On the test-time zero-shot generalization of vision-language models: Do we really need prompt learning?

Maxime Zanella* UCLouvain UMons
Ismail Ben Ayed ÉTS Montréal

code: https://github.com/MaxZanella/MTA

2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) | 979-8-3503-5300-6/24/$31.00 ©2024 IEEE | DOI: 10.1109/CVPR52733.2024.02245

Abstract

*The development of large vision-language models, notably CLIP, has catalyzed research into effective adaptation techniques, with a particular focus on soft prompt tuning. Conjointly, test-time augmentation, which utilizes multiple augmented views of a single image to enhance zero-shot generalization, is emerging as a significant area of interest. This has predominantly directed research efforts toward test-time prompt tuning. In contrast, we introduce a robust **M**eanShift for **T**est-time **A**ugmentation (MTA), which surpasses prompt-based methods without requiring this intensive training procedure. This positions MTA as an ideal solution for both standalone and API-based applications. Additionally, our method does not rely on ad hoc rules (e.g., confidence threshold) used in some previous test-time augmentation techniques to filter the augmented views. Instead, MTA incorporates a quality assessment variable for each view directly into its optimization process, termed as the inlierness score. This score is jointly optimized with a density mode seeking process, leading to an efficient training- and hyperparameter-free approach. We extensively benchmark our method on 15 datasets and demonstrate MTA's superiority and computational efficiency. Deployed easily as plug-and-play module on top of zero-shot models and state-of-the-art few-shot methods, MTA shows systematic and consistent improvements.*

1. Introduction

Vision-language models, pretrained on vast sets of image-text pairs, have emerged as powerful tools for learning cross-modal representations [25, 30, 46, 60, 62, 63]. The joint feature space of visual and textual features enables zero-shot recognition, without any task-specific data. For instance, given a set of candidate classes, one can create textual descriptions, with the so-called prompt [34], $\mathbf{p}_k$ = ''`a photo of a` [`class`$_k$]", and get its corresponding embedding representation $\mathbf{t}_k = \theta_t(\mathbf{p}_k)$ with the language encoder. Similarly, an image $\mathbf{x}$ is projected in the same embedding space $\mathbf{f} = \theta_v(\mathbf{x})$ using the visual encoder. Then, one can classify this image by measuring the similarity between these two encoded modalities and predicting the class corresponding to the most similar embedding, $\hat{k} = \arg\max_k \mathbf{f}^t \mathbf{t}_k$.

Despite their impressive capabilities, these models still encounter substantial challenges and may yield unsatisfactory responses in complex situations [17, 46]. These issues are particularly pronounced when confronted with the pragmatic constraints of real-world scenarios, where labeled data can be scarce (i.e., few-shot scenarios [52]) or completely absent (i.e., zero-shot scenarios [28]), thus limiting their broader usage. Consequently, there has been a growing interest in enhancing the test-time generalization faculties of these vision-language models [14, 16, 36, 38, 40, 45]. Empirical findings indicating that improved textual descriptions can positively impact zero-shot predictions [46] have sparked interest in refining prompt quality for downstream tasks. Originating in the NLP community [23, 26, 49], soft prompt learning, which utilizes learnable continuous tokens as input [29], has rapidly gained popularity. Building on this momentum, CoOp [68] stands out as the seminal work for prompt tuning in vision-language models. Since then, prompt tuning has appeared as the prominent approach for adapting vision-language models [63] across both unsupervised [14, 24, 38] and few-shot scenarios [3, 5, 12, 35, 59, 67–69].

In parallel, test-time augmentation, which has been extensively used in the computer vision community [32, 50, 64], is now emerging in the vision-language field, with a focus on prompt tuning [14, 36, 38]. Instead of exploiting a single image $\mathbf{x}$, *test-time prompt tuning* techniques leverage multiple embeddings $(\mathbf{f}_p)_{1\leq p\leq N}$, each derived from a different augmented view $(\mathbf{x}_p)_{1\leq p\leq N}$ of the same original im-

*Corresponding author: maxime.zanella@uclouvain.be

This work was partly supported by the Walloon Region (Service Public de Wallonie Recherche, Belgium) under grant n°2010235 (ARIAC by DigitalWallonia4.ai).

2575-7075/24/$31.00 ©2024 IEEE
DOI 10.1109/CVPR52733.2024.02245

age $\mathbf{x}$. Afterwards, the prompt is optimized by forcing consistency of the predictions among these different views [38]. The final classification step is then performed by computing the similarity between the original image encoding and the *optimized* textual embedding $\mathbf{t}_k^*$, $\hat{k} = \arg\max_k \mathbf{f}^t \mathbf{t}_k^*$. These novel research directions underscore the increasing attention in enhancing these models' robustness, especially in zero-shot scenarios, through data augmentation at test-time. Alongside this expanding literature, we ask the following question: *Can we improve the image representation* $\mathbf{f}$ *directly in the embedding space, achieving superior results in a way that is more efficient than prompt tuning?*

Concurrently, there has been a surge in the use of proprietary and closed APIs that encapsulate advanced machine learning functionalities, often termed *black boxes* due to their limited transparency, offering little insight into their internal mechanisms or architectures. Yet, they are crucial in executing a wide spectrum of tasks in vision and NLP, introducing new challenges in model adaptation [51]. The field of NLP, in particular, has seen an emerging literature on few-shot adaptation of black-box models [8], driven by the reality that large-scale models (e.g., GPT family [2, 41], Palm [6]) are only accessible via APIs and their pretrained weights are not publicly available. Optimizing prompts, which necessitates gradient computation from output back to input, a memory-intensive and time-consuming process, is impractical in the context of API-reliant applications. In contrast, our approach does not require extra assumptions about the model's internal states or architecture, making it suitable for black-box applications.

Contributions. In this work, we introduce a robust multimodal **M**eanShift **T**est-time **A**ugmentation (MTA), which enhances the zero-shot generalization of CLIP models, leveraging different augmented views of a given image. Unlike current prompt tuning solutions, which rely on heavy training procedures and *ad hoc* thresholds to discard degenerated views, MTA uses only the final embedding state and directly integrates an *inlierness* assessment of the augmented views into its optimization process. Our objective function is efficiently solvable using iterative block coordinate descent updates, and relaxes the need for training the model's parameters or prompts. Empirically, we demonstrate that MTA surpasses state-of-the-art prompt-tuning alternatives, while being time and memory efficient. Our key contributions are as follows:

1. We propose a robust MeanShift formulation, which automatically manages augmented views in test-time augmentation scenarios by optimizing *inlierness* variables. Used as a versatile *plug-and-play* tool, MTA improves the zero-shot performances of various models on a large variety of classification tasks, without any hyperparameter tuning.
2. We report comprehensive evaluations and comparisons to the existing test-time prompt tuning techniques on 15 datasets, showing MTA's highly competitive performances, although it operates in limited access (i.e., final embedding) and training-free mode. This makes MTA suitable for both standalone and API-based applications.
3. Deployed easily atop current state-of-the-art few-shot learning methods, MTA brings consistent improvements, a benefit not observed with test-time prompt tuning.

2. Related works

Vision-language models adaptation. Large scale vision-language models have shown excellent results in several vision tasks [63]. This success has created interest in developing adaptation techniques that capitalize their general knowledge [57]. Among these, prompt tuning [29] has emerged as the primary method for adapting CLIP-like models, at test-time based on data augmentations [14, 36, 38] or with few labeled samples [3, 5, 12, 35, 59, 67–69]. CoOp [68] optimizes learnable continuous tokens attached to the class name, while CoCoOp [67] trains a neural network to generate instance-conditioned tokens based on the image. Further efforts include ProGrad [69], which guides prompts toward predefined handcrafted ones based on gradients, whereas PLOT [5] aligns learned prompts with finer-grained visual features via an optimal transport formulation. Beyond soft prompt tuning, other strategies involve using hierarchical word structures to create more semantically refined class descriptions [16, 40], or exploiting other large scale models to generate more detailed prompts [45, 54, 66] or new images by diffusion mechanisms [14, 66].
Contrastingly, methods such as CLIP-Adapter [15] offer an alternative strategy by learning feature adapters. However, there has been limited effort in developing black-box methods [42], which can effectively capitalize the knowledge of these models while only accessing their final embedding state. Examples include zero-shot prediction with parameter-free plug-in attention [18], or few-shot settings with Tip-Adapter [65] using a cache model.
Our experiments demonstrate that our robust MeanShift algorithm significantly enhances the performances in zero-shot scenarios, without relying on soft prompt tuning, while respecting the black-box constraints. Additionally, we report increased performances when applied atop of various aforementioned few-shot methods, without requiring further training or hyperparameter tuning.

Test-time augmentation. Data augmentation during training is widely recognized for its capacity to enhance model robustness [20, 21]. Also, its utility extends to test-time applications [32, 50, 64]. In particular, test-time augmentation can be used on a single image [64] to adapt models with an entropy minimization term. The latter is of-

ten used in the context of unsupervised adaptation [31, 55], but is deployed differently in this augmentation setting, enforcing consistent predictions across the various augmented views. This idea is further developed for vision-language models with test-time prompt tuning (TPT) [38], where a prompt is optimized to make consistent predictions among light augmentations inspired by Augmix [20]. DiffTPT [14] builds up on this work by adding generated images from Stable Diffusion [48] to acquire more diverse views. Both works show improvements when selecting a subset of the augmented views. Specifically, TPT utilizes only the 10% most confident views, and DiffTPT measures the similarity with the original image, keeping the unconfident but correctly classified augmentations. We also demonstrate that filtering the augmented views can substantially improve test-time augmentation techniques. Additionally, our method does not rely on arbitrary hard thresholds or rules as in TPT and DiffTPT; instead, we directly integrate the weighting of the augmented views in our optimization procedure thanks to *inlierness* variables.

3. Robust multi-modal MeanShift

Similarly to the test-time generalization setting recently introduced in TPT [38], let us assume that we are given a set of image samples $(\mathbf{x}_p)_{1\leq p\leq N}$, which correspond to N distinct augmented views of a given test sample $\mathbf{x}$. It is important to note that our method is applicable on top of any type of augmentations. Let $\mathbf{f}_p = \theta_v(\mathbf{x}_p)$ denote the vision-encoded feature embedding corresponding to augmented sample $\mathbf{x}_p$, θ_v being the vision encoder of the CLIP pre-trained model. A straightforward way to use the ensemble of augmentations is to perform the zero-shot prediction based on their mean embedding, thereby giving exactly the same importance to all augmented samples, independently of the structure of the data. However, the augmentations may include degenerated views, which correspond to *outliers*, e.g., in the form of isolated data points or small regions with little structure within the feature space. Such outliers may bias global statistics like the mean.

3.1. Formulation

We hypothesize that *robust statistics* like the modes of the density of the set of augmented views could provide better representations. Augmented views presenting major characteristics of the concept to be recognized are likely to be projected close to the original image and close to each others. This motivates our formulation, which could be viewed as a novel robust and multi-modal extension of the popular MeanShift algorithm [9], an unsupervised procedure for finding the modes of the distribution of a given set of samples. In our case, we explicitly model and handle the potential presence of outliers in the estimation of the kernel density of the set of feature vectors $(\mathbf{f}_p)_{1\leq p\leq N}$. To do so, we introduce a latent assignment vector $\mathbf{y} = (y_p)_{1\leq p\leq N} \in \Delta^{N-1}$, with $\Delta^{N-1} = \{\mathbf{y} \in [0,1]^N \mid \mathbf{1}^t\mathbf{y} = 1\}$ the probability simplex, and propose to minimize the following objective function:

$$\min_{\mathbf{m},\mathbf{y}} \mathcal{L}(\mathbf{m},\mathbf{y}) \quad \text{s.t.} \quad \mathbf{y} \in \Delta^{N-1} \quad \text{with}$$
$$\mathcal{L}(\mathbf{m},\mathbf{y}) = -\sum_{p=1}^{N} y_p K(\mathbf{f}_p - \mathbf{m}) - \frac{\lambda}{2}\sum_{p,q} w_{p,q} y_p y_q$$
$$- \lambda_{\mathbf{y}} H(\mathbf{y}) \tag{1}$$

In the following, we describe the notations occurring in our model in Eq. (1), as well as the effect of each of its terms:

Robust KDE (first term) K is a kernel function measuring a robust affinity between $\mathbf{f}_p$ and $\mathbf{m}$, e.g., a Gaussian kernel [4]: $K(\mathbf{f}_p - \mathbf{m}) \propto \exp(-\|\mathbf{f}_p - \mathbf{m}\|^2/h^2)$, where h is the kernel bandwidth. When variables y_p are fixed to 1 $\forall p$, the first term reduces to the kernel density estimate (KDE) of the distribution of features at point $\mathbf{m}$. Clearly, minimizing this term w.r.t $\mathbf{m}$ yields the standard MeanShift algorithm for finding the mode of the density (i.e., the point maximizing it). In our model, the additional latent variable y_p evaluates the *inlierness* of the p^{th} augmented view, i.e., the model's belief in $\mathbf{f}_p$ being an inlier or not within the whole set of augmented-view embeddings $(\mathbf{f}_p)_{1\leq p\leq N}$. Score $y_p \in [0,1]$ is high when the model considers the p^{th} sample as an inlier, enabling it to contribute more in the KDE evaluation in (1), and small (closer to 0) otherwise.

Text-knowledge guided quadratic term (second term) This term encourages samples with nearby text-based zero-shot predictions to have similar *inlierness* scores y_p. Specifically, we construct the pairwise affinities $w_{p,q}$ in the second term of (1) from both the text and vision embeddings as follows. Let $\mathbf{s}_p \in \mathbb{R}^K$ denotes the text-driven softmax prediction based on the zero-shot text embedding for the p^{th} sample, i.e., the k^{th} component of $\mathbf{s}_p$ is given by:

$$s_{p,k} = \frac{\exp l_{p,k}}{\sum_{j=1}^{K} \exp l_{p,j}};\ l_{p,k} = \tau \mathbf{f}_p^t \mathbf{t}_k \tag{2}$$

where τ is the temperature scaling parameter of the CLIP model. Affinities $w_{p,q}$ are given by:

$$w_{p,q} = \mathbf{s}_p^t \mathbf{s}_q \tag{3}$$

The Shannon entropy (third term) $H(\mathbf{y})$ is the Shannon entropy defined over simplex variables as follows:

$$H(\mathbf{y}) = -\sum_{p=1}^{N} y_p \ln y_p \tag{4}$$

This term acts as a barrier function, forcing latent variable $\mathbf{y}$ to stay within the probability simplex. Furthermore, this entropic regularizer is necessary to avoid a trivial solution minimizing the quadratic term (i.e., $y_p = 1 \quad \exists p$.), as it pushes the solution toward the middle of the simplex (i.e., $y_p = 1/N \ \forall p$).

3.2. Block coordinate descent optimization

Our objective in (1) depends on two types of variables: $\mathbf{m}$ and $\mathbf{y}$. Therefore, we proceed with block-coordinate descent alternating two sub-steps: one optimizing (1) w.r.t $\mathbf{y}$ and keeping density modes $\mathbf{m}$ fixed, while the other minimizes (1) w.r.t $\mathbf{m}$ with the *inlierness* variables fixed.

Optimization w.r.t $\mathbf{y}$ via the concave-convex procedure When $\mathbf{m}$ is fixed, our objective $\mathcal{L}(\mathbf{y}, \mathbf{m})$ could be minimized efficiently w.r.t $\mathbf{y}$ using the Concave-Convex Procedure (CCCP) [61], with convergence guarantee. At each iteration, we update the current solution $\mathbf{y}^{(n)}$ as the minimum of a tight upper bound on $\mathcal{L}$, which ensures the objective does not increase. For the sum of concave and convex functions, as for our sub-problem, the CCCP replaces the concave part by its linear first-order approximation at the current solution, which is a tight upper bound, while keeping the convex part. For (1), the quadratic term could be written as $\mathbf{y}^t W \mathbf{y}$, with $W = [w_{i,j}]$. It is easy to see that this term is concave as affinity matrix W is positive semi-definite, whereas the remaining part of $\mathcal{L}$ is convex. Therefore, we replace this quadratic term by $\mathbf{y}^t W^t \mathbf{y}^{(n)}$, obtaining, up to an additive constant, the following tight bound:

$$\mathcal{L}(\mathbf{y}, \mathbf{m}) \overset{c}{\leq} -\sum_{p=1}^{N} y_p K(\mathbf{f}_p - \mathbf{m}) - \lambda \mathbf{y}^t W^t \mathbf{y}^{(n)} - \lambda_{\mathbf{y}} H(\mathbf{y}) \tag{5}$$

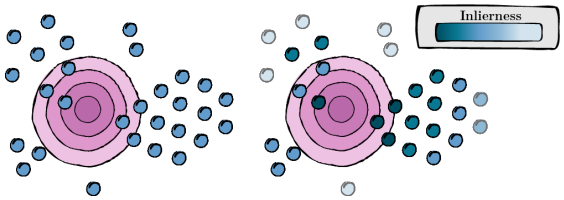
Solving the Karush-Kuhn-Tucker (KKT) conditions for minimizing bound (9), s.t. simplex constraint $\mathbf{y} \in \Delta^{N-1}$, gives the following updates for $\mathbf{y}$:

$$y_p^{(n+1)} = \frac{\exp\left((K(\mathbf{f}_p - \mathbf{m}) + \lambda \sum_{q=1}^{N} w_{p,q} y_q^{(n)})/\lambda_{\mathbf{y}}\right)}{\sum_{j=1}^{N} \exp\left((K(\mathbf{f}_j - \mathbf{m}) + \lambda \sum_{q=1}^{N} w_{j,q} y_q^{(n)})/\lambda_{\mathbf{y}}\right)} \tag{6}$$

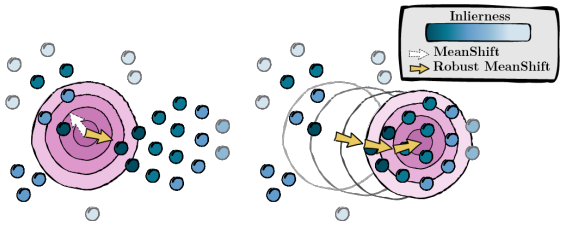
which have to be iterated until convergence. The complete derivation of Eq. (6) is provided in Appendix A.

Optimization w.r.t $\mathbf{m}$ via fixed-point iterations This sub-step fixes $\mathbf{y}$, and minimizes the objective in (1) w.r.t the density modes $\mathbf{m}$. Setting the gradient of $\mathcal{L}$ w.r.t $\mathbf{m}$ to 0 yields the following necessary condition for a minimum, which takes the form of a fixed-point equation:

$$\mathbf{m} - g(\mathbf{m}) = 0; \quad g(\mathbf{m}) = \frac{\sum_{p=1}^{N} y_p K(\mathbf{f}_p - \mathbf{m}) \mathbf{f}_p}{\sum_{p=1}^{N} y_p K(\mathbf{f}_p - \mathbf{m})} \tag{7}$$



(a) ***Inlierness scores distribution.*** Blue gradation represents the *inlierness* score: the darker, the higher.



(b) ***Inlierness-density mode seeking.*** The white arrow represents the update direction of the traditional MeanShift; instead our robust MeanShift follows the orange ones.

Figure 1. Interpretation of Eqs. (6) and (8) in (a) and (b) respectively. Our robust MeanShift alternatively solves these 2 equations until convergence. A pseudocode is available in Appendix C.

The solution to (7) could be obtained by the following fixed-point iterations:

$$\mathbf{m}^{l+1} = g(\mathbf{m}^l) = \frac{\sum_{p=1}^{N} y_p K(\mathbf{f}_p - \mathbf{m}^l) \mathbf{f}_p}{\sum_{p=1}^{N} y_p K(\mathbf{f}_p - \mathbf{m}^l)} \tag{8}$$

This yields a Cauchy sequence $\{\mathbf{m}^l\}_{l \in \mathbb{N}}$, which converges to a unique value: $\mathbf{m}^* = \lim_{l \to \infty} \mathbf{m}^{l+1} = \lim_{l \to \infty} g(\mathbf{m}^l) = g(\lim_{l \to \infty} \mathbf{m}^l) = g(\mathbf{m}^*)$, and $\mathbf{m}^*$ is the unique solution of the fixed-point in (7) (Appendix B).

Final prediction The class prediction is computed by the cosine similarity between the mode minimizing our objective in Eq. (1), i.e., $\mathbf{m}^*$ obtained at convergence, and the encoded prompts of each class k, i.e., $\mathbf{t}_k$:

$$\hat{k} = \arg\max_k \ (\mathbf{m}^*)^t \mathbf{t}_k$$

Interpretation Eqs. (6) and (8) can be nicely interpreted in Figure 1. Sub-figure 1a shows how the *inlierness* scores are spreading. A data point is given a high *inlierness* score if it is close to the mode and/or close to other data points with high *inlierness* scores. Therefore, *inlierness* scores are spreading iteratively from data points close to the mode toward other data points controlled by their affinity relations. Sub-figure 1b shows the update direction followed by traditional MeanShift, i.e., updates (8) but with fixed $y_p = 1, \forall p$. The orange arrows follow the update of our robust MeanShift, i.e. joint updates in (6) and (8). Our mode update is directed toward dense regions with high *inlierness* scores, thus avoiding close regions with few or isolated data points.

Table 1. Zero-shot methods on ImageNet datasets. We use "`a photo of a` [`class`$_k$]" as prompt. Majority vote of the 80 handcrafted prompts of [46] is used for final prediction when Ensemble is specified. CoOp is a pretrained prompt on 16-shots ImageNet. We highlight the best and second best results by **bolding** and underlining them, respectively. ✓ for training-free methods at test-time, ✗ otherwise.

Method		ImageNet	-A	-V2	-R	-Sketch	Average
CLIP [46]	✓	66.73	47.87	60.86	73.98	46.06	59.11
CLIP + Ensemble [46]	✓	68.38	49.95	62.1	77.38	47.96	61.15
TPT [38]	✗	68.94	54.63	63.41	77.04	47.97	62.40
MTA (Ours)	✓	69.29	57.41	63.61	76.92	48.58	63.16
MTA + Ensemble (Ours)	✓	**70.08**	**58.06**	**64.24**	**78.33**	**49.61**	**64.06**
CoOp [68]	✓	71.51	49.71	64.20	75.21	47.99	61.72
TPT + CoOp [38]	✗	73.61	57.85	66.69	77.99	49.59	65.14
MTA + CoOp (Ours)	✓	**73.99**	**59.29**	**66.97**	**78.2**	**49.96**	**65.68**

Table 2. Zero-shot methods on 10 fine-grained classification datasets. We highlight the best result by **bolding**. +E. means that majority vote with the 80 handcrafted prompts of [46] is used for final prediction.

Method	SUN397	Aircraft	EuroSAT	Cars	Food101	Pets	Flower102	Caltech101	DTD	UCF101	Average
CLIP [46]	62.59	23.67	42.01	65.48	83.65	88.25	67.44	93.35	44.27	65.13	63.58
CLIP + E. [46]	66.02	23.88	**48.8**	66.14	83.83	**88.42**	67.8	93.87	46.04	66.77	65.16
TPT [38]	65.41	23.1	42.93	66.36	84.63	87.22	**68.86**	94.12	**46.99**	68.00	64.76
MTA (Ours)	64.98	25.32	38.71	68.05	84.95	88.22	68.26	94.13	45.59	68.11	64.63
MTA +E. (Ours)	**66.67**	**25.2**	45.36	**68.47**	**85.00**	88.24	68.06	**94.21**	45.9	**68.69**	**65.58**

Table 3. Comparison of augmentation strategies: RandomCrop Vs Diffusion-based. Note that the test set of each dataset is reduced to 1000 samples for computation reasons and batch size of 128 was used as in the DiffTPT paper [14] (64 randomly cropped and 63 diffusion-generated images for Diffusion). Hence reported performance can vary from Table 1. We highlight the best result by **bolding**.

Augmentation	Method	ImageNet	-A	-V2	R	-Sketch	Average
RandomCrop	TPT [38]	68.15	51.23	**66.17**	76.88	49.31	62.35
	MTA	**69.11**	**55.27**	65.71	**77.48**	**50.23**	**63.56**
Diffusion	DiffTPT [14]	67.83	53.43	**65.18**	**76.85**	50.2	62.7
	MTA	**69.18**	**54.5**	64.81	76.82	**51.09**	**63.28**

4. Experimental settings

4.1. Datasets

To evaluate our proposed MTA, we follow the setting of previous works [38, 68]. We assess our method on ImageNet [11] and its four variants (ImageNet-A [22], ImageNet-V2 [47], ImageNet-R [21], ImageNet-Sketch [56]) to measure robustness to natural domain shifts. Additionally, we also consider 10 datasets for fine-grained classification of scenes (SUN397 [58]), aircraft types (Aircraft [37]), satellite imagery (EuroSAT [19]), automobiles (StanfordCars [27]), food items (Food101 [1]), pet breeds (OxfordPets [43]), flowers (Flower102 [39]), general objects (Caltech101 [13]), textures (DTD [7]) and human actions (UCF101 [53]). These diverse datasets provide a comprehensive benchmark for visual classification tasks.

4.2. Implementation details

No hyperparameter tuning. In the MeanShift algorithm, the bandwidth is a sensitive hyperparameter that can cause the algorithm to become stuck in small, locally dense areas if set too low, or escape significant dense regions if set too high [10]. We utilize the Gaussian kernel [4] as described in Section 3 and adopt a variable bandwidth [10] in which each point is assigned a unique bandwidth value h_p^2. The bandwidth of a point is estimated with a ratio ρ of its closest neighbors: $h_p^2 = \frac{1}{\rho(N-1)} \sum_{q \in I_p} \|\mathbf{f}_p - \mathbf{f}_q\|^2$ with ρ set to 0.3 inspired by [44]. Here, I_p represents the set of indices corresponding to the neighbors of point p. Initial guess of the mode can also impact the final solution and is set to the embedding of the original (i.e., non-augmented) image. Affinities are based on the text prediction as described in Section 3. Finally, λ and $\lambda_{\mathbf{y}}$ are set to 4 and 0.2 respectively

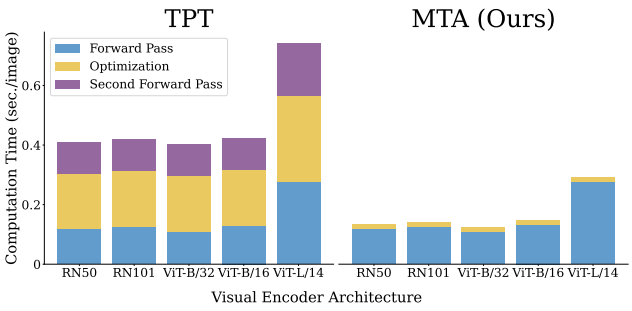


Figure 2. Runtime in seconds per image on ImageNet for TPT and MTA with 5 different backbones: RN50 (ResNet-50), RN101 (ResNet-101), ViT-B/32, ViT-B/16 and ViT-L/14. Experiences were performed on a single A100 40Gb GPU.

and remain fixed for every CLIP visual encoder and dataset. This leads to a hyperparameter-free method across all experiments. Unless otherwise mentioned, we report top-1 accuracy using the ViT-B/16 backbone.

Comparison methods. We employ the term ensemble for the majority vote among the 80 predefined handcrafted prompts of CLIP [46]. For the zero-shot scenario, TPT is performed with one step as suggested in their work [38]. We evaluate the zero-shot setting [28] with the basic "`a photo of a` [$\texttt{class}_k$]" as prompt initialization and with the ensemble for final prediction if mentioned. Note that combining ensemble with TPT is not straightforward as the goal is to use the optimized prompt for prediction, hence we only use ensemble with our approach. The weights of CoOp [68] are from 16 shots Imagenet with 4 tokens. For the few-shot scenario, the number of tokens of CoOp [68] and ProGrad [69] are specified in the caption of each table. As proposed in [33], to establish a fair comparison between few-shot methods, we limit the number of validation samples to $min(n, 4)$ where n is the number of training shots, notably for Tip-Adapter and Tip-Adapter-F [65] that tune hyperparameters. Because of the randomness inherent in test-time augmentation or some training procedures (e.g., prompt tuning), each reported performance is the averaged top-1 accuracy of three different random seeds. More detailed performances are available in Appendix D.

Test-time augmentation. Our proposed method uses random cropping (RandomCrop) as augmented view generator identical to the one in TPT [38]. We also study at the end of Section 5 the impact of a more complex data augmentation based on diffusion as done in DiffTPT [14]. Note that TPT and DiffTPT are using slightly stronger augmentations for the 10 fine-grained classification datasets inspired by AugMix [20]. In our case, for a more realistic zero-shot setting, we keep the simple RandomCrop for all the 15 datasets.

Table 4. Averaged top-1 accuracy on ImageNet and its 4 variants for different visual encoders of CLIP. Hyperparameters are kept identical across datasets and visual encoder architectures. +E. stands for Ensemble.

Architecture	ResNet-50	ResNet-101	ViT-B/32	ViT-B/16	ViT-L/14
CLIP	44.11	49.48	50.65	59.11	70.65
Ensemble	46.23	51.58	52.23	61.12	72.6
TPT	47.23	52.16	53.27	62.40	73.67
MTA	47.22	53.15	55.17	63.16	73.88
MTA + E.	**48.35**	**54.23**	**56.1**	**64.06**	**74.71**

5. Zero-shot

MTA globally outperforms TPT while respecting the black-box constraints and running nearly three times as fast. Table 1 demonstrates MTA's stable improvement over TPT on ImageNet and its variants with both the basic "`a photo of a` [$\texttt{class}_k$]" and CoOp's pretrained prompt. Moreover, Table 2 indicates that, except for the EuroSAT dataset [19], MTA consistently enhances baseline performance across fine-grained datasets. It also surpasses TPT when combined with ensembling, which respects the black-box assumption. TPT is not able to outperform the majority vote strategy in average for the 10 fine-grained classification datasets, questioning its applicability for a larger variety of tasks. Finally, we compare runtimes on ImageNet in Figure 2. MTA is nearly three times faster than TPT, notably due to the quick optimization step and the removed second forward pass.

MTA benefits from improved prompt strategy. Table 1 and 2 concur in suggesting that the improvements brought by our approach complement those from refined prompt strategies, i.e., basic ensemble of handcrafted prompts or soft pretrained prompts. This may imply that leveraging both modalities in the same optimization process, as suggested in [33] for the few-shot setting, could further enhance performance.

MTA bridges the transferability gap of pretrained prompts due to better image representation. We use CoOp pretrained on ImageNet to measure the domain generalization ability of our method as in [68] in Table 1. Although the gains of CoOp over the ensembling strategy remain ambiguous, the combination of CoOp with MTA significantly enhances accuracy for all datasets. This notable improvement suggests that the mode found by MTA is able to highlight the generalization ability of pretrained prompts. This illustrates an additional use case of MTA, which can be used in synergy with fine-tuning for a specific task (e.g., prompt tuning). This aspect is emphasized in Section 6.

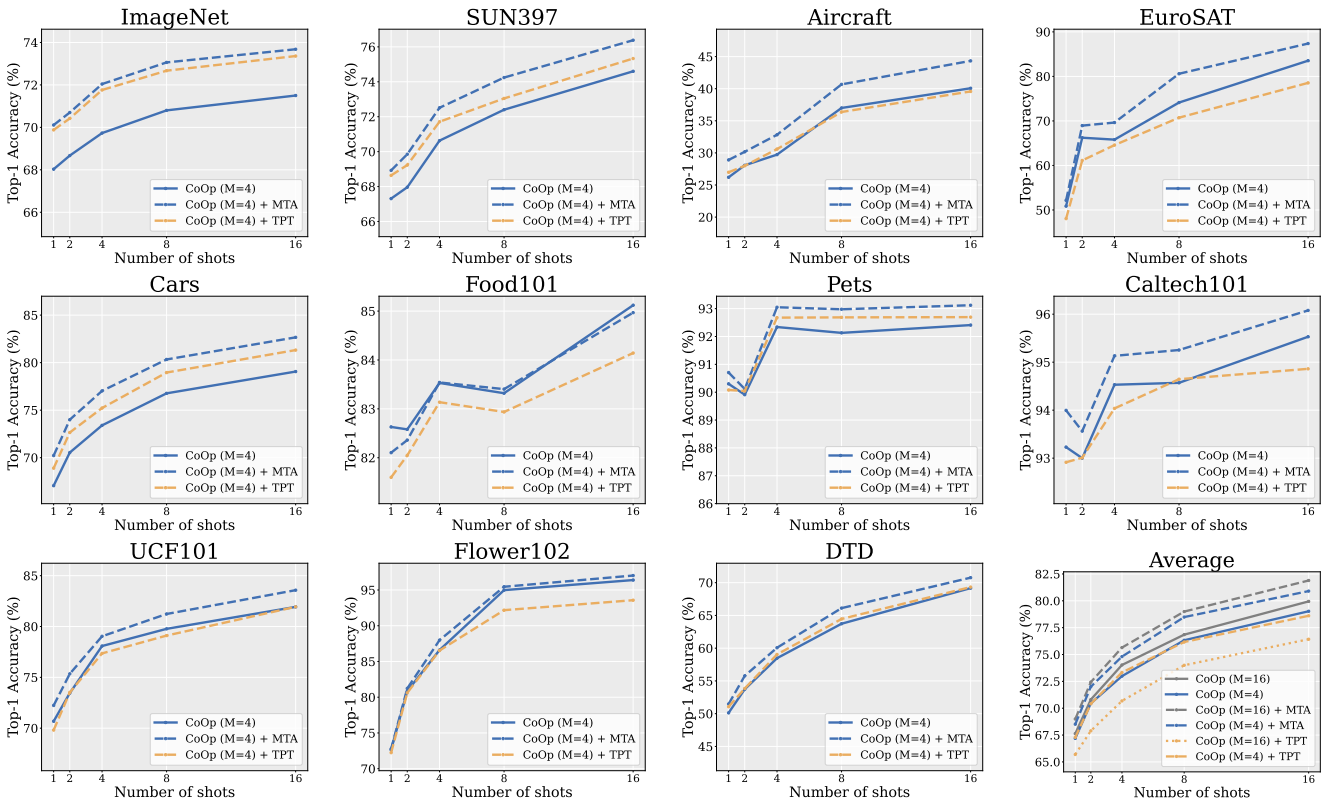


Figure 3. Few-shot learning results on the 10 fine-grained datasets and ImageNet. We compare MTA and TPT when added on top of CoOp prompts (M=4 tokens) for increasing number of shots. Averaged top-1 accuracy over the 11 datasets is shown on the bottom right, we additionally show the averaged top-1 accuracy for CoOp with M=16 tokens.

MTA generalizes across visual encoder architectures with the same hyperparameters. Generalization across various model architectures is a desirable attribute of zero-shot methods, as it eliminates the need for laborious and computationally demanding hyperparameter tuning. This becomes increasingly critical in the context of the current scaling trend, where model complexity is rapidly expanding [2, 6]. Table 4 shows consistent improvement on the baseline for five different visual backbones of CLIP with fixed hyperparameters. It demonstrates the generalization ability of our method across architectures and model scales.

MTA is applicable to other data augmentation strategies. We explore the compatibility of MTA with different types of data augmentation. Specifically, we follow the protocol of DiffTPT [14] to generate augmented views from cropping and diffusion model [48]. We employ a batch of 128 images composed of the original one, 63 diffusion-generated and 64 randomly cropped views. Since generating images by diffusion is more computationally intensive than RandomCrop (generating 63 images takes approximately 2 minutes on a A100 40Gb GPU), we present these results separately. Inspired by the observations of DiffTPT about the reliance of the images generated by diffusion, we increase ρ to a slightly less restrictive value of 0.5 for the purpose of this experiment. Otherwise, we keep the same hyperparameters λ and λ_y and do not treat the two kinds of augmentation differently. As demonstrated in Table 3, our method surpasses DiffTPT on average, further evidencing its effectiveness across a broad range of applications.

6. Few-shot

Fine-grained learned prompts benefit from MTA but not from TPT. As depicted in Figure 3, MTA improves CoOp performances across shots and datasets with notably Aircraft 40.07% $\rightarrow$ 44.33% (**+ 4.26%**), EuroSAT 83.53% $\rightarrow$ 87.38% (**+ 3.85%**), and Cars 79.06% $\rightarrow$ 82.66% (**+ 3.6%**) with 16 shots. In contrast, TPT generally diminishes performance across most datasets, highlighting its limitations in enhancing prompts for fine-grained tasks. While the accuracy for 16 tokens is on average higher for MTA and lower for TPT, we report the 4 tokens results in alignment with the TPT paper [38] to maintain fairness. Detailed performances for 16 tokens and individual datasets are available in Appendix D.

Table 5. Improvement of few-shot learning methods on ImageNet when MTA is added on top. CoOp and ProGrad are using 16 tokens. Δ highlights the gain. ✓ for training-free methods, ✗ otherwise.

Shots		1	2	4	8	16
Tip-Adapter [65]	✓	68.94	69.18	69.75	70.15	70.51
Tip-Adapter-F [65]	✗	69.36	69.95	70.74	71.82	73.39
CoOp [68]	✗	65.7	66.97	68.83	70.57	71.87
ProGrad [69]	✗	67.01	69.06	70.15	71.25	72.14
Tip-Adapter + MTA	✓	71.05	71.12	71.4	71.9	72.05
Δ		(+2.11)	(+1.94)	(+1.65)	(+1.75)	(+1.54)
Tip-Adapter-F + MTA	✗	71.35	71.48	72.17	73.18	74.23
Δ		(+1.99)	(+1.53)	(+1.43)	(+1.36)	(+0.84)
CoOp + MTA	✗	67.82	69.1	71.05	72.85	74.20
Δ		(+2.12)	(+2.13)	(+2.22)	(+2.28)	(+2.33)
ProGrad + MTA	✗	69.27	71.39	72.50	73.66	74.41
Δ		(+2.26)	(+2.33)	(+2.35)	(+2.41)	(+2.27)

MTA can be applied atop few-shot learning methods. As demonstrated in Table 5, applying MTA to prompt-based and adapter-based few-shot learning methods on ImageNet results in notable performance improvements. Specifically, prompt-based methods see an average gain exceeding **2%** across shots, with adapter-based methods showing slightly lower yet significant improvements. Indeed, our affinity term, outlined in Eq. 3, can greatly benefit from refined prompts. Nevertheless, we can notice that a training-free approach, Tip-Adapter, combined with MTA is competitive with prompt tuning methods without MTA. This could present a compelling trade-off: opting for more intensive computational efforts during training or at test-time.

7. Ablation study

Number of augmented views. In Figure 4, the accuracy for MTA and MTA + Ensemble increases as the number of augmented views grows until reaching a plateau around 128. Even when we restrict the number of augmented views to 16, our method still brings about **2.3**% gain, which makes it a useful tool for lighter applications, up until **4.4**% gain for batches of 128 views. We can observe similar gains when combined with ensemble of prompts, with **1.9**% gain for batches of 16 up until **3.1**% for larger batches.

***Inlierness* scores.** We compare performance with equal weights on each augmented view (i.e., traditionnal MeanShift), with a confidence threshold as in TPT [38] and with our *inlierness* formulation in Table 6. The *inlierness* formulation yields better performance on average over the 15 datasets. Note that the relatively high score of confidence threshold is mainly due to peak performance on ImageNet-A, a trend not consistent on other datasets, see Appendix D. Additionally, Appendix C contains a study of λ and $\lambda_{\mathbf{y}}$ showing their interdependent relation and importance.

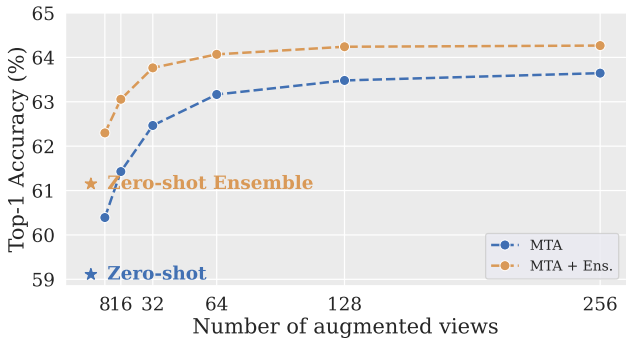


Figure 4. Averaged top-1 accuracy of MTA with and without majority vote for final prediction on the 5 ImageNet variants with increasing number of augmented views.

Affinity measure. Rather than using the affinity measure based on text described in Section 3, one could use only visual features $(\mathbf{f}_p)_{1\leq p\leq N}$ to compute the affinity between two embeddings $w_{p,q} = \mathbf{f}_p^t\mathbf{f}_q$. For the sake of fairness, we reperform a comprehensive hyperparameter search (λ ranging from 0.1 to 10 and $\lambda_{\mathbf{y}}$ from 0.1 to 0.5). Best performance is reported in Table 6. We observe that vision-based affinity measure degrades performance in comparison to text-based affinity measure.

Table 6. Ablation study on two main components of MTA: the *inlierness* scores and the affinity measure. Reported value is the averaged top-1 accuracy over the 15 datasets studied in this work.

Baseline	CLIP	62.09
Filtering Strategy	MeanShift (no *inlierness* scores)	59.93
	Confidence thresh. (10%)	63.27
	Inlierness scores	**64.14**
Affinity Measure	Vision-based	63.15
	Text-based	**64.14**

8. Conclusion

In this work, we have investigated a novel approach to handle test-time augmentation for vision-language models. Our **M**eanShift for **T**est-time **A**ugmentation (MTA) is based on a robust generalization of a well-known mode seeking algorithm, and operates solely on the final embeddings, in a training-free manner. Extensive experiments demonstrate that our method not only surpasses test-time prompt tuning alternatives but also runs significantly faster. Without any other requirements, MTA can easily be deployed in a zero-shot manner and atop few-shot learning methods. We believe our work could serve as a starting point to broaden the current research focus on improving zero-shot vision-language models, and to investigate more efficient approaches, beyond prompt learning.

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] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020. 2, 7

[3] Adrian Bulat and Georgios Tzimiropoulos. Lasp: Text-to-text optimization for language-aware soft prompting of vision & language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 23232–23241, 2023. 1, 2

[4] Miguel Á. Carreira-Perpiñán. Gaussian mean-shift is an em algorithm. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29:767–776, 2007. 3, 5

[5] Guangyi Chen, Weiran Yao, Xiangchen Song, Xinyue Li, Yongming Rao, and Kun Zhang. Plot: Prompt learning with optimal transport for vision-language models. In *The Eleventh International Conference on Learning Representations*, 2022. 1, 2

[6] Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. Palm: Scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311*, 2022. 2, 7

[7] 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. 5

[8] Pierre Colombo, Victor Pellegrain, Malik Boudiaf, Victor Storchan, Myriam Tami, Ismail Ben Ayed, Celine Hudelot, and Pablo Piantanida. Transductive learning for textual few-shot classification in api-based embedding models. *arXiv preprint arXiv:2310.13998*, 2023. 2

[9] Dorin Comaniciu and Peter Meer. Mean shift analysis and applications. In *Proceedings of the seventh IEEE international conference on computer vision*, pages 1197–1203. IEEE, 1999. 3, 13

[10] Dorin Comaniciu, Visvanathan Ramesh, and Peter Meer. The variable bandwidth mean shift and data-driven scale selection. In *Proceedings Eighth IEEE International Conference on Computer Vision. ICCV 2001*, pages 438–445. IEEE, 2001. 5

[11] 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. 5

[12] Mohammad Mahdi Derakhshani, Enrique Sanchez, Adrian Bulat, Victor Guilherme Turrisi da Costa, Cees GM Snoek, Georgios Tzimiropoulos, and Brais Martinez. Variational prompt tuning improves generalization of vision-language models. *arXiv preprint arXiv:2210.02390*, 2022. 1, 2

[13] 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

[14] Chun-Mei Feng, Kai Yu, Yong Liu, Salman Khan, and Wangmeng Zuo. Diverse data augmentation with diffusions for effective test-time prompt tuning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2704–2714, 2023. 1, 2, 3, 5, 6, 7

[15] 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*, pages 1–15, 2023. 2

[16] Yunhao Ge, Jie Ren, Andrew Gallagher, Yuxiao Wang, Ming-Hsuan Yang, Hartwig Adam, Laurent Itti, Balaji Lakshminarayanan, and Jiaping Zhao. Improving zero-shot generalization and robustness of multi-modal models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11093–11101, 2023. 1, 2

[17] Gabriel Goh, Nick Cammarata, Chelsea Voss, Shan Carter, Michael Petrov, Ludwig Schubert, Alec Radford, and Chris Olah. Multimodal neurons in artificial neural networks. *Distill*, 6(3):e30, 2021. 1

[18] Ziyu Guo, Renrui Zhang, Longtian Qiu, Xianzheng Ma, Xupeng Miao, Xuming He, and Bin Cui. Calip: Zero-shot enhancement of clip with parameter-free attention. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 746–754, 2023. 2

[19] 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, 6

[20] Dan Hendrycks*, Norman Mu*, Ekin Dogus Cubuk, Barret Zoph, Justin Gilmer, and Balaji Lakshminarayanan. Augmix: A simple method to improve robustness and uncertainty under data shift. In *International Conference on Learning Representations*, 2020. 2, 3, 6

[21] 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. 2, 5

[22] 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. 5

[23] Z hong, D Friedman, and D Chen. Factual probing is [mask]: Learning vs. learning to recall. In *Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, 2021. 1

[24] Tony Huang, Jack Chu, and Fangyun Wei. Unsupervised prompt learning for vision-language models. *arXiv preprint arXiv:2204.03649*, 2022. 1

[25] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International conference on machine learning*, pages 4904–4916. PMLR, 2021. 1

[26] Z Jiang, F Xu, J Araki, and G Neubig. How can we know what language models know. In *Association for Computational Linguistics (ACL)*, 2020. 1

[27] 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

[28] Christoph H Lampert, Hannes Nickisch, and Stefan Harmeling. Learning to detect unseen object classes by between-class attribute transfer. In *2009 IEEE conference on computer vision and pattern recognition*, pages 951–958. IEEE, 2009. 1, 6

[29] Brian Lester, Rami Al-Rfou, and Noah Constant. The power of scale for parameter-efficient prompt tuning. *arXiv preprint arXiv:2104.08691*, 2021. 1, 2

[30] Liunian Harold Li, Pengchuan Zhang, Haotian Zhang, Jianwei Yang, Chunyuan Li, Yiwu Zhong, Lijuan Wang, Lu Yuan, Lei Zhang, Jenq-Neng Hwang, et al. Grounded language-image pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10965–10975, 2022. 1

[31] Jian Liang, Dapeng Hu, and Jiashi Feng. Do we really need to access the source data? source hypothesis transfer for unsupervised domain adaptation. In *International conference on machine learning*, pages 6028–6039. PMLR, 2020. 3

[32] Jian Liang, Ran He, and Tieniu Tan. A comprehensive survey on test-time adaptation under distribution shifts. *arXiv preprint arXiv:2303.15361*, 2023. 1, 2

[33] Zhiqiu Lin, Samuel Yu, Zhiyi Kuang, Deepak Pathak, and Deva Ramanan. Multimodality helps unimodality: Cross-modal few-shot learning with multimodal models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19325–19337, 2023. 6

[34] Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Computing Surveys*, 55(9): 1–35, 2023. 1

[35] Yuning Lu, Jianzhuang Liu, Yonggang Zhang, Yajing Liu, and Xinmei Tian. Prompt distribution learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5206–5215, 2022. 1, 2

[36] Xiaosong Ma, Jie Zhang, Song Guo, and Wenchao Xu. Swapprompt: Test-time prompt adaptation for vision-language models. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. 1, 2

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

[38] Shu Manli, Nie Weili, Huang De-An, Yu Zhiding, Goldstein Tom, Anandkumar Anima, and Xiao Chaowei. Test-time prompt tuning for zero-shot generalization in vision-language models. In *NeurIPS*, 2022. 1, 2, 3, 5, 6, 7, 8

[39] 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

[40] Zachary Novack, Julian McAuley, Zachary Chase Lipton, and Saurabh Garg. Chils: Zero-shot image classification with hierarchical label sets. In *International Conference on Machine Learning*, pages 26342–26362. PMLR, 2023. 1, 2

[41] OpenAI. Gpt-4v(ision) technical work and authors, 2023. 2

[42] 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

[43] 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

[44] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. Scikit-learn: Machine learning in python. *Journal of machine learning research*, 12(Oct):2825–2830, 2011. 5

[45] Sarah Pratt, Ian Covert, Rosanne Liu, and Ali Farhadi. What does a platypus look like? generating customized prompts for zero-shot image classification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15691–15701, 2023. 1, 2

[46] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision, 2021. 1, 5, 6, 13

[47] 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. 5

[48] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10684–10695, 2022. 3, 7

[49] T Shin, Logan R. L. IV Razeghi, Y, E Wallace, and S Singh. Autoprompt: Eliciting knowledge from language models with automatically generated prompts. In *Empirical Methods in Natural Language Processing (EMNLP)*, 2020. 1

[50] Connor Shorten and Taghi M Khoshgoftaar. A survey on image data augmentation for deep learning. *Journal of big data*, 6(1):1–48, 2019. 1, 2

[51] Irene Solaiman. The gradient of generative ai release: Methods and considerations, 2023. 2

[52] Yisheng Song, Ting Wang, Puyu Cai, Subrota K Mondal, and Jyoti Prakash Sahoo. A comprehensive survey of few-shot

learning: Evolution, applications, challenges, and opportunities. *ACM Computing Surveys*, 2023. 1

[53] 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

[54] Vishaal Udandarao, Ankush Gupta, and Samuel Albanie. Sus-x: Training-free name-only transfer of vision-language models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2725–2736, 2023. 2

[55] Dequan Wang, Evan Shelhamer, Shaoteng Liu, Bruno Olshausen, and Trevor Darrell. Tent: Fully test-time adaptation by entropy minimization. In *International Conference on Learning Representations*, 2021. 3

[56] 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. 5

[57] Mitchell Wortsman, Gabriel Ilharco, Jong Wook Kim, Mike Li, Simon Kornblith, Rebecca Roelofs, Raphael Gontijo Lopes, Hannaneh Hajishirzi, Ali Farhadi, Hongseok Namkoong, et al. Robust fine-tuning of zero-shot models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7959–7971, 2022. 2

[58] 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

[59] Hantao Yao, Rui Zhang, and Changsheng Xu. Visual-language prompt tuning with knowledge-guided context optimization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6757–6767, 2023. 1, 2

[60] Lewei Yao, Runhui Huang, Lu Hou, Guansong Lu, Minzhe Niu, Hang Xu, Xiaodan Liang, Zhenguo Li, Xin Jiang, and Chunjing Xu. Filip: Fine-grained interactive language-image pre-training. *arXiv preprint arXiv:2111.07783*, 2021. 1

[61] Alan L Yuille and Anand Rangarajan. The concave-convex procedure (cccp). *Advances in neural information processing systems*, 14, 2001. 4

[62] Xiaohua Zhai, Xiao Wang, Basil Mustafa, Andreas Steiner, Daniel Keysers, Alexander Kolesnikov, and Lucas Beyer. Lit: Zero-shot transfer with locked-image text tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18123–18133, 2022. 1

[63] Jingyi Zhang, Jiaxing Huang, Sheng Jin, and Shijian Lu. Vision-language models for vision tasks: A survey. *arXiv preprint arXiv:2304.00685*, 2023. 1, 2

[64] Marvin Zhang, Sergey Levine, and Chelsea Finn. Memo: Test time robustness via adaptation and augmentation. In *Advances in Neural Information Processing Systems*, pages 38629–38642. Curran Associates, Inc., 2022. 1, 2

[65] 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, 6, 8

[66] Renrui Zhang, Xiangfei Hu, Bohao Li, Siyuan Huang, Hanqiu Deng, Yu Qiao, Peng Gao, and Hongsheng Li. Prompt, generate, then cache: Cascade of foundation models makes strong few-shot learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15211–15222, 2023. 2

[67] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Conditional prompt learning for vision-language models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 1, 2

[68] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *International Journal of Computer Vision (IJCV)*, 2022. 1, 2, 5, 6, 8

[69] 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. 1, 2, 6, 8