learning benchmarks such as ImageNet [3]. This success has motivated the investigation of methods to adapt

vision-language models (VLMs) to unseen tasks, circumventing the need for training a model anew, either through prompt optimization [12, 13, 36, 37], low-rank adaptation [30], or adapters in the embedding space [29, 34]. These latter methods are of particular interest as they do not require access to the model weights—so-called black-box methods [21]—making them suitable for API-based applications when few labeled shots are available.

Subsequently, the test-time adaptation (TTA) paradigm, where no labels are available, has attracted significant attention in the Vision-Language community, aiming at exploiting unlabeled data to further improve these performances. Notably, TTA has been deployed through methods which require encoding a large number of augmented views for each image [24, 31] or rely on transductive settings that demand relatively large data batches to discover patterns among instances [10, 19, 32]. These limitations can be overcome when TTA is cast in an online setting, where data arrives in batches as small as one with the possibility of retaining information from one batch to the next ones. Very recent works, such as TDA [11] and DMN [35], utilize cache models that are iteratively updated with incoming data. However, their performance depends on some dataset-specific hyper-parameters in their intricate prediction rule that are adjusted specifically for their experiments. This observation is not new and was recently highlighted in a study [25] on related cache-based methods [34] in the few-shot setting. To mitigate this important practical deployment issue, we propose Online Gaussian Adaptation (OGA) which models the likelihoods of observed visual features with multivariate Gaussian distributions and combines them with the zero-shot priors, yielding a principled *Maximum A Posteriori* (MAP) prediction rule (with no need for *hyper-parameter* tuning). Our approach achieves superior performances, as depicted in Figure 1 and Table 1.

Additionally, our study reveals that, despite their growing popularity, online test-time adaptation (OTTA) methods for VLMs lack rigorous evaluation frameworks. For instance, TDA [11] and DMN [35] evaluate performance using at most three random seeds, even though Figure 1 and Table 1 demonstrate significant variance in measured accuracy across random runs. We propose increasing the number of runs to mitigate variability in comparisons arising from the stochastic nature of data stream generation. Furthermore, we argue that the average accuracy metric is insufficient to accurately compare methods, as it fails to account for *tail risk*, where methods may exhibit much worse accuracies for a small proportion of runs. This behavior could render a method undesirable in practice. Thus, we recommend reporting an additional metric, which we term *Expected Tail Accuracy* (ETA). ETA represents the average accuracy below the lower 10th-percentile, capturing performance in worst-case scenarios.

Contributions. We summarize our contributions as follows:

- We propose Online Gaussian Adaptation (OGA), an OTTA method that models the likelihoods of observed visual features with multivariate Gaussian distributions and combines them with zero-shot priors into an elegant *Maximum A Posteriori* (MAP) prediction rule with *fixed hyper-parameters* across all datasets. Our method delivers strong performance, fits in the blackbox framework and is computationally efficient, as it does not require backpropagation.
- Similar to previous works, we report performances of OTTA methods when applied to zero-shot VLMs. OGA outperforms state-of-the-art methods on most datasets and runs. Additionally, we compare methods atop popular few-shot methods, a very convenient way to combine offline few-shot learning with efficient online adaptation which has been overlooked so far in OTTA.
- Finally, we advocate for more rigorous evaluation procedures in this domain, emphasizing the need for multiple runs to account for variability and introducing *Expected Tail Accuracy* (ETA) as a metric to assess performance in worst-case scenarios.

2. Related work

Fine-tuning of VLMs. One main design choice that differentiates fine-tuning methods is the set of parameters they tune, from input textual tokens [9, 17, 24, 36], hidden layers [30], additional parameters at the output of the text or vision encoder [5, 29], to adapters as memory banks [11, 34, 35]. Others operate directly in the embedding space, for example with a mode-seeking algorithm [31]. One notable group of such methods, sometimes called black-box methods in the literature [21, 31], is undoubtedly cache-based methods. These methods stem from the initial work of Tip-Adapter [34], which explicitly combines logits from zero-shot prediction with similarity scores derived from a memory bank. Other notable advances in black-box methods include the recent successes of Gaussian modeling in few-shot learning [27], transductive settings [14, 32, 33] or human-in-the-loop frameworks [7]. Inspired by these recent developments in related fields, we propose to refine class representations by modeling the likelihoods of observed visual features with multivariate Gaussian distributions. We then use the resulting posterior probabilities obtained from these likelihoods and the zero-shot priors to yield a principled *Maximum A Posteriori* (MAP) prediction rule that is mathematically sound.

Test-Time Adaptation of VLMs. The major distinction between current TTA methods lies in how they process the incoming data. One group of methods operates on a single image with data augmentations at test time, such as

TPT [24] which relies on prompt optimization for each individual image. MTA [31] avoids prompt tuning and optimizes a mean-shift-inspired objective function. However, these methods substantially increase computational requirements. In transductive learning, another branch of unsupervised learning, VLMs are directly adapted to the testing data. For example, EM-Dirichlet [19] optimizes a maximum likelihood estimator of a Dirichlet distribution directly in the prediction space. ZLaP [10] proposes propagating zero-shot labels based on a similarity graph of the representation of each instance. TransCLIP [14, 32, 33] suggests adding a text-based regularization derived from a Kullback-Leibler divergence term in an expectation-maximization-like objective function. One major drawback of these methods is that they rely on relatively large batch sizes, and require multiple samples of the same class within a batch to effectively leverage relationships between instances.

Online Test-Time Adaptation of VLMs. OTTA approaches treat incoming data as a stream, retaining information from one batch to the next ones. A nascent work is [17], although it does not fit in the blackbox framework and uses a computationally expensive strategy combining prompt tuning and augmentations. More recent works provide a highly efficient solution to these issues by maintaining a small cache of selected samples to iteratively improve a prediction rule. TDA [11] and DMN [35] both use a similar minimal-entropy filtering strategy to fill their cache and a prediction rule directly inspired by Tip-Adapter. However, these methods rely on hyper-parameters that are difficult to tune for each new benchmark. In contrast, our approach is practical, using just one hyper-parameter fixed across backbones and datasets to weight the learned likelihoods obtained from the Gaussian models.

3. Preliminaries

To understand recent adaptation methods for vision-language models (VLMs), we start by defining the core components of the classification pipeline. At its foundation, a VLM encodes both images and textual descriptions into a shared embedding space, enabling comparison and alignment. These descriptions are tokenized into textual inputs \mathbf{c}_k , where $1 < k \leq K$ (K the number of classes), which are then transformed by the textual encoder into normalized embeddings \mathbf{t}_k on the unit-hypersphere. The image \mathbf{x}_i , where $i = 1, \dots, N$, is processed by the visual encoder to produce embeddings $\mathbf{f}_i \in \mathbb{R}^d$, where d is the dimension of the embedding space. These embeddings are also normalized to lie on the unit-hypersphere, facilitating direct comparison between images and class descriptions. With this shared embedding space, the cosine similarity between textual and visual representations $\mathbf{f}_i^\top \mathbf{t}_k$ forms the basis for classification tasks.

Zero-shot prediction. Deploying VLMs in a zero-shot setting is one of the simplest and most direct ways to perform downstream tasks, leveraging the pre-training process described in [23]. To classify an image, the similarity between the image embedding and each class embedding is measured using cosine similarity, producing logit scores

$$l_{i,k} = \mathbf{f}_i^\top \mathbf{t}_k. \quad (1)$$

These logits can be transformed into probabilistic predictions through a softmax function, which computes the posterior probability of class k given the test image \mathbf{x}_i

$$y_{i,k} = \frac{\exp(l_{i,k}/\tau)}{\sum_j^K \exp(l_{i,j}/\tau)} \quad (2)$$

where τ is the softmax temperature parameter that controls the sharpness of the probability distribution. The image \mathbf{x}_i can then be classified by selecting the class with the highest posterior probability: $\hat{k} = \operatorname{argmax}_k y_{i,k}$.

Few-shot adaptation. When few shots are available, they can be used to learn richer representations of the classes in the textual embedding space. This is done either (i) by fine-tuning the input prompts (so as to minimize the cross-entropy loss computed on the few available shots), as in prompt-tuning methods like CoOp [36]; or (ii) by updating a set of additional parameters called adapters [34] typically directly at the output of the model such as TaskRes [29]. Respectively, we have:

$$\mathbf{c}_k^{\text{CoOp}} = (\mathbf{v}_k^1, \dots, \mathbf{v}_k^M, [\text{class}_k]); \quad \mathbf{t}_k^{\text{TaskRes}} = \mathbf{t}_k + \alpha \mathbf{b}_k \quad (3)$$

where $(\mathbf{v}_k^l)_{1 \leq l \leq M}$ are trainable text tokens, $[\text{class}_k]$ is the fixed class tokens, \mathbf{b}_k class-wise learnable parameters, and α a scaling hyper-parameter. Observe that prompt tuning incur heavy computational load for fine-tuning and might be hard to optimize, since every gradient update of the text input requires back-propagating through the entire model¹. Note that our method is orthogonal to those advances in the few-shot learning community, in fact we show that our proposed OGA and other OTTA methods can be applied atop of them (see Table 4 with CoOp and TaskRes), offering a very convenient approach where few-shot supervised learning is done offline (potentially with heavy computation) with further adaptation done online using an efficient OTTA method.

Cache model. One of the first works to use a cache for VLMs adaptation is Tip-Adapter [34], which stores few-shot samples. In its training-free version, it directly utilizes

¹We refer to the runtime studies of [11, 31].

the cache for final predictions by combining zero-shot similarities with cache similarities to compute adapted logits,

$$l_{i,k} = \mathbf{f}_i^T \mathbf{t}_k + \alpha \sum_m \exp(-\beta(1 - \mathbf{f}_i^T \mathbf{f}_m^{(k)})) \quad (4)$$

with $\mathbf{f}_m^{(k)} \in \mathbb{R}^d$ the m^{th} sample held in the cache for the k^{th} class, α and β being hyper-parameters. This formula was later used in an online setting by TDA [11]. Note that, unlike Tip-Adapter, TDA relies on pseudo-labels rather than ground truth labels, as it focuses on zero-shot adaptation. A major drawback of these Tip-Adapter-based methods is their dependence on dataset-specific hyper-parameters (α and β) that are carefully tuned for each downstream task [25]. This is done via intensive searches over validation sets, requiring additional labeled samples which reduces their portability to new tasks. Our approach tackles this limitation by relying on a single hyper-parameter, fixed across backbones and datasets, as explained in the next section.

4. Online Gaussian Adaptation

This section introduces our proposal to improve the zero-shot capabilities of a pre-trained VLM, based on a set of samples whose classes are predicted with high confidence. In an online setting, those samples are continuously collected along the stream, to fill in and then update a cache memory. In practice, we select the samples with the smallest zero-shot prediction entropy, i.e. those reliably labeled by the zero-shot classifier. The selected samples are then used to estimate a model of the image features class-conditional likelihoods as multivariate Gaussian distributions. The likelihoods are subsequently combined with the zero-shot prediction, considered as a prior, to estimate the class posterior for a new sample, using a prediction rule derived from Bayes formula. The main steps involved in this process—namely class posterior estimation, Gaussian parameters estimation, and online selection of reliable samples—are detailed below.

Gaussian modeling. Modeling the normalized visual features with a single multivariate Gaussian for each class has been proven effective both for zero-shot and few-shot adaptation of VLMs [7, 14, 27, 32, 33]. We adopt this framework to model the image feature likelihoods conditioned on the class, while presenting additional motivation in Supplementary Material Section E. Hence, for the feature \mathbf{f}_i associated to image i , we have $p_{i,k} = p(\mathbf{f}_i | c_i = k) = p(\mathbf{f}_i | \boldsymbol{\mu}_k, \Sigma, k)$, following a multivariate normal distributions with shared covariance Σ . Formally,

$$p_{i,k} \propto \exp\left(-\frac{1}{2}(\mathbf{f}_i - \boldsymbol{\mu}_k)^T P(\Sigma)(\mathbf{f}_i - \boldsymbol{\mu}_k)\right) \quad (5)$$

where $P(\Sigma)$ is an estimator of the precision matrix Σ^{-1} .

Pseudo-Bayesian adaptation rule. Our proposed adaptation rule is derived from the class posterior probabilities given by the Bayes rule. This posterior reads as

$$p(c_i = k | \mathbf{f}_i) = \frac{p_{i,k} \cdot p(c_i = k)}{p(\mathbf{f}_i)} = \frac{p_{i,k} \cdot p(c_i = k)}{\sum_{l=1}^K p_{i,l} \cdot p(c_i = l)}. \quad (6)$$

In absence of prior knowledge about class probability, the prior $p(c_i = k)$ is generally chosen as $1/K$ to model the features distribution as a balanced mixture of multivariate normals. However, in the case of VLMs, we propose to leverage the knowledge obtained from the zero-shot predictions by using the soft labels $y_{i,k}$ as priors, which yields

$$p(c_i = k | f_i) = \frac{p_{i,k} \cdot y_{i,k}}{\sum_{l=1}^K p_{i,l} \cdot y_{i,l}}. \quad (7)$$

Interestingly, one could remark that Eq. (7) yields a *Maximum A Posteriori* (MAP) estimator for each of the sample. To better control the degree to which the initial zero-shot prediction is modified by the Gaussian likelihoods, we introduce an hyper-parameter ν

$$p(c_i = k | f_i) = \frac{p_{i,k}^\nu \cdot y_{i,k}}{\sum_{l=1}^K p_{i,l}^\nu \cdot y_{i,l}}. \quad (8)$$

We use the same *fixed* value of $\nu = 0.05$ across all datasets and backbones, and investigate its impact in our ablation study (see Figure 3).

Gaussian parameters update. Whenever the cache memory is updated, we also update the Gaussian parameters. First, the centroids $\boldsymbol{\mu}_k$ are updated as the mean of the cached samples for the k^{th} class. Then, the shared covariance matrix is updated using the cached samples as

$$\Sigma = \frac{1}{n-1} \sum_{k=1}^K \sum_m (\mathbf{f}_m^{(k)} - \boldsymbol{\mu}_k)(\mathbf{f}_m^{(k)} - \boldsymbol{\mu}_k)^T \quad (9)$$

where n is the total number of samples in the cache and $\mathbf{f}_m^{(k)}$ the m^{th} cached sample for class k . Note that since we store a relatively low (typically at most 8) number of samples per class, the total number of samples used for estimating Σ can be lower or on the same order of magnitude as the embedding space dimension d . Therefore, in the case where we have less than $4d$ samples in our cache, we use the Bayes-Ridge estimator of [16] which reads as

$$P = d(n_t \Sigma + \text{tr}(\Sigma) I_d)^{-1}. \quad (10)$$

When more than $4d$ samples are in the cache, we revert to using the inverse of Σ as $P(\Sigma)$. More details are provided in the ablation study in Table 7.

Online selection of samples. Similarly to [11], the samples are selected to fill in the cache according to their zero-shot entropy. More specifically, we compute the zero-shot Shannon entropy for a single sample from its zero-shot soft labels as $e_i = -\sum_{k=1}^K \log(y_{i,k})y_{i,k}$. If the sample’s entropy is lower than that of at least one cached sample for the class matching its pseudo-label, we replace the cached sample with the highest entropy with this new one. This process builds a low-entropy cache for each class as the model encounters new data.

5. Experimental setting

Datasets. We follow the settings of previous works [36] and use ImageNet [3] as well as 10 other datasets: SUN397 [28] for classification of scenes, Aircraft [18] for aircraft types, EuroSAT [8] for satellite imagery, StanfordCars [15] for cars models, Food101[1] for food items, Pets [22] for pet types, Flower102 [20] for flowers species, Caltech101 [4] for a variety of general objects, DTD [2] for textures types and UCF101 [26] for actions recognition.

Backbones. We use CLIP with a ViT-B/16 visual architecture for most results presented in the main paper, and provide results with 4 other backbones in Table 2 and Tables 10, 11, 12, 13 (Supplementary Material).

Data stream generation. We generate i.i.d. data streams from the test set of each dataset, and then run the methods on the full stream with batch size 32. For each dataset, the methods are compared on the *same* 100 runs. In our ablation study, we provide further results for our approach for batch sizes 1, 64 and 128 in Table 8.

Competitors. We compare our approach to two recent state-of-the-art works in OTTA, namely TDA (CVPR ’24) [11] and DMN (CVPR ’24) [35]. For the sake of fairness, we use the same total cache size of $8K$ samples for every methods, where K is the number of classes. For TDA, the positive cache has size 5 while the negative cache is set to size 3 for each class.

Data augmentations. We note that our competitors use many computationally expensive augmentations in some settings. Since we do not propose to include such costly computations, we also do not use augmentations when running our competitors methods, so that we can compare performance at similar computational cost. Note that we also report the results of a non-online TTA method, MTA [31], which relies on several augmentations of each image for informational purpose.

Prompts. First, we show results when applied on top of the zero-shot model with (i) handcrafted prompts (provided in Table 9a (Supplementary Material)) and (ii) an ensemble of prompts (provided in Table 9b (Supplementary Material)). Then, we compare the methods when run on top of few-shot adapted models with (i) prompt-tuning method CoOp [36] and (ii) adapter method TaskRes [29]. This comprehensive benchmarking highlights the broad applicability of OTTA methods and more specifically OGA across diverse scenarios. We aim to inspire other works to adopt a similar broad benchmarking methodology in future research.

Hyper-parameters. Our approach is dependent on a hyper-parameter ν (see Eq. (8)). We use the *same fixed value* $\nu = 0.05$ across all datasets, backbones and batch sizes and investigate its impact in Section 7.

Evaluation metrics. We report the average accuracy across 100 runs to mitigate variability in comparison due to the stochastic effects of data streams generation, which was not done in previous studies [11, 35] despite variability in results as demonstrated in Figure 1 and Table 1a. Moreover, we argue that the latter metric is not sufficient to accurately compare methods and is not robust to *tail risk*, where methods could show much worse accuracies for a small proportion of runs. The latter could make a method undesirable in practice. Therefore, we introduce a metric which we call *Expected Tail Accuracy* (ETA) and is the average of accuracies in the 10% worst cases, i.e.,

$$\text{ETA} = \frac{10}{N_{\text{runs}}} \sum_{r=1}^{N_{\text{runs}}} \text{acc}^{(r)} \times \mathbb{1}(\text{acc}^{(r)} \leq \text{acc}_{0.1}) \quad (11)$$

where $\text{acc}^{(r)}$ is the accuracy of run r and $\text{acc}_{0.1}$ the accuracy such that 10% of the runs fall below, and report this additional metric. Note that our approach does not contain any design choice for specifically mitigating these worst case accuracies and we just advocate for better performance reporting practices.

6. Results and discussion

Atop zero-shot. Table 1a shows that OGA performs better than OTTA competitors on 9 out of 11 datasets on average over 100 runs. For the two remaining datasets, our method still places second best. Note that each method is tested using the *same* 100 runs for each dataset, and that we use the *same fixed hyper-parameters* for all datasets. Overall, this proves the effectiveness of our approach. Now we analyse the results to the light of our proposed metric ETA. Notice in Tables 1a and 1b that on several datasets (ImageNet, SUN397, StanfordCars, Pets), the ETA of our

Table 1. All methods are tested on the same 100 runs for each datasets with the same handcrafted prompts of Table 9a (Supplementary Material). The best metric is marked in **bold** while the second best is underlined. For our method named OGA, we show the difference Δ Competitor with the best competitor.

(a) We report the average accuracy as well as the standard deviation over the 100 runs for each method and each dataset. As a reference, we provide the results of a non-online TTA method which relies on augmentations, namely MTA [31].

	ImageNet	SUN397	Aircraft	EuroSAT	StanfordCars	Food101	Pets	Flower102	Caltech101	DTD	UCF101	AVERAGE
Zero-Shot	66.74	62.55	24.87	48.25	65.53	85.88	89.10	70.81	93.35	43.32	67.54	65.3
MTA (CVPR '24)	69.3	64.8	27.4	46.9	68.0	87.2	89.4	71.7	94.0	44.4	69.0	66.6
TDA (CVPR '24)	<u>67.9</u> ±0.07	64.9±0.13	<u>24.7</u> ±0.38	63.4±1.19	66.5±0.22	<u>85.8</u> ±0.07	89.8±0.28	72.7±0.32	93.4 ±0.36	45.0±0.41	70.5±0.35	<u>67.7</u>
DMN (CVPR '24)	67.0±0.10	<u>64.9</u> ±0.17	24.0±0.39	<u>64.0</u> ±0.82	<u>67.0</u> ±0.30	83.9±0.10	<u>89.9</u> ±0.30	73.3 ±0.34	92.6±0.42	44.7±0.62	<u>71.2</u> ±0.41	67.5
OGA (ours)	68.5 ±0.11	66.0 ±0.20	25.3 ±0.38	64.5 ±0.76	67.8 ±0.21	86.1 ±0.07	91.7 ±0.30	<u>72.7</u> ±0.38	<u>93.2</u> ±0.42	45.8 ±0.54	71.6 ±0.37	68.5
Δ Competitor	+0.6	+1.2	+0.5	+0.5	+0.9	+0.3	+1.8	-0.6	-0.2	+0.7	+0.4	+0.8

(b) We report the average accuracy over the 10 worst runs for each method and each dataset, i.e., the ETA (Equation 11).

	ImageNet	SUN397	Aircraft	EuroSAT	StanfordCars	Food101	Pets	Flower102	Caltech101	DTD	UCF101	AVERAGE
Zero-Shot	66.74	62.55	24.87	48.25	65.53	85.88	89.10	70.81	93.35	43.32	67.54	65.3
TDA (CVPR '24)	<u>67.8</u>	<u>64.6</u>	<u>24.0</u>	61.0	66.1	<u>85.7</u>	<u>89.3</u>	<u>72.1</u>	92.5	<u>44.3</u>	69.9	<u>67.0</u>
DMN (CVPR '24)	66.8	64.6	23.3	<u>62.6</u>	<u>66.4</u>	83.7	89.3	72.7	91.6	43.6	<u>70.5</u>	66.8
OGA (ours)	68.3	65.7	24.6	63.2	67.4	85.9	91.2	71.9	<u>92.2</u>	44.9	71.0	67.9

Table 2. We show results averaged over the 11 datasets for 4 different backbones, using 100 runs per method per dataset and the handcrafted prompts of Table 9a (Supplementary Material).

	ViT-B/32	ViT-L/14	ResNet50	ResNet101
Zero-Shot	61.9	72.6	58.7	59.5
TDA (CVPR '24)	<u>62.3</u>	73.5	<u>59.3</u>	60.6
DMN (CVPR '24)	61.8	<u>73.7</u>	58.6	<u>61.0</u>
OGA (ours)	62.9	74.3	59.8	61.6

method is higher than the average accuracy of our competitors, i.e. the worst 10% runs for our method still ranks higher than the average of our competitors.

Moreover, Table 1a shows the ETA of all methods are lower than the zero-shot performance of CLIP on the Aircraft dataset, indicating that they quite often deliver performance below zero-shot. This breakdown demonstrates the value of ETA in providing deeper insights into the results. We also report the accuracy of a non-online state-of-the-art TTA method, MTA [32], which relies on multiple augmentations of the input images and does not retain information from samples. This shows how casting the problem of TTA in an online setting can be highly beneficial, with a striking example being the large gain of more than 15 points of accuracy on EuroSAT. Meanwhile, Figure 2 shows the percentage of runs for which OGA achieves a higher accuracy than TDA and DMN for each dataset. Observe that for 5 datasets (ImageNet, SUN397, StanfordCars, Pets and UCF101), our approach achieves a higher accuracy than TDA for all of the 100 runs used for testing. In comparison with DMN, our method yields a higher accuracy for all of the runs for 6 datasets (ImageNet, SUN397, Aircraft, StanfordCars,

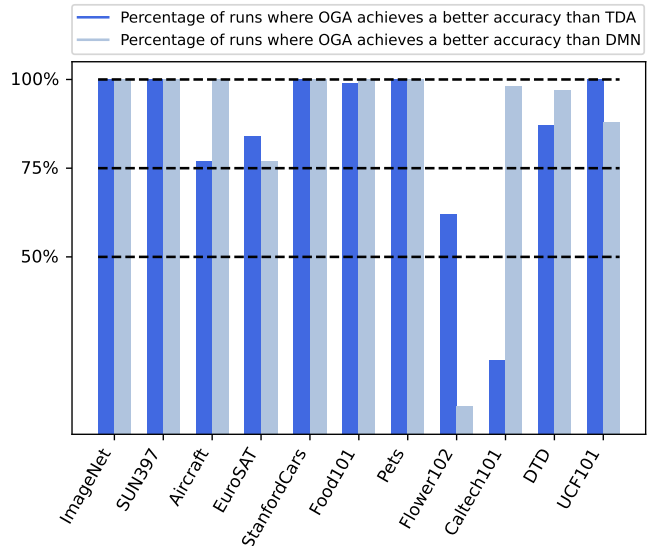


Figure 2. For each dataset, we show the percentage of runs for which our method OGA achieves a higher accuracy than competitors. The experimental setting is the same as the one in Table 1.

Food101, Pets). Finally, we compare the three methods in the same setting but with the ensemble of prompts of Table 9b (Supplementary Material) in the Table 3. In this experiment, our method ranks first for 8 datasets out of 11, and second on the remaining three. Therefore, our approach is robust to changes in the prompts used for zero-shot predictions, a finding further confirmed in the next paragraph.

Table 3. We report the averaged accuracy over the same 100 runs for each method and each dataset. We use the ensemble of prompts of Table 9b (Supplementary Material).

	ImageNet	SUN397	Aircraft	EuroSAT	StanfordCars	Food101	Pets	Flower102	Caltech101	DTD	UCF101	AVERAGE
Zero-Shot	68.73	66.17	23.10	50.54	66.05	85.59	87.90	67.07	93.87	45.15	67.59	65.6
TDA (CVPR '24)	<u>69.3</u> ±0.06	<u>67.5</u> ±0.11	<u>23.1</u> ±0.30	57.3 ±0.68	66.8±0.23	<u>85.3</u> ±0.06	<u>87.9</u> ±0.22	68.7±0.39	94.0 ±0.33	46.4±0.39	69.7±0.33	<u>66.9</u>
DMN (CVPR '24)	68.2±0.10	66.9±0.16	22.7±0.34	51.7±1.25	<u>67.5</u> ±0.28	83.5±0.09	87.8±0.33	71.0 ±0.43	93.2±0.46	<u>47.0</u> ±0.61	<u>70.6</u> ±0.43	66.4
OGA (ours)	69.4 ±0.11	67.9 ±0.16	23.2 ±0.39	<u>54.2</u> ±1.38	68.1 ±0.20	85.6 ±0.07	89.4 ±0.26	<u>69.2</u> ±0.40	<u>93.6</u> ±0.40	47.9 ±0.44	71.4 ±0.41	67.3

Table 4. Results atop popular few-shot methods. For each few-shot method, we train three adapted models with different seeds and run OTTA methods on the same 100 runs per seed. We report the averaged accuracy and standard deviation over the resulting 300 runs. The best metric is marked in **bold** while the second best is underlined.

(a) CoOp [36] is a popular prompt-tuning method for few-shots adaptation, which adds learnable tokens to to the texts defining the classes (see Equation 3).

	ImageNet	SUN397	Aircraft	EuroSAT	StanfordCars	Food101	Pets	Flower102	Caltech101	DTD	UCF101
1 shot											
CoOp	65.7	66.9	20.8	56.4	67.5	84.3	90.2	78.3	92.5	50.1	71.2
+ TDA (CVPR '24)	<u>66.8</u> ±0.04	<u>68.2</u> ±0.07	<u>21.9</u> ±0.24	<u>61.7</u> ±0.38	68.1±0.15	<u>84.6</u> ±0.05	<u>90.4</u> ±0.14	80.7±0.23	92.9 ±0.29	51.5±0.21	73.1±0.17
+ DMN (CVPR '24)	66.5±0.07	68.2±0.11	21.6±0.24	62.0 ±0.55	<u>69.0</u> ±0.21	83.7±0.07	89.9±0.20	82.7 ±0.24	92.6±0.34	<u>52.0</u> ±0.39	<u>73.8</u> ±0.34
+ OGA (ours)	67.6 ±0.07	69.1 ±0.12	22.1 ±0.20	61.6±0.59	69.6 ±0.13	85.2 ±0.05	91.3 ±0.17	<u>81.0</u> ±0.30	<u>92.8</u> ±0.33	52.5 ±0.34	74.3 ±0.25
4 shots											
CoOp	68.8	69.7	30.8	69.7	74.4	84.3	<u>92.5</u>	92.2	94.5	59.4	77.5
+ TDA (CVPR '24)	<u>69.4</u> ±0.04	<u>70.6</u> ±0.06	<u>31.2</u> ±0.20	73.7±0.34	<u>74.8</u> ±0.13	<u>84.9</u> ±0.03	92.4±0.12	92.9±0.14	<u>94.5</u> ±0.30	60.9±0.19	78.9±0.17
+ DMN (CVPR '24)	68.6±0.06	70.5±0.09	31.0±0.25	<u>73.8</u> ±0.40	74.6±0.18	84.0±0.07	91.7±0.17	93.4 ±0.14	94.4±0.31	<u>61.3</u> ±0.28	<u>79.1</u> ±0.28
+ OGA (ours)	69.7 ±0.06	71.5 ±0.11	31.7 ±0.24	75.3 ±0.38	76.1 ±0.12	84.9 ±0.07	93.0 ±0.14	<u>92.9</u> ±0.23	94.5 ±0.31	61.6 ±0.27	79.8 ±0.21

(b) TaskRes [29] is a popular adapter method which adds a bias to the text embedding of each class (see Equation 3).

	ImageNet	SUN397	Aircraft	EuroSAT	StanfordCars	Food101	Pets	Flower102	Caltech101	DTD	UCF101
1 shot											
TaskRes	69.6	68.1	<u>31.2</u>	65.7	<u>69.1</u>	84.5	<u>90.1</u>	81.6	<u>93.6</u>	53.4	71.8
+ TDA (CVPR '24)	70.1 ±0.06	<u>69.3</u> ±0.08	30.7±0.22	<u>69.5</u> ±0.44	68.9±0.15	<u>84.9</u> ±0.04	90.1±0.15	83.6±0.30	93.9 ±0.27	<u>55.3</u> ±0.27	<u>73.1</u> ±0.20
+ DMN (CVPR '24)	68.7±0.09	68.2±0.12	30.2±0.26	69.5±0.54	68.9±0.18	83.4±0.09	89.4±0.19	85.9 ±0.29	92.8±0.38	54.4±0.44	72.8±0.30
+ OGA (ours)	<u>69.9</u> ±0.09	69.4 ±0.14	31.5 ±0.24	70.6 ±0.53	70.9 ±0.12	85.5 ±0.09	91.3 ±0.18	<u>84.1</u> ±0.33	<u>93.4</u> ±0.35	55.8 ±0.40	73.9 ±0.26
4 shots											
TaskRes	<u>71.0</u>	<u>72.8</u>	<u>33.2</u>	73.9	<u>76.1</u>	<u>86.1</u>	<u>91.9</u>	85.0	<u>94.8</u>	59.6	75.5
+ TDA (CVPR '24)	71.3 ±0.05	73.2 ±0.06	32.9±0.24	76.1 ±0.31	75.3±0.15	85.9±0.04	91.6±0.14	87.3±0.29	94.9 ±0.31	<u>61.3</u> ±0.30	<u>76.6</u> ±0.20
+ DMN (CVPR '24)	69.6±0.08	71.5±0.11	32.1±0.26	73.5±0.43	74.5±0.19	83.9±0.09	90.6±0.18	88.7 ±0.28	94.2±0.36	59.5±0.41	75.9±0.30
+ OGA (ours)	70.7±0.09	72.6±0.13	33.5 ±0.26	<u>74.4</u> ±0.49	77.4 ±0.12	86.2 ±0.07	92.3 ±0.18	<u>87.3</u> ±0.36	94.7±0.30	61.8 ±0.37	77.2 ±0.26

Atop few-shot. In Table 4, we report the results atop two popular few-shot adaptation methods. For CoOp (Table 4a), a prompt-learning method, our approach yields the strongest improvement, performing better on average for 8 datasets out of 11 in the 1-shot setting and for 10 out of 11 datasets in the 4-shot setting. For TaskRes (Table 4b), an adapter method, our approach also achieves the highest overall accuracy gain, ranking first for 8 datasets out of 11 in the 1-shot setting. In the 4-shot setting, our method achieves highest accuracy for 6 datasets out of 11. Interestingly, we observe that the few-shot adaptation reduces the variability of OTTA method on nearly every dataset. Finally, we see that in the vast majority of the cases, the OTTA methods improve over the few-shot adapted model, which proves the benefits of using OTTA atop adapted models.

With other backbones. A summary of results obtained with two other ViT based (ViT-L/14 and ViT-B/32) and CNN based (ResNet101 and ResNet50) backbones is shown in Table 2, while the detailed results are available in the

Supplementary Material in Tables 10, 11, 12, 13, and 14. All backbones show similar trends.

7. Ablation studies

Likelihood weighting hyper-parameter ν . Our method uses the same fixed hyper-parameter $\nu = 0.05$ (see Equation 8) for all experiments and datasets. It controls the degree to which the Gaussian likelihood is pushed away from the uniform distribution. Therefore, when $\nu = 0$, our MAP degenerates to the zero-shot prior. Following, it is expected that higher values of ν are detrimental when the Gaussian modeling is poor (e.g., at the beginning of a run). Figure 3 illustrates that our choice of hyper-parameter is essentially a trade-off between mitigating early transitory effects, when the cache is either empty or filled with poor quality samples, and end point accuracy. This interesting observation could pave the way for improving our method by designing an adaptive rule to adapt ν as a function of the state of the cache.

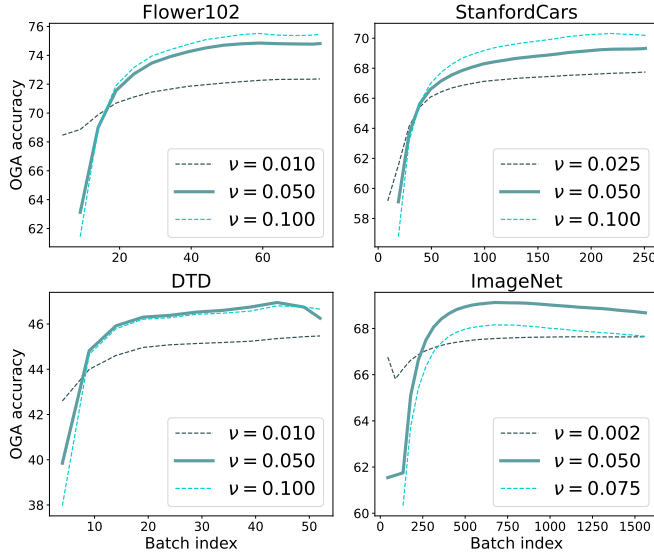


Figure 3. We show the dynamic of the accuracy of our OGA method as it starts from an empty cache, averaged on 100 runs. At regular intervals, we evaluate the accuracy of OGA on the complete test set.

Zero-shot prior. We show results obtained without the zero-shot prior (i.e., $\nu \rightarrow \infty$) in Table 5, highlighting its relevance across all backbones.

Table 5. Ablation study on the zero-shot prior. We report the averaged accuracy over the 11 datasets for each backbone.

	ViT-B/32	ViT-B/16	ViT-L/14	ResNet50	ResNet101
OGA w/o prior	60.35	65.5	73.0	57.3	59.2
OGA (ours)	62.8	68.5	74.7	59.8	61.6

Size of the cache. We show results with different cache sizes in Table 6, i.e., the maximum number of cached samples per class. This illustrates how the cache size is a trade-off between diversity and contamination with incorrectly labeled samples.

Table 6. Ablation study on the size of the cache for our method. We report the averaged accuracy over the 11 datasets.

OGA w/ cache size	4	8	16	32
AVERAGE	67.8	68.5	68.0	67.0

Precision matrix estimation. We show that it is beneficial to use different estimators depending on the number of samples in the cache. To do so, we run our method either

with only the Ridge estimator or only the (pseudo-)inverse, and present results in Table 7.

Table 7. Ablation study on the use of two different estimators instead of one. We report the averaged accuracy over the 11 datasets.

OGA w/	Ridge and inverse	only inverse	only Ridge
AVERAGE	68.5	66.6	68.3

Batch size. In all experiments, we process the data streams in batches of 32 samples. In Table 8, we show that our method is able to process the streams sample by sample and that it benefits from increased batch sizes. The latter is due to the fact that cache-based methods are quicker to fill their cache with quality samples when the batch size increases, as the cache is updated before predicting. Note our approach still achieves a higher average accuracy in batch size 1 (detailed results in Table 15 (Supplementary Material)) compared to our competitors in batch size 32.

Table 8. Ablation study on the batch size used to process the data streams. We report the averaged accuracy over the 11 datasets.

OGA w/ batch size	1	32	64	128
AVERAGE	68.4	68.5	68.5	68.6

8. Conclusion

In this study, we proposed Online Gaussian Adaptation (OGA), a method for the online test-time-adaptation of VLMs. Our method uses a modeling of the class-conditional likelihoods of visual features with multivariate Gaussians, which are estimated from low-entropy samples collected along the data stream. We compared our approach to state-of-the-art methods with a rigorous evaluation protocol, inspired by the significant variability in the measured accuracy observed between runs. Using 100 runs per dataset and our proposed *Expected Tail Accuracy* (ETA) metric which captures the performance in worst-case scenarios, we showed that our method delivers strong performance with *fixed hyper-parameter* across datasets and backbones. Lastly, we showed that applying OTTA methods on top of few-shot learning methods, either prompt-tuning or adapter, is highly beneficial. We hope our work will encourage more rigorous and diverse evaluation practices in the OTTA community, and facilitate deployment of OTTA methods in real-world multimodal applications.

Future works. As highlighted in our ablation study, an interesting avenue to explore is the design of an adaptive rule for our hyper-parameter ν (Equation 8) depending on the state of the cache and the strength of the zero-shot prior.

9. Acknowledgments

C. Fuchs is funded by the MedReSyst project, supported by FEDER and the Walloon Region. M. Zanella is funded by the Walloon region under grant No. 2010235 (ARIAC by DIGITALWALLO-NIA4.AI). Part of the computational resources have been provided by the Consortium des Équipements de Calcul Intensif (CÉCI), funded by the Fonds de la Recherche Scientifique de Belgique (F.R.S.-FNRS) under Grant No. 2.5020.11 and by the Walloon Region.

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. 5
- [2] Mircea Cimpoi, Subhansu 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. 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, 2009. 1, 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. 5
- [5] Peng Gao, Shijie Geng, Renrui Zhang, Teli Ma, Rongyao Fang, Yongfeng Zhang, Hongsheng Li, and Yu Qiao. Clip-adapter: Better vision-language models with feature adapters. *International Journal of Computer Vision*, 132(2): 581–595, 2024. 2
- [6] Onur C Hamsici and Aleix M Martinez. Spherical-homoscedastic distributions: The equivalency of spherical and normal distributions in classification. *Journal of Machine Learning Research*, 8(7), 2007. 3
- [7] Zongbo Han, Jialong Yang, Junfan Li, Qinghua Hu, Qianli Xu, Mike Zheng Shou, and Changqing Zhang. Dota: Distributional test-time adaptation of vision-language models. *arXiv preprint arXiv:2409.19375*, 2024. 2, 4
- [8] 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. 5
- [9] Tony Huang, Jack Chu, and Fangyun Wei. Unsupervised prompt learning for vision-language models. *arXiv preprint arXiv:2204.03649*, 2022. 2
- [10] Yannis Kalantidis, Giorgos Tolias, et al. Label propagation for zero-shot classification with vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23209–23218, 2024. 2, 3
- [11] Adilbek Karmanov, Dayan Guan, Shijian Lu, Abdulmotaheb 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, 3, 4, 5
- [12] Muhammad Uzair Khattak, Hanoona Rasheed, Muhammad Maaz, Salman Khan, and Fahad Shahbaz Khan. Maple: Multi-modal prompt learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19113–19122, 2023. 2
- [13] Muhammad Uzair Khattak, Syed Talal Wasim, Muzammal Naseer, Salman Khan, Ming-Hsuan Yang, and Fahad Shahbaz Khan. Self-regulating prompts: Foundational model adaptation without forgetting. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 15190–15200, 2023. 2
- [14] Karim El Khoury, Maxime Zanella, Benoît Gérin, Tiffanie Godelaine, Benoît Macq, Saïd Mahmoudi, Christophe De Vleeschouwer, and Ismail Ben Ayed. Enhancing remote sensing vision-language models for zero-shot scene classification. *arXiv preprint arXiv:2409.00698*, 2024. 2, 3, 4
- [15] 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. 5
- [16] Tatsuya Kubokawa and Muni S Srivastava. Estimation of the precision matrix of a singular wishart distribution and its application in high-dimensional data. *Journal of Multivariate Analysis*, 99(9):1906–1928, 2008. 4
- [17] Xiaosong Ma, Jie Zhang, Song Guo, and Wenchao Xu. Swapprompt: Test-time prompt adaptation for vision-language models. *Advances in Neural Information Processing Systems*, 36, 2024. 2, 3
- [18] Subhansu Maji, Esa Rahtu, Juho Kannala, Matthew Blaschko, and Andrea Vedaldi. Fine-grained visual classification of aircraft. *arXiv preprint arXiv:1306.5151*, 2013. 5
- [19] Ségolène Martin, Yunshi Huang, Fereshteh Shakeri, Jean-Christophe Pesquet, and Ismail Ben Ayed. Transductive zero-shot and few-shot clip. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 28816–28826, 2024. 2, 3
- [20] 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. 5
- [21] Yassine Ouali, Adrian Bulat, Brais Matinez, and Georgios Tzimiropoulos. Black box few-shot adaptation for vision-language models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15534–15546, 2023. 2
- [22] 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. 5

- [23] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 1, 3
- [24] Manli Shu, Weili Nie, De-An Huang, Zhiding Yu, Tom Goldstein, Anima Anandkumar, and Chaowei Xiao. Test-time prompt tuning for zero-shot generalization in vision-language models. *Advances in Neural Information Processing Systems*, 35:14274–14289, 2022. 2, 3
- [25] Julio Silva-Rodríguez, Sina Hajimiri, Ismail Ben Ayed, and Jose Dolz. A closer look at the few-shot adaptation of large vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 23681–23690, 2024. 2, 4
- [26] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012. 5
- [27] Zhengbo Wang, Jian Liang, Lijun Sheng, Ran He, Zilei Wang, and Tieniu Tan. A hard-to-beat baseline for training-free clip-based adaptation. In *The Twelfth International Conference on Learning Representations*, 2024. 2, 4
- [28] 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. 5
- [29] Tao Yu, Zhihe Lu, Xin Jin, Zhibo Chen, and Xinchao Wang. Task residual for tuning vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10899–10909, 2023. 2, 3, 5, 7
- [30] Maxime Zanella and Ismail Ben Ayed. Low-rank few-shot adaptation of vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 1593–1603, 2024. 2
- [31] Maxime Zanella and Ismail Ben Ayed. On the test-time zero-shot generalization of vision-language models: Do we really need prompt learning? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23783–23793, 2024. 2, 3, 5, 6
- [32] Maxime Zanella, Benoît Gérin, and Ismail Ben Ayed. Boosting vision-language models with transduction. *Neural Information Processing Systems (NeurIPS)*, 2024. 2, 3, 4, 6
- [33] Maxime Zanella, Fereshteh Shakeri, Yunshi Huang, Houda Bahig, and Ismail Ben Ayed. Boosting vision-language models for histopathology classification: Predict all at once. In *International Workshop on Foundation Models for General Medical AI*, pages 153–162. Springer, 2024. 2, 3, 4
- [34] Renrui Zhang, Wei Zhang, Rongyao Fang, Peng Gao, Kunchang Li, Jifeng Dai, Yu Qiao, and Hongsheng Li. Tip-adapter: Training-free adaption of clip for few-shot classification. In *European conference on computer vision*, pages 493–510. Springer, 2022. 2, 3
- [35] Yabin Zhang, Wenjie Zhu, Hui Tang, Zhiyuan Ma, Kaiyang Zhou, and Lei Zhang. Dual memory networks: A versatile adaptation approach for vision-language models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 28718–28728, 2024. 2, 3, 5
- [36] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *International Journal of Computer Vision*, 130(9):2337–2348, 2022. 2, 3, 5, 7
- [37] Beier Zhu, Yulei Niu, Yucheng Han, Yue Wu, and Hanwang Zhang. Prompt-aligned gradient for prompt tuning. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 15659–15669, 2023. 2