

A Lost Opportunity for Vision-Language Models: A Comparative Study of Online Test-time Adaptation for Vision-Language Models

Mario Döbler * Robert A. Marsden * Tobias Raichle Bin Yang
University of Stuttgart

{mario.doebler, robert.marsden, bin.yang}@iss.uni-stuttgart.de

Abstract

In the realm of deep learning, maintaining model robustness against distribution shifts is critical. This paper investigates test-time adaptation strategies for vision-language models, with a specific focus on CLIP [37] and its variants. Through a systematic exploration of prompt-based techniques and existing test-time adaptation methods, the study aims to enhance the adaptability and robustness of vision-language models in diverse real-world scenarios. The investigation includes an analysis of prompt engineering strategies, such as hand-crafted prompts, prompt ensembles, and prompt learning techniques. We introduce a vision-text-space ensemble that significantly boosts the average performance compared to a text-space-only ensemble. Additionally, our comparative study delves into leveraging existing test-time adaptation methods originally designed for image classification tasks. Experimental evaluations conducted across various datasets and model architectures demonstrate the efficacy of different adaptation strategies. We further give insights into the importance of updating the vision encoder and whether it is beneficial to update the text encoder. Code is available at: <https://github.com/mariodoebler/test-time-adaptation>

1. Introduction

In the rapidly evolving field of deep learning, the robustness of models against distribution shifts remains a critical challenge. If the data distribution at test-time deviates from the training distribution, the performance can decrease significantly. This challenge is prevalent in most practical deep learning applications due to the difficulty of accurately replicating testing conditions during training. An intuitive answer to shifts in distribution is an extensively trained model across diverse datasets that can be adapted to a wide range of downstream tasks. Such models are

nowadays termed as foundation models. They are known to exhibit superior generalization abilities, setting them apart from conventional models. Current vision-language models, like CLIP [37], have shown strong zero-shot performance across a variety of computer vision benchmarks.

In this work, we study the task of online test-time adaptation (TTA) for vision-language (VL) models, with a specific focus on CLIP and its variants. We explore various strategies and methodologies aimed at enabling these models to adapt dynamically to distribution shifts encountered during inference. Our investigation encompasses both prompt-based approaches, which involve modifying the input prompts provided to the model, and existing TTA methods borrowed from the domain of image classification. By systematically evaluating these approaches across a range of datasets and scenarios, we aim to provide insights into the efficacy and practical applicability of different TTA strategies for vision-language models. Through our exploration, we seek to contribute to the development of more robust and adaptable vision-language models capable of performing reliably in diverse real-world settings.

We summarize our main contributions as follows:

- We discuss a broad range of possibilities to adapt vision-language foundation models at test-time - from various prompting strategies to applying existing test-time adaptation methods.
- We introduce a vision-text-space ensemble that is optimization-free and outperforms test-time prompt tuning.
- Our broad comparative study shows the potential of existing test-time adaptation methods for enhancing the robustness of vision-language models. Choosing a good method leads to significant improvements across a broad range of datasets and models.

2. Related Work

2.1. Foundation Models

“Foundation model” is a general notion of systems with broad zero-shot capabilities that can be adapted for specific

*Equal contribution.

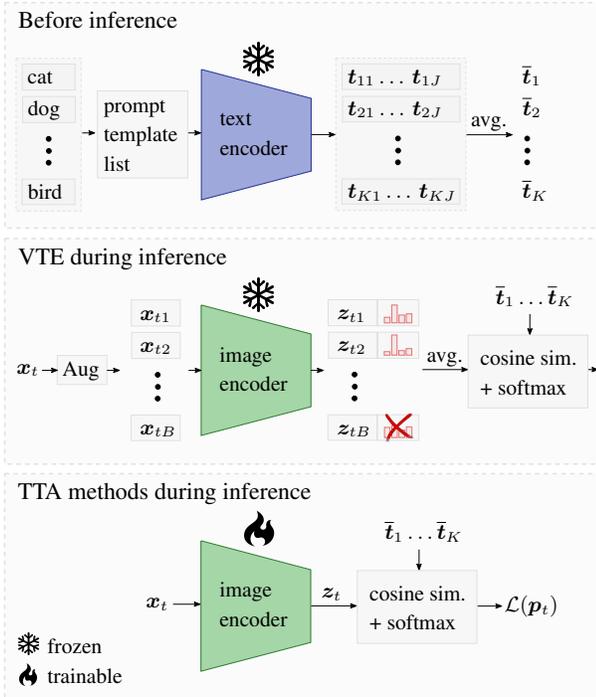


Figure 1. Overview of the proposed VTE approach and the application of existing TTA methods for VLMs. Before inference, an average text representation \bar{t}_k for each of the K classes is extracted by mapping a list of prompts into the text embedding space. During inference, VTE uses test-time augmentation and entropy based filtering. In the case of applying TTA methods, only the parameters of the vision encoder are updated.

purposes, e.g., via fine-tuning. Most notably, this encompasses large language models (LLMs) and multimodal models, such as large vision-language models (VLMs). LLMs are systems capable of understanding and generating language; popular examples include [4, 8, 45]. VLMs combine visual and textual information, enabling them to comprehend and generate content that encompasses both modalities. Several VLM architectures have been proposed: dual-encoder architectures [20, 37], encoder-decoder architectures [6, 49], unified transformer architectures [1, 25], and many more. In this work we investigate test-time adaptation for VLMs and mainly focus on CLIP [37], as it is still the most representative VLM. Additionally, we report results for EVA-CLIP [44] that proposed improved training techniques for CLIP at scale.

2.2. Online Test-time Adaptation

Online test-time adaptation adapts the model to an unknown domain shift directly during inference, leveraging all available test samples. One successful line of work recalculates the batch normalization (BN) statistics during test-time [41] to mitigate covariate shift caused by corruption. Although

updating only the BN statistics is computationally efficient, it has its limitations, especially regarding natural domain shifts. As a result, recent TTA methods additionally incorporate model weight updates through self-training. TENT [46], for example, showcased that minimizing the entropy with respect to the batch normalization parameters can successfully improve the performance for single-target adaptation. Building upon this idea, EATA [33] introduces a loss weighting and filtering scheme that accounts for the reliability and diversity of a sample. Furthermore, they use elastic weight consolidation [21] to mitigate catastrophic forgetting [30] on the initial training domain. However, accessing data from the initial training domain may not always be feasible in practical scenarios. To prevent a model from collapsing to a trivial solutions induced by confidence maximization, [26, 31] apply diversity regularizers. Other works, such as [5, 9] also employ contrastive learning to mitigate a domain shift during test-time.

While certain TTA methods focus solely on adapting to a single domain, real-world scenarios often involve encountering multiple domain shifts. Thus, [48] introduced continual test-time adaptation, where a model is adapted to a sequence of diverse domains. While self-training-based approaches such as [46] can be also utilized in the continual setting, they can be susceptible to error accumulation [29, 48]. To address this, [48] proposes weight and augmentation-averaged predictions alongside a stochastic restore mechanism to mitigate catastrophic forgetting. RMT [9] proposes a robust mean teacher to handle multiple domain shifts, while GTTA [28] uses mixup and style-transfer to artificially create intermediate domains.

Recent research has tackled even more challenging scenarios, such as dealing with temporally correlated data. LAME [3] focuses on adapting the model’s output using Laplacian adjusted maximum-likelihood estimation. On the other hand, NOTE [14], RoTTA [51], and DAB [11] introduce a buffer to simulate an i.i.d. stream. To handle large and noisy gradients that can promote model collapse, SAR [34] proposes a sharpness-aware and reliable entropy minimization method. Building upon SAR, DeYO [23] incorporates a confidence metric that measures the extent to which the probability of pseudo-label decreases after applying an image transformation that distorts the shape of the objects.

In the work of [29], recent TTA methods are evaluated on a broad range of possible TTA scenarios, termed Universal TTA. Their proposed method ROID [29] puts emphasis on using certainty and diversity weighting to prevent model collapse during adaptation and introduces weight ensembling at test-time to prevent the adapted model from overfitting on the current distribution shift. CMF [24] builds upon the framework of ROID and replaces weight ensembling by continual momentum filtering. It utilizes the Kalman filtering algorithm to find a source model that is robust against

catastrophic forgetting while maintaining high flexibility.

Due to their multimodality and zero-shot generalization capabilities, VLMs offer new possibilities for TTA. One approach that focuses on adapting the prompt space is test-time prompt tuning (TPT) [42]. It is loosely based on the supervised context optimization (CoOp) [52]. The basic idea of TPT is to optimize the prompt context via entropy minimization of an augmented batch based on a single sample. In this work we want to give new perspectives on how to deal with VLMs, namely CLIP, in the context of online test-time adaptation.

3. Prompts and Vision-space Ensembles

Vision-language foundation models such as CLIP [37] aim to learn a joint embedding space for the vision and language modality. This is achieved by aligning the representations of images and their associated textual descriptions through contrastive learning. To extract the embeddings, CLIP leverages a separate encoder for the vision and text modality, denoted here as f_{vision} and f_{text} , respectively. After successfully training the encoders on typically millions of image-text pairs, the learned joint embedding space allows to associate similar concepts across modalities, resulting in cross-modal understanding.

In the case of zero-shot classification with a hand-crafted prompt, the procedure involves the following steps. Let $\mathbf{x}_t \in \mathbb{R}^{H \times W \times C}$ be the current test image at time step t with height H , width W , and C channels, and $\mathbf{z}_t = f_{\text{vision}}(\mathbf{x}_t)$ denote its corresponding representation. In addition, let $\{\mathbf{t}_1, \mathbf{t}_2, \dots, \mathbf{t}_K\}$ be a textual representation for each of the K classes, obtained by embedding short phrases (templates) like *"a photo of a {classname}."* into the text embedding space. Now, to determine the class label for an image, its representation \mathbf{z}_t is first paired with each of the K text representations $(\mathbf{t}_k, \mathbf{z}_t)$. Then, the cosine similarity $s_k = \text{sim}(\mathbf{t}_k, \mathbf{z}_t)$ is computed for each pair. The final model prediction simply corresponds to the class with the highest similarity score or highest softmax probability. The latter can be computed with

$$p_{tk} = \frac{\exp(\text{sim}(\mathbf{t}_k, \mathbf{z}_t)/\tau)}{\sum_{j=1}^K \exp(\text{sim}(\mathbf{t}_j, \mathbf{z}_t)/\tau)} \quad (1)$$

where τ is a temperature. It's worth noting that the text embeddings $\{\mathbf{t}_1, \mathbf{t}_2, \dots, \mathbf{t}_K\}$ are typically precomputed once before inference, ensuring efficiency during prediction.

3.1. Prompt Engineering

While using a simple phrase like *a photo of a {classname}*, can already work exceptionally well, the performance of VL models heavily depends on the utilized prompt and its encoded representation [52]. Thus, writing better hand-crafted prompts can significantly improve the performance

of the model. For instance, in the context of ImageNet-R [17] and a ViT-B-16 model pretrained by OpenAI, employing the prompt template *"depiction of a {classname}."* reduces the error rate from 26.0% to 23.6%. Similarly, for datasets like EuroSAT [15] that contain low-resolution satellite photos, using a prompt such as *a blurry satellite photo of {classname}*, decreases the error rate from 58.5% to 46.3%. These examples underscore the importance of well-designed prompts to optimize model performance.

Instead of relying on a single prompt template, Radford et al. [37] also proposed to use a list of J different templates. An example can look like the following list [*"a photo of a {classname}."*, *"a sketch of a {classname}."*, *"a painting of a {classname}."*]. By averaging the text representations obtained from all templates for class k , i.e.,

$$\bar{\mathbf{t}}_k = \frac{1}{J} \sum_{j=1}^J \mathbf{t}_{kj}, \quad (2)$$

a text-based ensemble within the embedding space can be formed. In this case, the similarity scores are now computed with $s_k = \text{sim}(\bar{\mathbf{t}}_k, \mathbf{z}_t)$. While this has been found to not only consistently improve the results [37], it also avoids increasing the computational complexity and the memory requirements during inference. This efficiency is again due to the ability of precomputing $\bar{\mathbf{t}}_k$ prior to inference.

While the ensemble approach described earlier uses a predefined list of hand-crafted prompt templates, CuPL [36] introduces a novel strategy that harnesses the power of a large language model to generate a class-specific prompt list. Specifically, the LLM is asked to write descriptive sentences that encapsulate the discriminative features of the various classes. In the case of the category goldfish, the prompt list might look like [*"Most goldfish have a shiny gold or orange color."*, *"A goldfish in a bowl."*, ...]. These descriptive prompts help to improve the performance of the VL model without requiring any expert knowledge.

3.2. Learning Prompts

Zhou et al. [52] introduced context optimization to offline fine-tune CLIP-like vision-language models with a few labeled training examples $\{\mathbf{x}_i, \mathbf{y}_i\}_{i=1}^N$, where $\mathbf{y}_i \in \mathbb{R}^K$ is the one-hot encoded category of image \mathbf{x}_i . Unlike before, where the context of the prompt (such as *"a photo of a"*) was either fixed or manually tuned, it is now learnable. This involves representing the context with a few learnable token embeddings and then minimizing, for example, a cross-entropy (CE) loss according to

$$\mathcal{L}_{\text{CE}} = -\frac{1}{N} \sum_{i=1}^N \sum_{k=1}^K y_{ik} \log(p_{ik}). \quad (3)$$

Building upon the idea of learning the context prompts, TPT [42] exploits context optimization during test-time. The

Table 1. Online classification error rate (%) for CLIP with a ViT-B-16 and ViT-L-14 backbone pretrained by OpenAI. The models comprise 149.62 million parameters with 41.09 billion FLOPS and 427.62 million parameters with 175.33 billion FLOPS, respectively.

| | Method | Prompt | ImageNet | ImageNet-C | ImageNet-A | ImageNet-V2 | ImageNet-R | ImageNet-S | ImageNet-D100 | Flower102 | DTD | Pets | Cats | UCF101 | Caltech101 | Food101 | SUN397 | Aircraft | EuroSAT | Avg. | |
|----------|--------|--------------|-------------|-------------|-------------|-------------|-------------|-------------|---------------|-------------|-------------|-------------|-------------|-------------|------------|-------------|-------------|-------------|-------------|-------------|---|
| ViT-B-16 | Source | Single | 33.3 | 75.5 | 52.3 | 39.2 | 26.0 | 53.9 | 29.5 | 32.6 | 55.4 | 11.8 | 34.8 | 34.9 | 7.1 | 16.2 | 37.4 | 76.2 | 58.5 | 39.7 | |
| | | Ensemble | 31.7 | 73.8 | 49.9 | 38.1 | 22.5 | 51.7 | 27.5 | 34.2 | 54.6 | 11.8 | 33.6 | 32.6 | 7.0 | 15.5 | 34.6 | 76.5 | 51.8 | 38.1 | |
| | | CuPL | 30.4 | 73.3 | 49.3 | 36.7 | 22.9 | 51.0 | 28.0 | - | - | - | - | - | - | - | - | - | - | - | - |
| | | All Prompts | 30.3 | 73.0 | 49.0 | 36.9 | 22.0 | 50.6 | 27.2 | - | - | - | - | - | - | - | - | - | - | - | - |
| | TPT | a photo of a | 31.0 | 75.1 | 45.7 | 36.5 | 23.0 | 52.2 | 26.8 | 30.9 | 52.7 | 12.8 | 34.0 | 32.5 | 6.3 | 15.2 | 34.6 | 76.7 | 57.2 | 37.8 | |
| | VTE | Ensemble | 29.6 | 74.4 | 37.3 | 34.9 | 19.6 | 49.8 | 24.6 | 34.5 | 52.7 | 13.0 | 31.0 | 33.0 | 6.7 | 16.6 | 33.5 | 75.9 | 52.4 | 36.4 | |
| | VTE | All Prompts | 28.3 | 73.6 | 36.7 | 34.1 | 19.7 | 49.0 | 24.5 | - | - | - | - | - | - | - | - | - | - | - | |
| ViT-L-14 | Source | Single | 26.5 | 60.5 | 31.3 | 32.1 | 14.6 | 42.1 | 24.2 | 24.1 | 47.4 | 6.8 | 23.2 | 27.3 | 5.2 | 11.4 | 32.5 | 69.7 | 44.7 | 30.8 | |
| | | Ensemble | 24.5 | 58.6 | 29.3 | 30.1 | 12.2 | 40.4 | 22.4 | 24.4 | 43.3 | 7.0 | 22.1 | 25.0 | 5.5 | 10.8 | 30.9 | 68.1 | 39.3 | 29.1 | |
| | | CuPL | 23.4 | 57.7 | 28.2 | 29.2 | 12.3 | 40.0 | 22.5 | - | - | - | - | - | - | - | - | - | - | - | - |
| | | All Prompts | 23.5 | 57.5 | 28.2 | 29.0 | 11.8 | 39.7 | 22.1 | - | - | - | - | - | - | - | - | - | - | - | - |
| | TPT | a photo of a | 24.5 | 59.0 | 25.2 | 29.9 | 12.1 | 40.2 | 22.0 | 23.5 | 46.1 | 6.4 | 22.3 | 25.5 | 4.4 | 10.9 | 29.8 | 68.7 | 48.0 | 29.3 | |
| | VTE | Ensemble | 23.0 | 59.5 | 20.4 | 28.4 | 10.3 | 38.9 | 20.5 | 26.2 | 41.9 | 7.1 | 21.6 | 24.6 | 4.2 | 11.7 | 29.3 | 66.1 | 46.4 | 28.2 | |
| | VTE | All Prompts | 22.3 | 58.8 | 19.7 | 27.5 | 9.8 | 38.3 | 20.2 | - | - | - | - | - | - | - | - | - | - | - | |

procedure involves using test-time augmentation to create a batch $\{x_{ti}\}_{i=1}^B$ of $B = 64$ samples from a single test image. Then, the most confident $\rho = 10\%$ of the samples in terms of entropy $e_{ti} = \sum_{k=1}^K p_{tik} \log(p_{tik})$ are selected to minimize an entropy loss with respect to the trainable context parameters. This results in the following expression

$$\mathcal{L}_{\text{TPT}} = -\frac{1}{\rho B} \sum_{i=1}^B \sum_{k=1}^K [e_{ti} \leq \beta] p_{tik} \log(p_{tik}), \quad (4)$$

where $[\cdot]$ is the Iverson bracket and β is a threshold. After the context is updated one (or several) times, regular zero-shot classification can be performed. While [42] demonstrate the efficacy of this approach, the computational overhead is huge, since for each test image, there are 64 forward passes through the image encoder and at least 2 forward passes through the text encoder - one to learn a better context and another to acquire the new text representations.

3.3. Vision-Text-Space Ensemble

The effectiveness of methods like TPT depends on the creation of suitable training examples via test-time augmentation, which can be identified with confidence-based filtering, for example. Feng et al. [13] take this approach one step further by additionally exploiting Stable Diffusion [39] to generate new images that resemble the current test image. Since leveraging a diverse set of augmented test samples might also be helpful for the previous hand-crafted or LLM-based prompt ensemble approaches, we introduce a **Vision-Text-Space Ensemble (VTE)**, as depicted in Fig. 1, which creates an ensemble in both spaces. Following TPT, VTE utilizes the same test-time augmentation strategy and

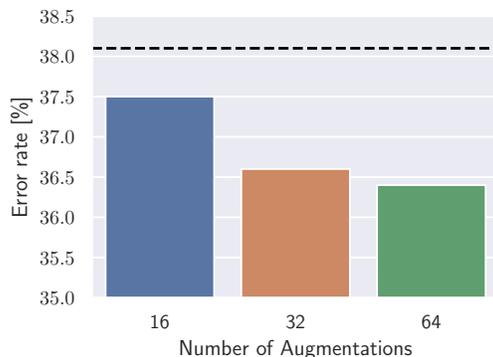


Figure 2. Average error rate of VTE with a ViT-B-16 backbone across all 17 datasets when using different numbers of augmentations during test-time. The dashed line indicates the performance of zero-shot CLIP with Ensemble prompts.

entropy-based confidence filtering to extract reliable samples from the artificially generated batch. The representations of the identified samples are then averaged via the equation

$$\bar{z}_t = \frac{1}{\rho B} \sum_{i=1}^B [e_{ti} \leq \beta] z_{ti}, \quad (5)$$

where \bar{z}_t is subsequently utilized to compute the similarity scores according to $s_k = \text{sim}(\hat{t}_k, \bar{z}_t)$. Note that this procedure is again optimization-free and, unlike TPT, does not require any forward passes through the text encoder during test-time.

3.4. Experiments

3.4.1 Datasets, Models, and Metric

We follow the continual test-time adaptation setting in [29] and evaluate the models’ robustness on ImageNet (validation set) and its variants. ImageNet-C [16] includes 15 types of corruptions with 5 severity levels applied to the validation images of ImageNet (IN). For the natural domain shifts, we consider ImageNet-R [17], ImageNet-Sketch [47], as well as ImageNet-D109, a variation of ImageNet-D [40] introduced in [29]. While ImageNet-R contains 30,000 examples depicting different renditions of 200 IN classes, ImageNet-Sketch contains 50 sketches for each of the 1,000 IN classes. Additionally, we report results for ImageNet-V2 [38] and ImageNet-A [18]. ImageNet-V2 is an independent test set containing 10,000 natural images covering all 1,000 IN classes. ImageNet-A comprises 7,500 adversarial examples for a subset of 200 IN categories.

To evaluate categories outside the ImageNet context, we follow [42] and report results for ten datasets, covering fine-grained classifications including species of plants or animals (Flowers102 [32], OxfordPets [35]), scenes (SUN397 [50]), textures (DTD [7]), food (Food101 [2]), transportation (StanfordCars [22], Aircraft [27]), human actions (UCF101 [43]), satellite images (EuroSAT [15]), and general objects (Caltech101 [12]).

We utilize six different architectures for our evaluation, encompassing CLIP with a ResNet-50 (RN50) backbone, as well as ViT [10] models of various sizes: ViT-B-16, ViT-L-14, and ViT-H-14, all pretrained by OpenAI. Furthermore, we consider EVA-02-B-16 and EVA02-L-14 from [44]. Our evaluation metric is based on the error rate.

3.4.2 Results

The results of the diverse prompt-based methods are illustrated in Table 1. Here, Source denotes employing zero-shot classification with different prompt strategies: utilizing the single prompt “a photo of a {classname}.”, an ensemble of hand-crafted prompt templates following [37], the CuPL [36] prompts generated by an LLM, and a combination of both ensemble and CuPL prompts referred to as “All Prompts”.

As shown in Table 1, all methods substantially improve on the single prompt baseline. While the LLM generated prompts of CuPL outperform the hand-crafted ensemble on five out of seven ImageNet variations for both architectures, even better results can be achieved by leveraging all prompts. This even outperforms the optimization based approach of TPT on four out of seven IN variations, while requiring only a fraction of its computational effort, i.e., one image forward versus 64 image forwards, 2 text forwards, and one backward. However, the best results are

achieved by our VTE approach, which significantly outperforms all other baselines. Although this comes at the cost of an increased computational complexity compared to the hand-crafted approaches, VTE is still faster than TPT during inference, since it does not require any text forwards or backwards. We also find, that the performance of VTE can be further improved by employing a better prompt list. For ViT-B-16, for example, using All Prompts decreases the average error from 38.6% to 38.0%.

In Fig. 2, we study the performance of VTE for different numbers of augmentations during test-time, employing the ensemble prompt and a ViT-B-16. Only applying 32 augmentations results in a mere 0.2% increase in error rate compared to using 64 augmentations. Moreover, even with just 16 augmentations, there is still a notable improvement of 0.6% in terms of average error rate compared to zero-shot classification with an ensemble prompt.

4. Updating Model Parameters with Test-time Adaptation

While the focus of the previous section has been on leveraging prompts and vision-space ensembles, in this section we want to put emphasis on a surprisingly underexplored topic, namely leveraging existing TTA methods for adapting vision-language models. The idea is straightforward, the text encoder of a CLIP model is frozen, allowing to precompute the text embeddings $\{t_1, t_2, \dots, t_K\}$, ensuring efficiency during prediction. Given an image embedding z_t , the cosine similarity can be computed $s_t = \text{sim}(t_k, z_t)$. Treating the cosine similarities as the network’s logits, the output probabilities can be received through Eq. (1). In this way we can treat any CLIP model as a common image classifier, enabling the application of any existing TTA method for image classification. Note that it is also possible to update the text encoder’s parameters, but for now, we limit our analysis to only updating the parameters of the image encoder. In the following experiments a batch size of 64 is employed.

4.1. Test-time Normalization for CLIP

First, we investigate the performance of BN-1 which recalculates the batch normalization (BN) statistics using the current test batch. While Schneider et al. [41] showed that recalculating the batch normalization statistics during test-time can significantly reduce the error rate for models pretrained on ImageNet, we investigate whether this is also the case for a CLIP model that was trained on millions of data samples covering a much broader data distribution. In Figure 3 the zero-shot performance (source) and BN-1 performance is illustrated for CLIP with a RN50 and RN101 backbone using a single prompt. It can be clearly seen that unlike for models pretrained on ImageNet, the

Table 2. Online classification error rate (%) for CLIP with a ViT-B-16 and ViT-L-14 backbone pretrained by OpenAI in a continual TTA setting. The models comprise 149.62 million parameters with 41.09 billion FLOPS and 427.62 million parameters with 175.33 billion FLOPS, respectively.

| Method | Prompt | ImageNet | ImageNet-C | ImageNet-A | ImageNet-V2 | ImageNet-R | ImageNet-S | ImageNet-D109 | Flower102 | DTD | Pets | Cats | UCF101 | Caltech101 | Food101 | SUN397 | Aircraft | EuroSAT | Avg. | |
|----------|--------|----------|-------------|-------------|-------------|-------------|-------------|---------------|-------------|-------------|-------------|-------------|-------------|-------------|------------|-------------|-------------|-------------|-------------|-------------|
| ViT-B-16 | Source | Ensemble | 31.7 | 73.8 | 49.9 | 38.1 | 22.5 | 51.7 | 27.5 | 34.2 | 54.6 | 11.8 | 33.6 | 32.6 | 7.0 | 15.5 | 34.6 | 76.5 | 51.8 | 38.1 |
| | TENT | Ensemble | 31.6 | 75.4 | 49.6 | 38.1 | 21.6 | 51.6 | 27.2 | 34.2 | 54.3 | 11.6 | 33.6 | 32.3 | 6.9 | 15.8 | 34.4 | 76.5 | 43.3 | 37.5 |
| | ETA | Ensemble | 32.1 | 69.1 | 49.3 | 38.3 | 21.7 | 50.9 | 26.9 | 34.0 | 54.4 | 11.3 | 33.5 | 32.1 | 7.0 | 16.0 | 33.9 | 76.2 | 50.7 | 37.5 |
| | SAR | Ensemble | 32.0 | 70.0 | 49.2 | 38.3 | 21.8 | 51.5 | 27.0 | 33.9 | 54.4 | 11.5 | 34.1 | 32.5 | 7.0 | 15.7 | 34.8 | 76.4 | 44.7 | 37.3 |
| | DeYO | Ensemble | 32.4 | 71.3 | 48.9 | 38.4 | 21.7 | 51.5 | 27.0 | 34.2 | 54.3 | 11.3 | 34.1 | 32.1 | 7.2 | 16.0 | 35.0 | 76.2 | 50.1 | 37.7 |
| | CMF | Ensemble | 31.9 | 66.1 | 49.6 | 38.3 | 20.9 | 50.4 | 25.7 | 33.8 | 54.4 | 11.5 | 33.8 | 32.0 | 6.9 | 15.7 | 33.6 | 76.3 | 36.5 | 36.3 |
| | ROID | Ensemble | 31.7 | 65.7 | 49.3 | 38.2 | 21.1 | 50.9 | 26.3 | 33.6 | 54.2 | 11.4 | 33.4 | 31.9 | 6.9 | 15.8 | 33.4 | 76.1 | 36.3 | 36.2 |
| ViT-L-14 | Source | Ensemble | 24.5 | 58.6 | 29.3 | 30.1 | 12.2 | 40.4 | 22.4 | 24.4 | 43.3 | 7.0 | 22.1 | 25.0 | 5.5 | 10.8 | 30.9 | 68.1 | 39.3 | 29.1 |
| | TENT | Ensemble | 24.6 | 56.1 | 29.3 | 30.3 | 12.1 | 40.1 | 22.1 | 24.4 | 43.2 | 6.9 | 22.2 | 24.9 | 5.6 | 10.8 | 30.8 | 68.0 | 36.5 | 28.7 |
| | ETA | Ensemble | 24.6 | 53.8 | 28.9 | 30.4 | 11.9 | 39.6 | 21.7 | 24.3 | 43.1 | 7.0 | 21.7 | 24.7 | 5.6 | 10.8 | 30.5 | 67.8 | 39.3 | 28.6 |
| | SAR | Ensemble | 24.6 | 54.9 | 28.9 | 30.3 | 11.8 | 39.8 | 21.9 | 24.3 | 43.2 | 6.9 | 22.1 | 24.9 | 5.6 | 10.7 | 30.4 | 68.0 | 36.2 | 28.5 |
| | DeYO | Ensemble | 24.6 | 54.3 | 28.6 | 30.5 | 11.6 | 39.6 | 21.3 | 24.3 | 43.1 | 6.8 | 21.8 | 24.9 | 5.7 | 10.9 | 30.7 | 67.9 | 38.1 | 28.5 |
| | CMF | Ensemble | 24.2 | 50.6 | 28.2 | 30.0 | 11.1 | 38.6 | 20.4 | 24.1 | 43.0 | 6.6 | 21.9 | 24.7 | 5.6 | 10.7 | 30.1 | 67.4 | 32.2 | 27.6 |
| | ROID | Ensemble | 24.3 | 51.4 | 28.4 | 30.0 | 11.6 | 39.3 | 21.5 | 23.9 | 43.3 | 6.6 | 21.8 | 24.7 | 5.6 | 10.7 | 30.1 | 67.6 | 32.3 | 27.8 |

average performance across the investigated datasets substantially decreases when applying BN-1. For RN50 the average error rate increases from 50.2% to 74.1% and for RN101 from 46.8% to 71.0%. This can be possibly attributed to much larger batch sizes and a much broader data distribution used during CLIP pretraining. A similar phenomenon is described in [29], where employing BN-1 for a regular ImageNet pretrained RN50 decreases the error rate on ImageNet-C from 82.0% to 68.6% in a continual TTA setting, but increases to 82.5% in a mixed-domains TTA setting, where all corruptions of ImageNet-C are randomly suffled within the test sequence. Since BN-1 is employed by most TTA methods during adaptation, we conclude that RN backbones are not feasible. Instead we focus our following analysis on vision transformer backbones that do not employ BN.

4.2. Are Existing TTA Methods Beneficial for Vision-language Foundation Models?

In this section, we take a deeper look into the performance of existing TTA methods applied to vision-language models, namely CLIP [37] and EVA-CLIP [44]. We evaluate influential and recent TTA methods: TENT [46], ETA [33], SAR [34], DeYO [23], CMF [24], and ROID [29] using the same adaptation setup and hyperparameters as proposed in the corresponding papers. We investigate ETA instead of EATA, since EATA requires access to samples from the source domain. In Table 2 we report the error rate for CLIP with ViT-B-16 and ViT-L-14 backbones in the continual TTA setting [48]. We decide on the continual TTA setting, since this also shows how TTA methods cope with multiple distribution shifts. Later, in Section 4.3, we take a look into

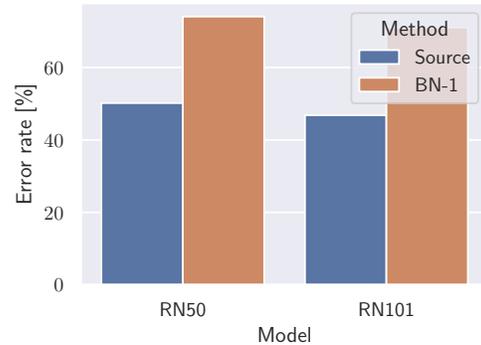


Figure 3. Average error rate for CLIP with a RN50 and RN101 backbone for both source and BN-1. As illustrated, the error rate drastically increases when the normalization statistics are recalculated during test-time.

a more challenging scenario, namely dealing with temporally correlated test sequences.

All TTA methods improve on average upon the zero-shot performance for ViT-B-16 and ViT-L-14. ROID and CMF show a comparable performance and show the best performance for most datasets. ROID decreases the error rate on average by 1.9% for ViT-B-16 and CMF by 1.5% for ViT-L-14. It is noteworthy that even for the already strong source performance, both CMF and ROID are on-par or better than the zero-shot model for each considered dataset. Both CMF and ROID even outperform VTE and TPT despite their much higher compute cost. Comparing ROID and TPT, ROID is absolutely 1.6% and 1.5% better using a ViT-B-16 and ViT-L-14, respectively.

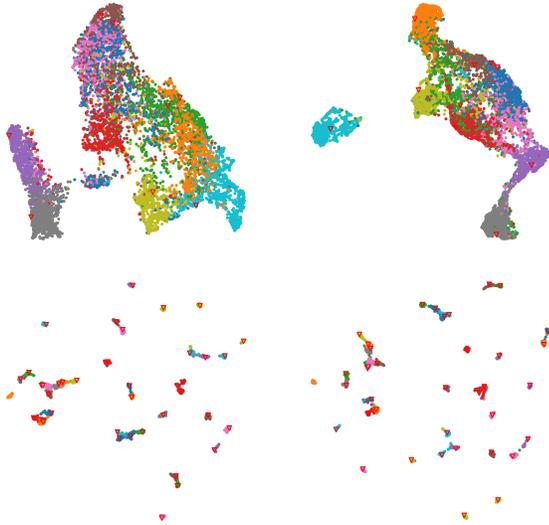


Figure 4. UMAP visualization for EuroSAT (top) and Pets (bottom) before (left) and after adaptation (right). To better align the text and image embeddings, we use a projection proposed in [19] before applying UMAP. The triangles illustrate the corresponding text ensemble embeddings.

The importance of updating the vision encoder for certain distribution shifts Taking a closer look at the individual performances, interestingly, ROID and CMF show a relatively high improvement on ImageNet-C and EuroSAT. For a ViT-B-16 they roughly improve absolutely 8% on ImageNet-C and 15% on EuroSAT compared to the source baseline. Getting insights into this phenomenon, we illustrate the feature space of the ViT-B-16 backbone before and after adaptation (adapted with ROID) for EuroSAT and compare it to the dataset Pets, where no significant improvement is seen. The UMAP visualization is shown in Figure 4. Comparing the low-dimensional space of EuroSAT and Pets before adaptation, it can be clearly seen that the zero-shot model has a much better class separation for Pets than for EuroSAT. For EuroSAT there is significant class overlap, hence, updating the vision encoder can result in a much more discriminative feature space. This undermines the importance and opportunity of adapting the vision encoder for data distributions where the zero-shot model has limited class separation. This also shows the limitations of prompt-based methods which simply work with a fixed image feature space.

Test-time adaptation remains beneficial for large models

A natural question that arises is whether the improvement for TTA methods diminishes for bigger models with a better initial performance. Therefore, in Figure 5, the error rate is illustrated for ViT-B-16 up to ViT-H-14. As one would expect, the performance gains through adaptation diminishes

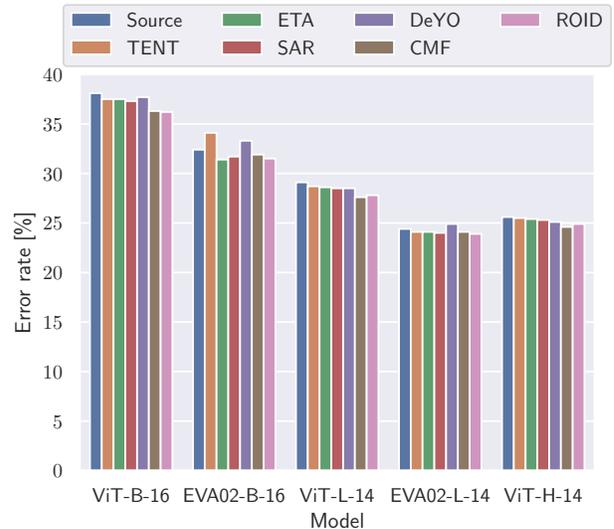


Figure 5. Comparison of different models sorted according to their number of model parameters from low (left) to high (right). The average error rate across all datasets is reported.

as the zero-shot performance improves. But, even for a ViT-H-14, all investigated TTA methods still improve upon the source performance with CMF taking the lead, reducing the error rate further by absolutely 1%. Interestingly, in contrast to the CLIP models by OpenAI, for the investigated EVA-CLIP models not all TTA methods, namely TENT and DeYO, can improve upon the zero-shot model.

4.3. Test-time Adaptation for Non-i.i.d. Data Streams

Since the previously investigated continual TTA setting might not apply to real-world online data streams, we additionally investigate a scenario with temporally correlated samples. In the correlated TTA setting the data of each domain is sorted by the class label rather than randomly shuffled, resulting in class-imbalanced batches. The results are reported in Table 3. For the more challenging correlated TTA setting, in contrast to the continual setting, not all TTA methods are capable to improve upon the source performance. Only CMF and ROID show a stable adaptation. Due to the employed prior correction¹ proposed in [29], CMF and ROID even report a better performance than in the continual setting.

4.4. Updating the Text Encoder

Up to now, only the parameters or a subset of the parameters of the vision encoder were updated. Additionally updating

¹In contrast to the original prior correction, we find that applying the prior correction in the output probability space instead of the logit space shows a more consistent performance for the investigated CLIP models.

Table 3. Online classification error rate (%) for CLIP with a ViT-B-16 and ViT-L-14 backbone pretrained by OpenAI in a correlated TTA setting. The models comprise 149.62 million parameters with 41.09 billion FLOPS and 427.62 million parameters with 175.33 billion FLOPS, respectively.

| | Method | Prompt | ImageNet | ImageNet-C | ImageNet-A | ImageNet-V2 | ImageNet-R | ImageNet-S | ImageNet-D109 | Flower102 | DTD | Pets | Cats | UCF101 | Caltech101 | Food101 | SUN397 | Aircraft | EuroSAT | Avg. |
|----------|--------|----------|-------------|-------------|-------------|-------------|-------------|-------------|---------------|-------------|-------------|------------|-------------|-------------|------------|------------|-------------|-------------|-------------|-------------|
| ViT-B-16 | Source | Ensemble | 31.7 | 73.8 | 49.9 | 38.1 | 22.5 | 51.7 | 27.5 | 34.2 | 54.6 | 11.8 | 33.6 | 32.6 | 7.0 | 15.5 | 34.6 | 76.5 | 51.8 | 38.1 |
| | TENT | Ensemble | 31.6 | 93.9 | 49.6 | 38.0 | 21.8 | 51.8 | 27.3 | 34.1 | 54.4 | 11.5 | 33.8 | 32.6 | 6.9 | 15.9 | 34.6 | 76.3 | 41.1 | 38.5 |
| | ETA | Ensemble | 33.7 | 88.4 | 51.4 | 38.1 | 23.0 | 53.6 | 30.9 | 34.1 | 54.3 | 11.5 | 33.8 | 32.7 | 7.0 | 20.5 | 34.5 | 76.5 | 51.7 | 39.7 |
| | SAR | Ensemble | 32.1 | 69.9 | 49.0 | 38.4 | 22.0 | 52.3 | 27.4 | 34.6 | 54.6 | 11.2 | 33.9 | 32.7 | 7.1 | 16.3 | 34.9 | 76.5 | 47.9 | 37.7 |
| | DeYO | Ensemble | 32.7 | 99.7 | 49.6 | 38.8 | 22.0 | 52.8 | 27.5 | 34.9 | 54.4 | 11.3 | 34.0 | 32.7 | 7.1 | 16.4 | 35.3 | 76.1 | 47.5 | 39.6 |
| | CMF | Ensemble | 25.5 | 59.3 | 41.0 | 36.3 | 11.4 | 47.1 | 20.3 | 32.8 | 48.8 | 7.7 | 28.3 | 28.1 | 5.5 | 8.1 | 25.7 | 74.0 | 40.7 | 31.8 |
| | ROID | Ensemble | 24.1 | 58.4 | 39.9 | 36.2 | 10.4 | 45.9 | 18.8 | 33.2 | 48.6 | 7.9 | 28.1 | 28.0 | 5.5 | 7.2 | 25.2 | 73.9 | 41.4 | 31.3 |
| ViT-L-14 | Source | Ensemble | 24.5 | 58.6 | 29.3 | 30.1 | 12.2 | 40.4 | 22.4 | 24.4 | 43.3 | 7.0 | 22.1 | 25.0 | 5.5 | 10.8 | 30.9 | 68.1 | 39.3 | 29.1 |
| | TENT | Ensemble | 24.5 | 53.5 | 29.2 | 30.2 | 12.1 | 40.1 | 22.1 | 24.4 | 43.4 | 7.0 | 22.1 | 24.9 | 5.6 | 10.8 | 30.9 | 68.1 | 36.1 | 28.5 |
| | ETA | Ensemble | 24.6 | 77.7 | 29.0 | 30.3 | 12.1 | 39.8 | 40.5 | 24.4 | 43.4 | 6.9 | 21.9 | 24.9 | 5.5 | 10.9 | 30.6 | 68.0 | 39.3 | 31.2 |
| | SAR | Ensemble | 27.2 | 60.7 | 29.1 | 30.3 | 12.2 | 44.1 | 22.4 | 24.3 | 43.6 | 6.9 | 23.0 | 25.0 | 5.7 | 11.1 | 31.6 | 67.9 | 42.7 | 29.9 |
| | DeYO | Ensemble | 24.6 | 55.9 | 28.8 | 30.6 | 11.8 | 39.8 | 21.4 | 24.2 | 43.3 | 6.9 | 22.0 | 25.0 | 5.7 | 11.1 | 31.0 | 67.9 | 40.3 | 28.8 |
| | CMF | Ensemble | 18.7 | 41.8 | 19.6 | 28.3 | 5.2 | 34.3 | 15.3 | 22.8 | 36.6 | 3.8 | 18.4 | 21.7 | 4.5 | 5.5 | 23.3 | 64.7 | 34.6 | 23.5 |
| | ROID | Ensemble | 17.6 | 41.7 | 19.7 | 28.1 | 4.9 | 33.3 | 15.1 | 22.9 | 36.8 | 3.8 | 18.4 | 21.8 | 4.5 | 5.1 | 22.8 | 64.4 | 39.1 | 23.5 |

the text encoder comes with a non-neglectable overhead. In this case, all text prompts have to be forwarded through the text encoder each update step. E.g., when using the common text prompt ensemble for ImageNet, this would require forwarding 80,000 text prompts each step and can quickly lead to an explosion in memory or compute requirement. Therefore, we restrict our ablation to using a single prompt for a ViT-B-16 backbone. Considering TENT, additionally updating the text encoder, decreases the performance on average by 1.8%. ROID improves on average by 0.2%, but compared to the ROID variant that employs the text ensemble, updating the text encoder with a single prompt is still 0.9% behind. Given these outcomes, we can conclude that updating the text encoder in addition to updating the vision encoder is not beneficial.

4.5. Limitations

A limitation of applying existing TTA methods to vision-language models is that they often require batches of data to perform updates. As discussed in [29] this is only partially true. Using networks that do not employ BN layers, such as VisionTransformer [10] allow to recover the batch TTA setting by simply accumulating the gradients of the last b test samples before updating the model. This comes with no computational overhead and even significantly reduces the memory requirement.

5. Conclusion

In this work, we explored the task of adapting vision-language models at test-time to accommodate distribution shifts. Our investigation led us through a comprehensive

analysis of both prompt-based approaches and existing test-time adaptation (TTA) methods applied to vision-language models, focusing particularly on CLIP and its variants. Our introduced vision-text-space ensemble shows to be the better option when compared to TPT. Our exploration of existing TTA methods revealed their potential for enhancing the robustness of vision-language models. Methods like ROID and CMF showcased impressive performance improvements across various datasets and model architectures.

References

- [1] Hangbo Bao, Li Dong, Songhao Piao, and Furu Wei. Beit: Bert pre-training of image transformers. *arXiv preprint arXiv:2106.08254*, 2021. 2
- [2] Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Food-101 – mining discriminative components with random forests. In *European Conference on Computer Vision*, 2014. 5
- [3] Malik Boudiaf, Romain Mueller, Ismail Ben Ayed, and Luca Bertinetto. Parameter-free online test-time adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8344–8353, 2022. 2
- [4] 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
- [5] Dian Chen, Dequan Wang, Trevor Darrell, and Sayna Ebrahimi. Contrastive test-time adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 295–305, 2022. 2
- [6] Xi Chen, Xiao Wang, Soravit Changpinyo, AJ Piergiovanni, Piotr Padlewski, Daniel Salz, Sebastian Goodman, Adam

- Grycner, Basil Mustafa, Lucas Beyer, et al. Pali: A jointly-scaled multilingual language-image model. *arXiv preprint arXiv:2209.06794*, 2022. 2
- [7] M. Cimpoi, S. Maji, I. Kokkinos, S. Mohamed, , and A. Vedaldi. Describing textures in the wild. In *Proceedings of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2014. 5
- [8] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. 2
- [9] Mario Döbler, Robert A Marsden, and Bin Yang. Robust mean teacher for continual and gradual test-time adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7704–7714, 2023. 2
- [10] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 5, 8
- [11] Mario Döbler, Florian Marencke, Robert A. Marsden, and Bin Yang. Diversity-aware buffer for coping with temporally correlated data streams in online test-time adaptation. In *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7665–7669, 2024. 2
- [12] 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 *CVPR Workshops*, 2004. 5
- [13] 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. 4
- [14] Taesik Gong, Jongheon Jeong, Taewon Kim, Yewon Kim, Jinwoo Shin, and Sung-Ju Lee. Note: Robust continual test-time adaptation against temporal correlation. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022. 2
- [15] 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 J. Sel. Top. Appl. Earth Obs. Remote. Sens.*, 2019. 3, 5
- [16] Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. *arXiv preprint arXiv:1903.12261*, 2019. 5
- [17] 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. 3, 5
- [18] 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
- [19] Xuefeng Hu, Ke Zhang, Lu Xia, Albert Chen, Jiajia Luo, Yuyin Sun, Ken Wang, Nan Qiao, Xiao Zeng, Min Sun, et al. Reclip: Refine contrastive language image pre-training with source free domain adaptation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2994–3003, 2024. 7
- [20] 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. 2
- [21] James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, et al. Overcoming catastrophic forgetting in neural networks. *Proceedings of the national academy of sciences*, 114(13):3521–3526, 2017. 2
- [22] Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d object representations for fine-grained categorization. In *ICCV Workshops*, 2013. 5
- [23] Jonghyun Lee, Dahuin Jung, Saehyung Lee, Junsung Park, Juhyeon Shin, Uiwon Hwang, and Sungroh Yoon. Entropy is not enough for test-time adaptation: From the perspective of disentangled factors. In *The Twelfth International Conference on Learning Representations*, 2024. 2, 6
- [24] Jae-Hong Lee and Joon-Hyuk Chang. Continual momentum filtering on parameter space for online test-time adaptation. In *The Twelfth International Conference on Learning Representations*, 2024. 2, 6
- [25] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, pages 12888–12900. PMLR, 2022. 2
- [26] 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. 2
- [27] S. Maji, J. Kannala, E. Rahtu, M. Blaschko, and A. Vedaldi. Fine-grained visual classification of aircraft. Technical report, 2013. 5
- [28] Robert A Marsden, Mario Döbler, and Bin Yang. Introducing intermediate domains for effective self-training during test-time. *arXiv preprint arXiv:2208.07736*, 2022. 2
- [29] Robert A Marsden, Mario Döbler, and Bin Yang. Universal test-time adaptation through weight ensembling, diversity weighting, and prior correction. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2555–2565, 2024. 2, 5, 6, 7, 8
- [30] Michael McCloskey and Neal J Cohen. Catastrophic interference in connectionist networks: The sequential learning problem. In *Psychology of learning and motivation*, pages 109–165. Elsevier, 1989. 2
- [31] Chaithanya Kumar Mummadi, Robin Huttmacher, Kilian Rambach, Evgeny Levinkov, Thomas Brox, and Jan Hendrik Metzen. Test-time adaptation to distribution shift by confi-

- dence maximization and input transformation. *arXiv preprint arXiv:2106.14999*, 2021. [2](#)
- [32] Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. In *Indian Conference on Computer Vision, Graphics and Image Processing*, 2008. [5](#)
- [33] Shuaicheng Niu, Jiayang Wu, Yifan Zhang, Yaofo Chen, Shijian Zheng, Peilin Zhao, and Mingkui Tan. Efficient test-time model adaptation without forgetting. In *The International Conference on Machine Learning*, 2022. [2](#), [6](#)
- [34] Shuaicheng Niu, Jiayang Wu, Yifan Zhang, Zhiqian Wen, Yaofo Chen, Peilin Zhao, and Mingkui Tan. Towards stable test-time adaptation in dynamic wild world. In *The Eleventh International Conference on Learning Representations*, 2023. [2](#), [6](#)
- [35] Omkar M. Parkhi, Andrea Vedaldi, Andrew Zisserman, and C. V. Jawahar. Cats and dogs. In *CVPR*, 2012. [5](#)
- [36] 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. [3](#), [5](#)
- [37] 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](#), [2](#), [3](#), [5](#), [6](#)
- [38] Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishal Shankar. Do imagenet classifiers generalize to imagenet? In *International conference on machine learning*, pages 5389–5400. PMLR, 2019. [5](#)
- [39] 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*, pages 10684–10695, 2022. [4](#)
- [40] Evgenia Rusak, Steffen Schneider, Peter Vincent Gehler, Oliver Bringmann, Wieland Brendel, and Matthias Bethge. Imagenet-d: A new challenging robustness dataset inspired by domain adaptation. In *ICML 2022 Shift Happens Workshop*, 2022. [5](#)
- [41] Steffen Schneider, Evgenia Rusak, Luisa Eck, Oliver Bringmann, Wieland Brendel, and Matthias Bethge. Improving robustness against common corruptions by covariate shift adaptation. *Advances in Neural Information Processing Systems*, 33:11539–11551, 2020. [2](#), [5](#)
- [42] 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. [3](#), [4](#), [5](#)
- [43] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. UCF101: A dataset of 101 human actions classes from videos in the wild. *CoRR*, abs/1212.0402, 2012. [5](#)
- [44] Quan Sun, Yuxin Fang, Ledell Wu, Xinlong Wang, and Yue Cao. Eva-clip: Improved training techniques for clip at scale. *arXiv preprint arXiv:2303.15389*, 2023. [2](#), [5](#), [6](#)
- [45] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023. [2](#)
- [46] 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. [2](#), [6](#)
- [47] Haohan Wang, Songwei Ge, Zachary Lipton, and Eric P Xing. Learning robust global representations by penalizing local predictive power. In *Advances in Neural Information Processing Systems*, pages 10506–10518, 2019. [5](#)
- [48] Qin Wang, Olga Fink, Luc Van Gool, and Dengxin Dai. Continual test-time domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7201–7211, 2022. [2](#), [6](#)
- [49] Zirui Wang, Jiahui Yu, Adams Wei Yu, Zihang Dai, Yulia Tsvetkov, and Yuan Cao. SimVLM: Simple visual language model pretraining with weak supervision. *arXiv preprint arXiv:2108.10904*, 2021. [2](#)
- [50] J. Xiao, J. Hays, K. A. Ehinger, A. Oliva, and A. Torralba. Sun database: Large-scale scene recognition from abbey to zoo. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2010. [5](#)
- [51] Longhui Yuan, Binhui Xie, and Shuang Li. Robust test-time adaptation in dynamic scenarios. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15922–15932, 2023. [2](#)
- [52] 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. [3](#)