CLIPArTT: Adaptation of CLIP to New Domains at Test Time

Gustavo A. Vargas Hakim* David Osowiechi*
Mehrdad Noori Milad Cheraghalikhani Ali Bahri Moslem Yazdanpanah
Ismail Ben Ayed Christian Desrosiers

ÉTS Montreal, Canada
International Laboratory on Learning Systems (ILLS)

Abstract

Pre-trained vision-language models (VLMs), exemplified by CLIP, demonstrate remarkable adaptability across zero-shot classification tasks without additional training. However, their performance diminishes in the presence of domain shifts. In this study, we introduce CLIP Adaptation duRing Test-Time (CLIPArTT), a fully test-time adaptation (TTA) approach for CLIP, which involves automatic text prompts construction during inference for their use as text supervision. Our method employs a unique, minimally invasive text prompt tuning process, wherein multiple predicted classes are aggregated into a single new text prompt, used as pseudo label *to re-classify inputs in a transductive manner. Additionally, we pioneer the standardization of TTA benchmarks (e.g., TENT) in the realm of VLMs. Our findings demonstrate that, without requiring additional transformations nor new trainable modules, CLIPArTT enhances performance dynamically across non-corrupted datasets such as CIFAR-100, corrupted datasets like CIFAR-100-C and ImageNet-C, alongside synthetic datasets such as VisDA-C. This research underscores the potential for improving VLMs' adaptability through novel test-time strategies, offering insights for robust performance across varied datasets and environments. The code can be found at:* https://github.com/dosowiechi/CLIPArTT.git

1. Introduction

Combining vision and language modalities for learning, namely a Vision Language model (VLM), has demonstrated an outstanding performance in different vision tasks [11, 13, 23]. Remarkably, these models are surprisingly effective at zero-shot generalization, where a new task can be outside the original scope of the training set, without any additional supervision to fine tune the model. Models such as CLIP [23] have then been employed in fields as diverse as video recognition [18], audio [8], and medical imaging [19].

As in other more traditional deep architectures (e.g., CNNs), CLIP is prone to performance degradation on domains to which it was not originally exposed. Recent research trends suggest that domain adaptation mechanisms can play an important role in deploying CLIP [14, 25]. The challenge, however, is adapting the model to new domains in an efficient manner, so that its attractive zero-shot capabilities is maintained without the need of retraining.

In this paper, CLIP is contextualized in the setting of Test-Time Adaptation, a challenging yet practical scenario of domain adaptation. In this scenario, a model needs to adapt to new data *on-the-fly* to cope with unknown distribution shifts, and without using any class supervision. Although an experimental boilerplate was standardized in recent years, CLIP has not been integrated to it yet. Additionally, we introduce a powerful adaptation technique that achieves *state-of-the-art* performance without a significant computational overhead. Comprehensive studies are performed on multiple datasets containing different types of domain shifts on several levels of severity, resulting in a total of 59 evaluation scenarios. Our main contributions can be summarized as follows:

- We propose CLIPArTT, a Test-Time Adaptation method for CLIP that adapts the VLM by updating normalization-layer parameters. This is achieved by combining multiple classes into a single new text prompt, which is then used as a pseudo-label.
- We introduce a new benchmark for Test-Time Adaptation on Vision-Language Models by implementing representative baselines, such as TENT.
- Through comprehensive experiments, we subject our CLIPArTT methodology to diverse and challenging Test-Time Adaptation scenarios, each characterized by

*Equal contribution. Correspondence to gustavo-adolfo.vargas-hakim.1.1@ens.etsmtl.ca, david.osowiechi.1@ens.etsmtl.ca

distinct types of domain shifts. The outcomes of these experiments highlight the superior performance of our approach when compared against other methodologies addressing similar challenges.

2. Related work

Test-Time Adaptation. TTA is a particular setting of Domain Adaptation, encompassing two main characteristics: (*a*) adapting a model to a target domain, with inputs coming as unlabeled data streams (i.e., batches), (*b*) without any access to the source domain samples. The former challenge complicates accurately estimating the target domain's distribution, while the latter impedes solving the problem by directly comparing measures of the domain distributions (e.g., feature means). Despite the challenging nature of the problem, the field has gained important momentum in recent years, providing insights on the possibilities and limitations of adapting pre-trained models.

A key focus in TTA methods is adapting batch normalization layers, which retain important source domain information. PTBN [21] adjusts batch statistics at test time, while TENT [28] refines the affine parameters using entropy minimization on predictions. Entropy minimization, used in various methods [6, 15, 17, 31], enhances model confidence without label supervision but often depends on image augmentations or large batches. For example, Text-Prompt Tuning (TPT) [25] learns text prompt adapters for CLIP using entropy minimization. However, this approach needs to perform several augmentations for each test sample, making it computationally expensive. Test-Time Distribution Normalization (TTDN) [33] normalizes test data to match the training distribution, but needs access to source data or approximations of the mean of each test batch. Our method fine-tunes normalization layers, leverages prediction confidence for text supervision, and avoids input augmentations.

Recent techniques have also sought to adapt CLIP in a gradient-free manner. CALIP [7] adapts the visual and text features bidirectionally through a parameter-free attention module. The features are later combined to obtain refined predictions. Although efficient, this method needs the features of the entire test set for conducting a hyperparameter search, which limits its generalization to very large datasets. TDA [12] dynamically builds a positive and a negative cache to adapt features through the Tip-Adapter strategy [32]. This method however requires to find specific hyperparameters for each dataset, as in the original Tip-Adapter approach.

Pseudo-labels have been central to previous TTA methods. SHOT [16] uses them in a regularization loss based on mutual information, optimizing the entire feature encoder. CoTTA [29] employs pseudo-labels in a student-teacher model with consistency loss between original and augmented inputs. PAD [30] enhances pseudo-labels by augmenting inputs and voting on predictions. Instead of simple class pseudo-labels, our CLIPArTT method exploits the text supervision to better predict the correct class. By combining visual and text information into a single pseudo-label guided by the most probable classes, we can direct the model towards a more certain prediction.

Other methods take less conventional approaches to TTA. LAME [2] refines classifier predictions transductively using feature similarity through the Laplacian. Test-time training (TTT) methods [5, 20, 22, 26, 27] train a sub-branch alongside the main network in an unsupervised manner to update the model, requiring training from scratch on the source domain. Like LAME, our method relates to Laplacian regularization but applies it differently. While LAME updates predictions without altering the model, our method uses it in a test-time adaptation loss to enforce consistency between embeddings of related batch samples. Unlike TTT methods, we do not require additional branches in the network or training this network from scratch.

Conformal Learning. CLIPArTT is also related to the field of conformal learning, where intervals of confidence for new predictions are derived from previous experience [24]. In a conformal prediction, a given level of certainty is assigned to a set $\mathcal{C} = \{c_1, ..., c_K\}$ with the K most plausible classes that an input can belong to [1]. We draw inspiration from this concept and build a conformal set of class predictions that can help adapting CLIP towards an accurate top-1 prediction. Our technique, however, stands out by not relying on image transformations such as in [14, 25] to filter out predictions.

3. Methodolody

We start by presenting the vanilla CLIP model for classification and explain how it can be extended to test-time adaptation using entropy minimization. Building on the limitations of this approach, we then introduce our CLIPArTT method that leverages class uncertainty and the relationship between samples in a batch.

3.1. CLIP-based classification

Contrastive Language-Image Pre-training (CLIP) [23] consists of a visual encoder $f_\theta^v(\cdot)$, mapping an image $\mathbf{x}$ to visual features $\mathbf{z}^v \in \mathbb{R}^D$, and a text encoder $f_\theta^t(\cdot)$ transforming text prompts $\mathbf{t}$ to text features $\mathbf{z}^t \in \mathbb{R}^D$. The visual and text encoders are trained jointly with a contrastive loss so that the feature embeddings of training images and their corresponding text prompt are close to each other, while those of different training examples are pushed apart.

In a classification task with K fixed classes, CLIP can be used to perform inference by encoding a pre-defined text prompt for each class, for example $\mathbf{t}_k$ = "`a photo of a {class k}`". For a new image $\mathbf{x}_i$, the probability of

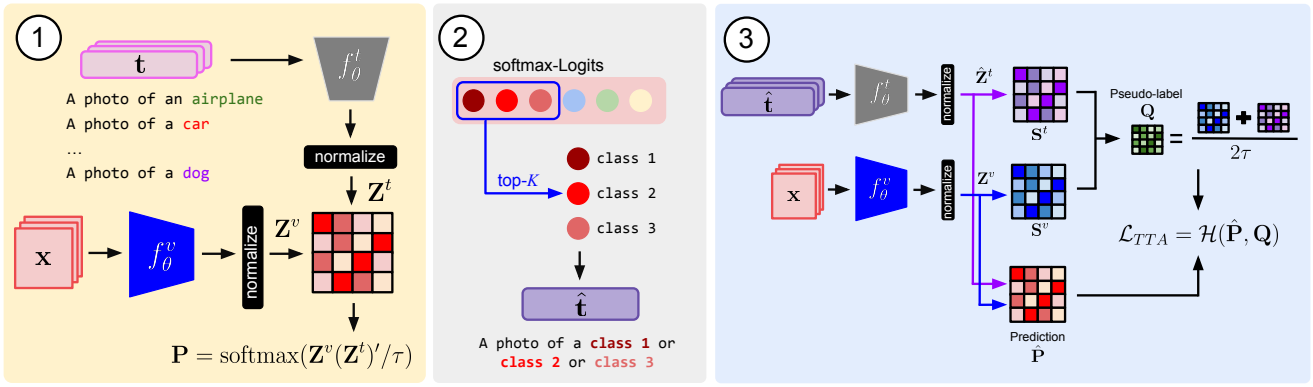

Figure 1. CLIPArTT pipeline overview: 1) Computing predictions from Image-Text Similarity, 2) generating a new text prompt by filtering the top-K class predictions, 3) with the new prompts, a pseudo-label $\mathbf{Q}$ is obtained by averaging the image-to-image and text-to-text similarity scores, while the prediction $\hat{\mathbf{P}}$ is computed as the image-to-text similarity. Cross-entropy is then used as the TTA loss.

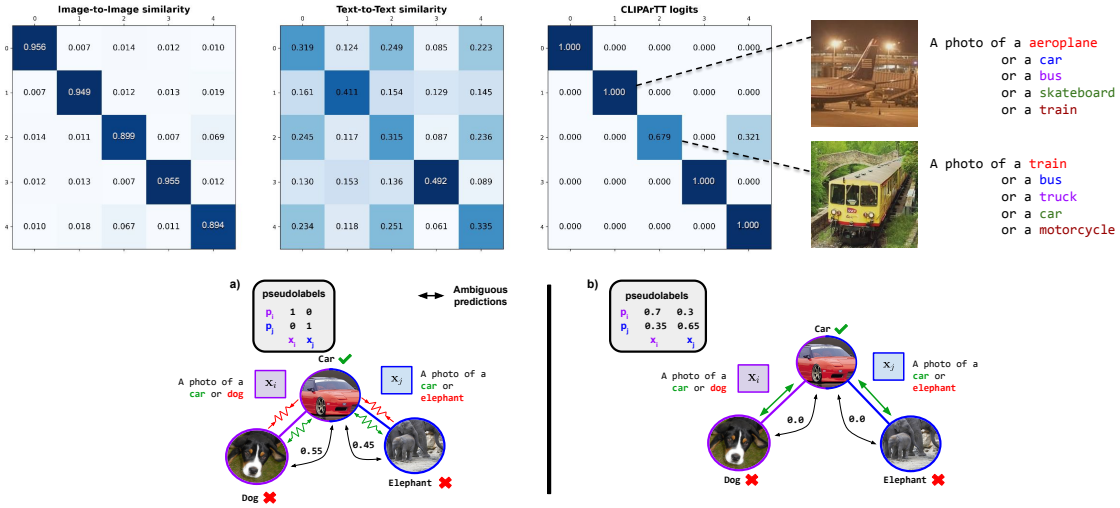

Figure 2. **Top**: Example of similarity matrices ($\mathbf{S}^v$, $\mathbf{S}^t$) and CLIPArTT softmax probabilities ($\mathbf{Q}$) for a batch of 5 examples and using $K = 5$ classes. **Bottom**: a) When using the identity matrix as pseudo-label for contrastive learning, the correct prediction is ambiguous, as the images are forced to both approaching and moving away from the right class. b) CLIPArTT uses soft pseudo-labels that smoothly guides the prediction towards the correct class by reducing the impact of ambiguities in the prompts.

belonging to class k is then estimated based on cosine similarity,

$$p_{ik} = \frac{\exp\left(\cos(\mathbf{z}_i^v, \mathbf{z}_k^t)/\tau\right)}{\sum_j \exp\left(\cos(\mathbf{z}_i^v, \mathbf{z}_j^t)/\tau\right)}, \quad \cos(\mathbf{z}, \mathbf{z}') = \frac{\mathbf{z}^\top \mathbf{z}'}{\|\mathbf{z}\|_2 \cdot \|\mathbf{z}'\|_2}, \tag{1}$$

where τ is a suitable softmax temperature.

The model in Eq. (1) can be used for test-time adaptation in various ways, the simplest one being entropy minimization as in TENT. This approach, which relies on the principle that the decision boundary lies in a low-density region of space, giving rise to a low prediction entropy, adapts the model parameters by minimizing entropy on a test batch of size B:

$$\mathcal{L}_{\text{TENT}}(\theta) = -\frac{1}{B} \sum_{i=1}^{B} \sum_{k=1}^{K} p_{ik} \log p_{ik} \tag{2}$$

However, this approach, as similar ones based on pseudo-labels [16, 29, 30], suffers from two important limitations. First, due to domain shifts, the model's prediction may be unreliable (e.g., giving the highest probability to the wrong class) and techniques such as entropy minimization or standard pseudo-label will only reinforce these errors during adaptation. Secondly, they assume that samples in a test batch are independent and do not directly leverage their semantic relationships.

	CIFAR10		CIFAR100	
	Top-1	Top-3	Top-1	Top-3
Original	88.74	97.79	61.68	80.92
Corrupted (-C)	59.22	82.43	29.43	46.61

Table 1. Accuracy (%) on CIFAR-10/100 and CIFAR-10/100-C datasets with Level 5 corruption for the top-1 or the top-3 predicted classes.

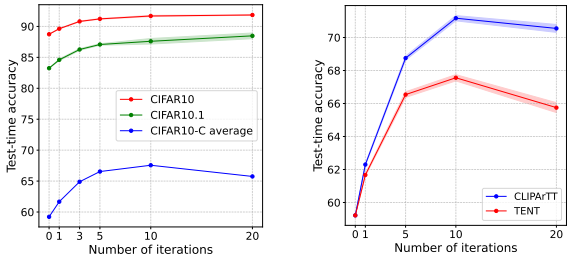

Figure 3. Evolution of CLIPArTT's accuracy during test-time adaptation. **Left**: For different versions of CIFAR10. **Right**: Compared to TENT on CIFAR10-C.

3.2. Our CLIPArTT method

The proposed CLIPArTT method (see Figure 1), addresses the above-mentioned limitations using two key insights. The first insight, inspired by conformal learning, is that the correct class is often among the top most probable ones, although the model's most confident prediction may not be always correct. This claim is supported by the results in Table 1 (first row: *Original*), showing that the correct class is within the top-3 predictions 97.79% of the times for CIFAR-10 (versus 88.74% within the top-1), and 80.92% of the times for CIFAR-100 (versus 61.68% within the top-1). A way to include information about multiple classes could, therefore, help adapting the model at test time. The second insight is that the similarity between the batch samples could be evaluated based on their visual and/or text embeddings. As we will show below, such similarities could be exploited in a strategy related to Stochastic Neighbor Embedding (SNE) and graph-Laplacian regularization.

Instance-specific multi-class prompt. Inspired by our first insight and recent work investigating CLIP's ability to perform compositional logical reasoning [3], we devise a novel technique to generate instance-specific prompts from the top-k predictions, with $1 \leq k \ll K$. Specifically, for an image $\mathbf{x}_i$, we estimate the CLIP-based class probabilities using Eq. (1) and then generate a new text prompt as $\hat{\mathbf{t}}_i$ = "`a photo of a {class` i_1`} or ... or {class` i_k`}`", where {`class` i_j} is the name of the class with j-th highest probability.

Transductive TTA. Next, we design a test-time adaptation loss that accounts for semantic relationships between batch samples. Let $\mathbf{Z}^v \in \mathbb{R}^{B\times D}$ and $\hat{\mathbf{Z}}^t \in \mathbb{R}^{B\times D}$ denote the *normalized* visual and instance-specific text embeddings of the samples within the test batch, respectively. We compute an image-to-image similarity matrix, $\mathbf{S}^v = \mathbf{Z}^v(\mathbf{Z}^v)^\top \in [-1,1]^{B\times B}$, and a text-to-text similarity matrix, $\mathbf{S}^t = \hat{\mathbf{Z}}^t(\hat{\mathbf{Z}}^t)^\top \in [-1,1]^{B\times B}$. The former measures the affinity between each pair of samples within the batch in terms of their visual characteristics (shapes, textures, etc.). The latter captures common (or related) classes in the top-k predictions of two samples, since it is computed using the instance-specific multi-class prompts. As illustrated in Figure 2 (*top*), a broad range of similarity values are obtained with this approach.

We deploy these two pairwise similarity matrices to compute pseudo-labels as follows:

$$\mathbf{Q} = \text{softmax}\big((\mathbf{S}^v + \mathbf{S}^t)/2\tau\big) \in [0,1]^{B\times B} \tag{3}$$

where the softmax operation is applied column-wise and the temperature $\tau = 0.01$ is used in all our experiments.

Let $\hat{\mathbf{P}}$ denote the zero-shot prediction matrix using our instance-specific multi-class text prompts:

$$\hat{\mathbf{P}} = \text{softmax}\big(\mathbf{Z}^v(\hat{\mathbf{Z}}^t)^\top/\tau\big),\ \hat{p}_{ij} = \frac{\exp\big(\cos(\mathbf{z}_i^v, \hat{\mathbf{z}}_j^t)/\tau\big)}{\sum_k \exp\big(\cos(\mathbf{z}_i^v, \hat{\mathbf{z}}_k^t)/\tau\big)} \tag{4}$$

This matrix, along with the pairwise pseudo-labels we introduced in Eq. (3), yield our final TTA loss based on cross-entropy:

$$\mathcal{L}_{\text{TTA}}(\theta) = -\frac{1}{B}\sum_{i=1}^{B}\sum_{j=1}^{B} q_{ij}\log\hat{p}_{ij}. \tag{5}$$

Unlike recent approaches like TPT [25], which adapt CLIP by learning a text prompt, we instead update the normalization-layer parameters of the *visual* encoder, which yields a light-weight, computationally efficient TTA method. We note that TPT requires to generate multiple augmentations during the forward pass, incurring a substantial overhead.

The mechanism of CLIPArTT is illustrated in Fig. 2. Our soft pseudo-label reduces the risk of increasing the ambiguity of the predictions. Using the identity matrix (i.e., hard pseudo-labels) leads to uncertain class assignments, as the ambiguity of the text prompts would both attract and repel the right class if it is present in the same text prompt for two different images.

3.3. Link to existing techniques

An important element of our TTA method is that it exploits the similarities between pairs of samples within the batch, in terms of visual features and features of text prompts representing the top-k classes. In this section, we draw a connection between the proposed method and two well-known techniques in machine learning: Stochastic Neighbor Embedding [9] and graph-Laplacian regularization [4, 10].

Stochastic Neighbor Embedding (SNE). This popular technique for dimensionality reduction estimates the local probability of a point $\mathbf{x}_j$ given a point $\mathbf{x}_i$ based on their Euclidean distance $d_{ij} = \|\mathbf{x}_i - \mathbf{x}_j\|_2$:

$$p_{ij} = \frac{\exp\left(-d_{ij}^2\right)}{\sum_{k \neq i} \exp\left(-d_{ik}^2\right)} \tag{6}$$

The goal is to find, for each $\mathbf{x}_i$, a low-dimensional embedding $\mathbf{y}_i$ such that the local probabilities q_{ij} computed using Eq. (6) on these embeddings, is similar to p_{ij}. This is achieved by minimizing the row-wise KL divergence between distribution matrices $\mathbf{P}$ and $\mathbf{Q}$. Our CLIPArTT method can be linked to SNE since $\cos(\mathbf{x}_i, \mathbf{x}_j) = -d_{ij}^2 + 2$ when $\mathbf{x}_i$, $\mathbf{x}_j$ are unit normalized, and KL divergence between $\mathbf{P}$ and $\mathbf{Q}$ is equal to the cross-entropy between these matrices minus the entropy of $\mathbf{P}$. In summary, our TTA loss in Eq. (5) ensures that the inter-modality (text-to-image) similarities of batch samples are aligned with their intra-modality ones (text-to-text and image-to-image).

Graph-Laplacian regularization is commonly-used in the context semi-supervised learning techniques [4] and label propagation [10]. In our case, the Laplacian regularizer is computed over a set of B nodes with predictions $\mathbf{Z} \in \mathbb{R}^{B \times D}$ as:

$$\mathcal{L}_{\text{reg}}(\mathbf{Z}) = \text{trace}\big(\mathbf{Z}^\top \mathbf{L}_{\text{W}} \mathbf{Z}\big) = \sum_{i,j} w_{ij} \|\mathbf{z}_i - \mathbf{z}_j\|_2^2, \tag{7}$$

where $\mathbf{L}_{\text{W}} \in \mathbb{R}^{B \times B}$ is the Laplacian matrix defined from edge weight matrix $\mathbf{W}$ as follows:

$$[\mathbf{L}_{\text{W}}]_{ij} = \begin{cases} d_{ii} = \sum_j w_{ij}, & \text{if } i = j \\ -w_{ij} & \text{else.} \end{cases} \tag{8}$$

This connection is made in the following proposition.

Proposition 1. *The TTA loss in Eq. (5) can be expressed as a Laplacian regularization over a bipartite graph with one set of nodes for image embeddings and another for text embeddings.*

Proof. Expanding the softmax in Eq. (5), we get

$$\mathcal{L}_{\text{TTA}} = -\frac{1}{B} \sum_{i=1}^{B} \sum_{j=1}^{B} q_{ij} \log \frac{\exp\left((\mathbf{z}_i^v)^\top \mathbf{z}_j^t / \tau\right)}{\sum_k \exp\left((\mathbf{z}_i^v)^\top \mathbf{z}_k^t / \tau\right)} \tag{9}$$

$$= \frac{1}{\tau B} \sum_i \sum_j q_{ij} \Big(-(\mathbf{z}_i^v)^\top \mathbf{z}_j^t + \tau \underbrace{\log \sum_k \exp\left((\mathbf{z}_i^v)^\top \mathbf{z}_k^t / \tau\right)}_{\text{LogSumExp (LSE)}} \Big) \tag{10}$$

Since the feature embeddings are normalized, we have $\|\mathbf{z}_i^v\|_2 = \|\mathbf{z}_j^t\|_2 = 1$ and thus $(\mathbf{z}_i^v)^\top \mathbf{z}_j^t = 2 - \|\mathbf{z}_i^v - \mathbf{z}_j^t\|_2^2$. Moreover, following the scaled LogSumExp (LSE) rule, we can bound the right-side term as follows:

$$\max_k \{(\mathbf{z}_i^v)^\top \mathbf{z}_k^t\} < \tau LSE \leq \max_k \{(\mathbf{z}_i^v)^\top \mathbf{z}_k^t\} + \tau \log B \tag{11}$$

For a small τ (we use a value of 0.01 in our method), the bounds tighten and we get that

$$LSE \approx \tfrac{1}{\tau} \max_k \{(\mathbf{z}_i^v)^\top \mathbf{z}_k^t\} = \tfrac{1}{\tau} (\mathbf{z}_i^v)^\top \mathbf{z}_i^t \tag{12}$$

The last equality comes from the hypothesis that, in a well trained model, the maximum similarity occurs between the image embedding of a sample and its corresponding text embedding. Our TTA loss can then be expressed as

$$\mathcal{L}_{\text{TTA}} \approx \frac{1}{\tau B} \Big(\sum_{i,j} q_{ij} \|\mathbf{z}_i^v - \mathbf{z}_j^t\|_2^2 + \sum_i (\mathbf{z}_i^v)^\top \mathbf{z}_i^t \Big) + const$$

$$= \frac{1}{\tau B} \text{trace}\big((\mathbf{Z}^v)^\top (\mathbf{L}_{\text{Q}} + \mathbf{I})\, \mathbf{Z}^t\big) + const \tag{13}$$

where $\mathbf{I}$ is the identity matrix. The modified Laplacian $\mathbf{L}_{\text{Q}} + \mathbf{I}$ enforces nodes with a high connection weight (q_{ij}) to have similar embeddings, while also avoiding embeddings to collapse into a single vector. □

4. Experimental Settings

We evaluate CLIPArTT's performance across a variety of TTA datasets, covering scenarios such as natural images, common corruptions, simulated images, video, and Domain Generalization benchmark. This comprehensive framework allows for a robust assessment of the model's adaptability to various challenges, including domain shifts and corruptions. For a detailed description of the datasets, please refer to the supplementary materials.

Test-time adaptation. For test-time adaptation, model updates are applied to all the Layer Normalization (LN) layers within the visual encoder. We employ the Adam optimizer with a fixed learning rate set to 10^{-3}. Throughout our experiments, a consistent batch size of 128 is utilized to maintain uniformity and enable effective comparisons across different scenarios. As in previous TTA works, a smaller learning rate of 10^{-4} is preferred to adapt to the 3D renderings split [27], as it represents a more aggressive shift.

Benchmarking. We compare CLIPArTT with *state-of-the-art* methods. Specifically, we utilize adapted versions of TENT [28] and LAME [2] tailored to CLIP. While the classifier logits are used in previous TTA research involving CNNs, we follow the standard practice and use the image-to-text similarity to obtain the final logits in CLIP. Only the visual encoder is optimized where needed. TENT now updates only the affine parameters in LN layers through entropy minimization directly on these logits. Notably, we employ 10 iterations for TENT adaptation, a choice informed by our findings, as evidenced in Fig 3. In LAME,

	$K=1$	$K=3$	$K=4$
CIFAR-10	89.80 ±0.05	90.04 ±0.13	**90.41 ±0.07**
CIFAR-10.1	85.37 ±0.17	**86.35 ±0.27**	86.07 ±0.21
CIFAR-10-C	70.79 ±0.15	**71.17 ±0.16**	70.99 ±0.15

Table 2. Accuracy (%) on CIFAR-10, CIFAR-10.1 and CIFAR-10-C datasets with Level 5 corruption for different number of K selected classes to create pseudo-labels.

	Iter = 1	Iter = 5	Iter = 10	Iter = 20
CIFAR-10	89.59 ±0.01	**90.54 ±0.09**	90.04 ±0.13	88.32 ±0.12
CIFAR-10.1	84.78 ±0.02	**86.67 ±0.06**	86.35 ±0.27	84.33 ±0.31
CIFAR-10-C	62.30 ±0.06	68.75 ±0.12	**71.17 ±0.16**	70.55 ±0.24

Table 3. Accuracy (%) on CIFAR-10, CIFAR-10.1 and CIFAR-10-C datasets with Level 5 corruption for different number of iterations to update the model at test-time.

the Laplacian regularizer is applied on the similarity between image features (obtained from the visual encoder). Finally, we include CLIP-tailored methods TTDN [33], TPT [25], TDA [12] and CALIP [7]. Both TDA and CALIP require hyperparameter tuning for optimal performance. To ensure fairness in our benchmarking process, we keep the hyperparameters of these approaches as reported in their original works. The batch-size for TPT is reduced to 32 due to its reliance on image augmentations and high memory requirements.

5. Results

In this section, we outline the experimental outcomes derived from CLIPArTT and provide a comprehensive analysis of results. We first show a series of exploratory ablations that later help conduct the final experiments on the different datasets and compare with *state-of-the-art* approaches.

5.1. Ablation Studies

Number of classes for new prompts. Determining the number of classes in building new prompts presents a critical aspect in our methodology. Table 1 illustrates various scenarios where $K = 3$ emerges as a favorable choice, exhibiting a remarkable accuracy above 90% across multiple instances of CIFAR-10-C datasets. This initial observation is further corroborated by the findings presented in Table 2. However, the decision becomes more nuanced when applied to CIFAR-100 datasets. While leveraging the top 3 classes contributes to enhanced accuracy, it does not guarantee optimal performance. Similarly, adopting a strategy akin to CIFAR-10, wherein 30% of the classes are chosen, proves impractical due to resulting lengthy sentences that are impossible to tokenize. Nonetheless, our analysis shows that $K = 3$ consistently yields superior performance, particularly evident in the average results across CIFAR-100-C. Consequently, we maintain $K = 3$ as the preferred configuration for all subsequent experiments on CIFAR datasets.

	Avg. Prompt	Image	Text	Image + Text
CIFAR-10	76.94 ±0.41	**90.18 ±0.02**	89.05 ±0.14	90.04 ±0.13
CIFAR 10.1	67.83 ±0.49	86.25 ±0.37	84.85 ±0.40	**86.35 ±0.27**
CIFAR-10-C	43.08 ±0.08	70.98 ±0.15	70.03 ±0.21	**71.17 ±0.16**

Table 4. Accuracy (%) on CIFAR-10, CIFAR-10.1 and CIFAR-10-C datasets with different targets.

	CLIP	BS = 16	BS = 32	BS = 64	BS = 128
CIFAR-10	88.74	85.89 ±0.19	88.25 ±0.15	89.48 ±0.15	90.04 ±0.13
CIFAR-10.1	83.25	81.55 ±0.53	84.00 ±0.31	85.40 ±0.08	86.35 ±0.27
CIFAR-10-C	59.22	64.72 ±0.23	67.70 ±0.23	69.82 ±0.20	71.17 ±0.16

Table 5. Accuracy (%) on CIFAR-10, CIFAR-10.1 and CIFAR-10-C datasets with Level 5 corruption for different number of batch-size.

Comparison against prompt averaging. The combination of K classes inside the template {`class` i_1} `or ... or` {`class` i_k}, poses a sensible question: how comparable is this alternative against a simple prompt averaging? In CLIPArTT, we hypothesize that combining the names of the most probable classes in a single text prompt better leverages CLIP's language understanding capabilities. Experiments conducted on CIFAR-10 and its variants are presented in Table 4, showing that our template, while simple, gives rise to an important accuracy gain with respect to prompt averaging.

Number of iterations. In line with our approach for TENT, we investigate the optimal number of adaptation iterations. As demonstrated in Table 3, iterating 10 times strikes the most favorable balance, exhibiting superior performance across the overall average on CIFAR-10-C and yielding optimal results for numerous corruption types. Nonetheless, it is imperative to acknowledge that employing a lower number of iterations (e.g., Iter = 5) may yield improved outcomes in scenarios with minimal or no distribution shift, while a higher number of iterations (e.g., Iter = 20) may be more effective in mitigating severe distribution shifts. As a result, 10 iterations are used for all forthcoming experiments.

Pseudo-label selection for adaptation. For the target, we employed a linear combination of both image-to-image and text-to-text similarities to encapsulate the information derived from both modalities. This combination proved to be more effective than each modality separately (see Table 4). While selecting solely the text-to-text similarity for adaptation performs well, it falls short of achieving optimal results. Conversely, utilizing only the image-to-image similarity mainly improves for the original CIFAR-10 dataset and few corruption types.

Performance of the model for different batch-sizes. TTA methods have historically exhibited limitations when applied with small batch sizes. In our study, we investigate

	CIFAR-100-C				ImageNet-C			
	CLIP	TENT	TDA	CLIPArTT	CLIP	TENT	TDA	CLIPArTT
Gaussian Noise	14.80	14.38 ±0.14	8.20 ±0.35	**25.32 ±0.14**	12.27	12.28 ±0.05	11.54 ±0.05	**20.18 ±0.61**
Shot noise	16.03	17.34 ±0.27	9.58 ±0.43	**27.90 ±0.05**	12.20	8.46 ±0.08	12.13 ±0.13	**20.94 ±0.08**
Impulse Noise	13.85	10.03 ±0.13	7.63 ±0.19	**25.62 ±0.09**	12.89	15.71 ±0.04	12.12 ±0.08	**19.95 ±0.02**
Defocus blur	36.74	49.05 ±0.07	25.59 ±0.41	**49.88 ±0.23**	21.60	20.61 ±7.05	21.39 ±0.08	**24.51 ±0.13**
Glass blur	14.19	3.71 ±0.07	9.83 ±0.56	**27.89 ±0.03**	10.84	17.20 ±0.08	10.74 ±0.06	**19.51 ±0.02**
Motion blur	36.14	46.62 ±0.27	28.92 ±0.18	**47.93 ±0.14**	17.85	24.60 ±0.04	18.45 ±0.06	**25.80 ±1.16**
Zoom blur	40.24	51.84 ±0.15	31.08 ±0.36	**52.70 ±0.06**	16.38	21.13 ±0.08	16.83 ±0.03	**24.00 ±0.06**
Snow	38.95	46.71 ±0.21	32.94 ±0.12	**49.72 ±0.01**	21.91	24.19 ±0.09	22.97 ±0.05	**27.26 ±0.06**
Frost	40.56	44.90 ±0.27	34.84 ±0.25	**49.63 ±0.12**	23.33	23.34 ±0.06	24.68 ±0.06	**26.56 ±0.08**
Fog	38.00	47.31 ±0.04	31.13 ±0.15	**48.77 ±0.04**	26.39	28.58 ±0.04	27.72 ±0.05	**34.68 ±0.02**
Brightness	48.18	60.58 ±0.18	42.36 ±0.10	**61.27 ±0.08**	45.87	47.16 ±0.01	48.06 ±0.04	**47.21 ±0.07**
Contrast	29.53	45.90 ±0.11	18.03 ±0.07	**48.55 ±0.24**	16.16	21.62 ±0.01	16.09 ±0.05	**23.38 ±0.12**
Elastic transform	26.33	33.09 ±0.08	18.88 ±0.24	**37.45 ±0.08**	16.04	17.17 ±0.01	17.75 ±0.01	**23.95 ±0.04**
Pixelate	21.98	26.47 ±0.09	14.59 ±0.30	**33.88 ±0.14**	28.00	33.56 ±1.71	30.03 ±0.05	**34.12 ±0.09**
JPEG compression	25.91	29.89 ±0.07	17.56 ±0.11	**36.07 ±0.32**	27.44	31.89 ±0.02	29.36 ±0.01	**32.23 ±0.12**
Average	29.43	35.19	22.08	**41.51**	20.61	23.17	21.32	**26.95**

Table 6. Accuracy (%) on CIFAR-100-C and ImageNet-C datasets with ViT-B/32 as visual encoder.

the performance of our model in this regard. As depicted in Table 5, performance improves significantly with increasing batch size. However, beyond a batch size of 8, no further performance gains are observed. This phenomenon can be attributed to the fact that, with smaller batch sizes, our method can potentially introduce uncertainty by using multiple classes. This leads to a lower performance, especially when the model is already confident. In subsequent experiments, we maintain a batch size of 128, consistent with prevailing practices in *state-of-the-art* methodologies.

5.2. Comparison on different datasets

In standard TTA, the concept of domain shift is strictly related to the source dataset used for pretraining. In the context of CLIP, the presence of domain shifts is less evident, as the source data contained an enormous amount of images that likely include different domains. For this reason, we categorize our experiments per dataset type.

Natural images. In the different Tables 7 and 8, CLIPArTT consistently enhances accuracy across CIFAR-10, CIFAR10.1, and CIFAR-100 datasets compared to the baseline (+2%, +3%, +8% respectively with ViT-B/32). This improvement suggests that uncorrupted datasets, which may not have been seen by the model during training, can benefit from adaptation. Consequently, a zero-shot model can leverage adaptation at test-time, relying solely on the model's predictions. However, while CLIPArTT outperforms the baseline, as well as LAME, TPT and TTDN, it is worth noting that TENT may yield superior results depending on the visual encoder and the number of classes. Entropy minimization used solely on confident results proves to be also effective. Interestingly, the corruptions in which TENT achieves a better performance tend to have a higher accuracy before adaptation, limiting the impact of error reinforcement on this approach. For a larger dataset such as ImageNet, most methods perform similarly, with the exception of TDA and CALIP, which improve CLIP performance by over 2%. Notably, both TDA and CALIP require hyperparameter tuning, and we used the recommended settings optimized for ImageNet.

Varied styles and textures. Next we compare methods on two datasets, PACS and OfficeHome, which contain images from very different domains and are often used as benchmark in domain generalization. Once again, these domains do not constitute real distribution shifts for CLIP which was trained with a broad range of datasets. As can be seen, the best performance is achieved by methods such as TTDN for this setting. Nevertheless, it is important to highlight CLIPArTT's robustness, as it consistently maintains or improves model performance after adaptation. For instance, on the PACS dataset, CLIPArTT achieves an accuracy of 93.95%, compared to 93.65% for the original CLIP.

Common corruptions. In the various referenced tables, including Tables 7 and 8, CLIPArTT consistently outperforms compared approaches across all visual encoders, resulting in enhancements of up to 13% on average. A more detailed examination reveals that CLIPArTT frequently surpasses TENT, particularly in instances where the baseline performs poorly. Indeed, TENT tends to compromise the model's performance under conditions of low baseline confidence. For instance, in Table 6, TENT experiences a 3% decrease compared to the baseline under *Impulse Noise*, while CLIPArTT exhibits a 12% improvement. In Figure 3 (*right*), it is evident that CLIPArTT achieves better performance more rapidly across varying numbers of iterations. Conversely, when the baseline is confident, CLIPArTT con-

	Shift	CLIP	LAME	TENT	TPT (BS=32)	TTDN	TDA	CALIP	CLIPArTT
CIFAR-10	✗	88.74	89.36 ±0.06	**91.69 ±0.10**	88.06 ±0.06	89.06 ±0.02	84.09 ±0.04	86.79 ±0.01	90.04 ±0.13
CIFAR-10.1	✗	83.25	81.22 ±0.33	**87.60 ±0.45**	81.80 ±0.27	83.77 ±0.06	78.98 ±0.37	81.02 ±0.03	86.35 ±0.27
CIFAR-100	✗	61.68	58.27 ±0.17	69.74 ±0.16	63.78 ±0.28	64.10 ±0.05	60.32 ±0.06	61.94 ±0.01	**69.79 ±0.04**
ImageNet	✗	61.81	61.17 ±0.02	62.13 ±0.02	60.74 ±0.16	61.85 ±0.02	**64.49 ±0.01**	58.50 ±0.00	61.39 ±0.09
VisDA-C (YT)	✗	84.45	84.72 ±0.01	84.41 ±0.02	82.95 ±0.01	83.29 ±0.03	84.47 ±0.08	**85.02 ±0.00**	83.46 ±0.03
PACS	✗	93.65	93.68 ±0.04	93.81 ±0.03	93.23 ±0.12	**94.77 ±0.07**	92.94 ±1.08	94.11 ±0.00	93.95 ±0.06
Office Home	✗	77.53	75.46 ±0.12	77.68 ±0.06	77.20 ±0.24	**79.76 ±0.04**	77.70 ±0.13	78.75 ±0.01	77.56 ±0.09
CIFAR-10-C	✓	59.22	50.57 ±0.30	67.56 ±0.19	56.80 ±0.22	60.13 ±0.07	48.00 ±0.68	56.58 ±0.0	**71.17 ±0.16**
CIFAR-100-C	✓	29.43	26.23 ±0.12	35.19 ±0.16	30.46 ±0.14	31.90 ±0.04	22.08 ±0.29	29.92 ±0.01	**41.51 ±0.15**
ImageNet-C	✓	20.61	20.02 ±0.16	23.17 ±0.54	21.10 ±0.06	21.28 ±0.02	21.32 ±0.06	21.00 ±0.00	**26.95 ±0.12**
VisDA-C (3D)	✓	84.51	84.75 ±0.17	83.85 ±0.03	78.45 ±0.05	80.45 ±0.01	84.22 ±0.10	83.81 ±0.08	**87.24 ±0.04**

Table 7. Accuracy (%) on CIFAR-10, CIFAR-10.1 and CIFAR-10-C datasets with ViT-B/32 as visual encoder. The Shift column denotes a domain shift from standard images recognized by CLIP: corruption in CIFAR-10-C, CIFAR-100-C and ImageNet-C datasets, synthetic images in VisDA-C (3D).

	Backbone	CLIP	TENT	CLIPArTT
CIFAR-10.1	ViT-B/16	84.00	88.52 ±0.33	**88.72 ±0.33**
	ViT-L/14	91.20	**92.22 ±0.25**	91.02 ±0.02
CIFAR-10-C	ViT-B/16	60.15	68.00 ±0.18	**73.22 ±0.17**
	ViT-L/14	76.04	**79.18 ±0.10**	78.06 ±0.15
CIFAR-100-C	ViT-B/16	32.01	37.90 ±0.18	**40.08 ±0.18**
	ViT-L/14	44.59	50.14 ±0.15	**52.52 ±0.18**

Table 8. Accuracy (%) on CIFAR-10, CIFAR-10.1, CIFAR-10-C, CIFAR-100 and CIFAR-100-C datasets with ViT-B/16 and ViT-L/14 as visual encoders.

sistently maintains a high performance. Finally, our method performs effectively on larger datasets with a greater number of classes, consistently improving upon the baseline by 6% on ImageNet-C.

Simulated images and video. The performance of out method on the YT dataset is suboptimal, a trend that can be observed in scenarios where the CLIP model already exhibits high confidence. However, our observations indicate that on the 3D dataset, CLIPArTT demonstrates a notable competitive advantage over other Test-Time Adaptation (TTA) methods and the baseline, achieving an improvement of nearly 3%. Additionally, while TENT – often considered one of the more robust methods in handling various corruptions – results in a performance decline relative to the baseline, CLIPArTT maintains its effectiveness.

5.3. Limitations

Our results indicate that CLIPArTT performs particularly well under significant domain shifts such as corruptions, though it is slightly less effective on natural images. While the performance of our method in these cases is not as strong as that of TENT, it remains stable without deteriorating CLIP's capabilities, whereas the performance of TENT degrades under severe domain shifts. It is important to mention that any dataset not included in CLIP's pre-training can be considered a potential domain shift. However, natural images are more likely to have been seen by CLIP during pre-training, making adaptation for them prone to overfitting. This may explain why CLIPArTT excels in more challenging scenarios. To further explore the potential limitations of our method, we have included two additional scenarios in the supplementary material: *a)* evaluation in a highly imbalanced setting, where only C randomly chosen classes are present in a batch at random, and *b)* open-set classification, where out-of-distribution images (i.e., from classes not present in the text prompts) are introduced into the batches.

6. Conclusion

We introduce CLIPArTT, a novel Test-Time Adaptation framework designed specifically for VLMs. By leveraging the model's predictions as new pseudo-labels, we effectively minimize cross-entropy at test-time, thereby enhancing model performance even within a zero-shot setting.

A comprehensive ablation study determined optimal hyperparameters and helped gain deeper insights into the various model configurations. Our experimental results demonstrate that CLIPArTT achieves highly competitive performance across TTA datasets, surpassing *state-of-the-art* approaches. While TENT remains a strong competitor, our model offers enhanced versatility by effectively addressing both natural and severe domain shifts, thus exhibiting robustness across various scenarios.

Exploring the potential of text prompts in classification is a promising avenue for future research. Investigating alternative methods to fine-tune these text prompts could yield valuable insights. Moreover, extending the study of Test-Time Adaptation to other scenarios, such as segmentation or object detection with Vision-Language Models (VLMs), holds significant promise for advancing our understanding of model adaptability and performance across diverse tasks.

References

[1] Anastasios N Angelopoulos and Stephen Bates. A gentle introduction to conformal prediction and distribution-free uncertainty quantification. *arXiv preprint arXiv:2107.07511*, 2021. 2

[2] Malik Boudiaf, Romain Mueller, Ismail Ben Ayed, and Luca Bertinetto. Parameter-free online test-time adaptation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8344–8353, 2022. 2, 5

[3] Justin Brody. On the potential of CLIP for compositional logical reasoning. *arXiv preprint arXiv:2308.15887*, 2023. 4

[4] Olivier Chapelle, Bernhard Scholkopf, and Alexander Zien. Semi-supervised learning. 2006. *Cambridge, Massachusettes: The MIT Press View Article*, 2, 2006. 4, 5

[5] Yossi Gandelsman, Yu Sun, Xinlei Chen, and Alexei A Efros. Test-time training with masked autoencoders. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho, editors, *Advances in Neural Information Processing Systems*, 2022. 2

[6] Sachin Goyal, Mingjie Sun, Aditi Raghunanthan, and Zico Kolter. Test-time adaptation via conjugate pseudo-labels. *Advances in Neural Information Processing Systems*, 2022. 2

[7] 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 Thirty-Seventh AAAI Conference on Artificial Intelligence and Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence and Thirteenth Symposium on Educational Advances in Artificial Intelligence*, AAAI'23/IAAI'23/EAAI'23. AAAI Press, 2023. 2, 6

[8] Andrey Guzhov, Federico Raue, Jörn Hees, and Andreas Dengel. AudioCLIP: Extending CLIP to image, text and audio. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 976–980. IEEE, 2022. 1

[9] Geoffrey E Hinton and Sam Roweis. Stochastic neighbor embedding. *Advances in neural information processing systems*, 15, 2002. 4

[10] Ahmet Iscen, Giorgos Tolias, Yannis Avrithis, and Ondrej Chum. Label propagation for deep semi-supervised learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5070–5079, 2019. 4, 5

[11] 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

[12] Adilbek Karmanov, Dayan Guan, Shijian Lu, Abdulmotaleb El Saddik, and Eric Xing. Efficient test-time adaptation of vision-language models, 2024. 2, 6

[13] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. *arXiv preprint arXiv:2304.02643*, 2023. 1

[14] Zhengfeng Lai, Noranart Vesdapunt, Ning Zhou, Jun Wu, Cong Phuoc Huynh, Xuelu Li, Kah Kuen Fu, and Chen-Nee Chuah. PADCLIP: Pseudo-labeling with adaptive debiasing in CLIP for unsupervised domain adaptation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 16155–16165, October 2023. 1, 2

[15] Jungsoo Lee, Debasmit Das, Jaegul Choo, and Sungha Choi. Towards open-set test-time adaptation utilizing the wisdom of crowds in entropy minimization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 16380–16389, October 2023. 2

[16] Jian Liang, Dapeng Hu, and Jiashi Feng. Do we really need to access the source data? Source hypothesis transfer for unsupervised domain adaptation. 2020. 2, 3

[17] Guoliang Lin, Hanjiang Lai, Yan Pan, and Jian Yin. Improving entropy-based test-time adaptation from a clustering view. *arXiv preprint arXiv:2310.20327*, 2023. 2

[18] Ziyi Lin, Shijie Geng, Renrui Zhang, Peng Gao, Gerard de Melo, Xiaogang Wang, Jifeng Dai, Yu Qiao, and Hongsheng Li. Frozen CLIP models are efficient video learners. In *European Conference on Computer Vision*, pages 388–404. Springer, 2022. 1

[19] Jie Liu, Yixiao Zhang, Jie-Neng Chen, Junfei Xiao, Yongyi Lu, Bennett A Landman, Yixuan Yuan, Alan Yuille, Yucheng Tang, and Zongwei Zhou. Clip-driven universal model for organ segmentation and tumor detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 21152–21164, 2023. 1

[20] Yuejiang Liu, Parth Kothari, Bastien van Delft, Baptiste Bellot-Gurlet, Taylor Mordan, and Alexandre Alahi. TTT++: When does self-supervised test-time training fail or thrive? *Neural Information Processing Systems (NeurIPS)*, 2021. 2

[21] Zachary Nado, Shreyas Padhy, D. Sculley, Alexander D'Amour, Balaji Lakshminarayanan, and Jasper Snoek. Evaluating prediction-time batch normalization for robustness under covariate shift. *arXiv:2006.10963 [cs, stat]*, Jan. 2021. arXiv: 2006.10963. 2

[22] David Osowiechi, Gustavo A. Vargas Hakim, Mehrdad Noori, Milad Cheraghalikhani, Ismail Ayed, and Christian Desrosiers. TTTFlow: Unsupervised test-time training with normalizing flow. In *2023 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 2125–2126, Los Alamitos, CA, USA, jan 2023. IEEE Computer Society. 2

[23] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 1, 2

[24] Glenn Shafer and Vladimir Vovk. A tutorial on conformal prediction. *Journal of Machine Learning Research*, 9(3), 2008. 2

[25] 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. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural*

Information Processing Systems, volume 35, pages 14274–14289. Curran Associates, Inc., 2022. 1, 2, 4, 6

[26] Yu Sun, Xiaolong Wang, Zhuang Liu, John Miller, Alexei A. Efros, and Moritz Hardt. Test-time training with self-supervision for generalization under distribution shifts. In *International Conference on Machine Learning (ICML)*, 2020. 2

[27] Gustavo A Vargas Hakim, David Osowiechi, Mehrdad Noori, Milad Cheraghalikhani, Ali Bahri, Ismail Ben Ayed, and Christian Desrosiers. ClusT3: Information invariant test-time training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6136–6145, 2023. 2, 5

[28] Dequan Wang, Evan Shelhamer, Shaoteng Liu, Bruno Olshausen, and Trevor Darrell. Tent: Fully test-time adaptation by entropy minimization. 2021. 2, 5

[29] 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, 3

[30] Qilong Wu, Xiangyu Yue, and Alberto Sangiovanni-Vincentelli. Domain-agnostic test-time adaptation by prototypical training with auxiliary data. In *NeurIPS 2021 Workshop on Distribution Shifts: Connecting Methods and Applications*, 2021. 2, 3

[31] Marvin Zhang, Sergey Levine, and Chelsea Finn. Memo: Test time robustness via adaptation and augmentation. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 38629–38642. Curran Associates, Inc., 2022. 2

[32] Renrui Zhang, Rongyao Fang, Peng Gao, Wei Zhang, Kunchang Li, Jifeng Dai, Yu Qiao, and Hongsheng Li. Tip-adapter: Training-free CLIP-adapter for better vision-language modeling. *arXiv preprint arXiv:2111.03930*, 2021. 2

[33] Yifei Zhou, Juntao Ren, Fengyu Li, Ramin Zabih, and Ser Nam Lim. Test-time distribution normalization for contrastively learned visual-language models. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems*, volume 36, pages 47105–47123. Curran Associates, Inc., 2023. 2, 6