Prompt-Based Memory Bank for Continual Test-Time Domain Adaptation in Vision-Language Models

1st Ran Wang
Australian Artificial Intelligence Institute
FEIT, University of Technology Sydney
Sydney, NSW, Australia
ran.wang-2@student.uts.edu.au

2nd Hua Zuo*
Australian Artificial Intelligence Institute
FEIT, University of Technology Sydney
Sydney, NSW, Australia
hua.zuo@uts.edu.au

3rd Zhen Fang
Australian Artificial Intelligence Institute
FEIT, University of Technology Sydney
Sydney, NSW, Australia
zhen.fang@uts.edu.au

4th Jie Lu
Australian Artificial Intelligence Institute
FEIT, University of Technology Sydney
Sydney, NSW, Australia
jie.lu@uts.edu.au

***Abstract*—In dynamic environments, the generalization capabilities of large-scale vision language models tend to decline. This is attributed to the evolving distribution of target domains over time, leading to misalignment between image and text pairings, affecting the model's performance. Addressing this, Test-Time Adaptation (TTA) has been proposed to adapt pre-trained source models to these changing target domains during testing phases. However, traditional TTA approaches, which are designed for a single changing scenario and mainly depend on self-training and entropy minimization, are easily affected by extreme and novel samples in long-term environments, leading to error accumulation and catastrophic forgetting. Although previous Continual Test-Time Adaptation (Continual TTA) methods based on the teacher-student framework can effectively address long-term adaptation issues, they are not feasible for large-scale vision language models due to their high memory requirements. To overcome these challenges, we introduce a novel approach: Prompt-based memory bank for Continual Test-Time Adaptation (PCoTTA). PCoTTA uniquely freezes the CLIP image and text encoders, focusing on updating and storing trainable prompts, significantly reducing memory usage. By implementing a stable pseudo-label strategy and high gradient sensitivity updating, PCoTTA effectively learns new knowledge. In long-term dynamically changing environments, PCoTTA demonstrates high stability and accuracy and achieves a good balance between learning new information and retaining existing knowledge, significantly enhancing the adaptability and generalization capabilities of the CLIP model. Through extensive experimental comparisons, PCoTTA surpasses the current state-of-the-art methods, achieving an average 2% improvement in accuracy for both test-time adaptation and continual test-time adaptation tasks.**

***Index Terms*—Transfer Learning, Vision-Language Model, Test-Time Domain Adaptation**

*Corresponding author. This work is supported by the Australian Research Council under Discovery Early Career Researcher Award DE220101075.

I. Introduction

Large-scale vision-language models [1, 2], such as Contrastive Language-Image Pre-training model (CLIP [3]), have demonstrated superior performance in zero-shot image classification tasks due to their text-image pairing capabilities [4]. Particularly, the incorporation of prompts into the text encoder, like *"a photo of a"* enables these large-scale vision-language models to swiftly adapt to various specific downstream tasks [5]. However, in practical testing scenarios, environments can undergo continual changes, such as in autonomous driving where weather conditions may shift from clear to rainy or snowy, or increased noise due to sensor aging over time [6, 7]. These factors can lead to temporal variations in the distribution of the target domain, which may in turn cause misalignment between text and image pairings, impacting the model's generalization performance [8].

To address this issue, Test-Time Adaptation (TTA) methods have been developed, aiming to utilize the source pre-trained model directly during the testing phase to adapt to these changing target domains [9]. TTA methods are designed to enhance the model's generalization performance on zero-shot tasks without retraining. However, it's important to note that traditional TTA approaches, primarily relying on self-training [10] and entropy minimization [11], are generally designed for short-term or single-scene adaptations rather than continuous, long-term adaption. As such, in dynamic, long-term environments, they often face limitations. They are prone to pseudo-label instability, which can lead to an accumulation of errors in extreme samples and catastrophic forgetting of initial knowledge when learning new samples [12, 13].

To mitigate these long-term adaptation issues, Continual Test-Time Adaptation (Continual TTA) aims to enable models

to continuously adapt to new data during testing without accessing the source data, enhancing their performance and accuracy in dynamic or constantly changing long-term environments [12]. Recent methods typically employ a teacher-student model [12, 13], addressing catastrophic forgetting and error accumulation through model averaging updates. However, this approach, requiring the simultaneous loading of two models, is unsuitable for large-scale language models. It significantly increases the memory burden, making it impractical in situations where high resource efficiency is required.

Accordingly, we introduce a novel strategy, **P**rompt-based memory bank for **C**ontinual **T**est-**T**ime **A**daptation (PCoTTA). PCoTTA contains two main components: the dynamic prompt-based memory bank and maximum gradient search. PCoTTA utilizes a frozen CLIP model and focuses on dynamically storing both new and old trainable prompts in a memory bank, reducing the need to load two separate models, such as teacher and student models, and consequently saving memory usage. It generates pseudo-labels using more stable entropy values and employs high gradient sensitivity for prompt updates. This innovative approach significantly reduces the memory requirements typically associated with large-scale vision language models, especially in scenarios where the test distribution is continually changing. Given that updates during testing can utilize pseudo-labels and the model's inherent capabilities, employing high gradient sensitivity facilitates rapid learning of knowledge beneficial for updates. This method is both efficient and swift in enhancing the model's generalization capabilities.

Specifically, we select the prompts most affected by the maximal gradients induced by new samples in the network. This selection process prioritizes model updates on samples that are most likely to disrupt the existing knowledge structure, effectively preserving old knowledge while acquiring new information. These updated prompts are retained in the memory bank for future reuse, thus enhancing the model's memory of past knowledge. Additionally, we reduce error accumulation through a weighted prediction mechanism and maintain initial prompts to decrease catastrophic forgetting, ensuring the model's stability in dynamic environments.

We demonstrate the efficacy of our proposed method in three image classification tasks of test-time adaptation and continual test-time adaptation, where PCoTTA outperforms the state-of-the-art methods by an average of 2% in accuracy. In these tasks, PCoTTA not only shows significant improvements in continual test-time adaptation over existing methods but also effectively mitigates the impact of catastrophic forgetting and error accumulation.

Our contributions are summarized as follows:

- We introduce a prompt-based memory bank approach that dynamically manages trainable prompts to efficiently adapt large-scale vision-language models to changing environments, effectively reducing memory usage.
- Our strategy employs entropy-based pseudo-labeling and model sensitivity to gradient when learning new knowledge for robust and targeted updates which mitigate error accumulation and catastrophic forgetting.
- Our approach is simple and easy to implement. To the best of our knowledge, we are the first to propose improving the generalization of large-scale vision-language models in a continuous testing environment.

II. Related Works

In this section, we introduce previous research on prompt tuning in vision-language models and continual test-time domain adaptation methods.

A. Prompt Tuning in Vision-Language Models

Vision-Language Models (VLMs) are designed to integrate visual and textual data through extensive pre-training, establishing cross-modal links [1, 3, 14]. These models typically consist of image and text encoders, utilizing self-supervised learning techniques such as contrastive loss to enhance efficiency. For instance, in VLMs like CLIP, by introducing text encoders, multimodal image classification methods using text-image pairs are employed, applicable to various zero-shot downstream tasks [4, 15–17]. Recent studies have started employing prompt learning techniques as an alternative to comprehensive fine-tuning, enabling VLMs to adapt more effectively to specific applications while reducing the number of parameters. Prompt learning, originating from Natural Language Processing (NLP), is an effective method to tailor large-scale models for specific tasks without retraining model parameters, using textual prompts [5, 18, 19]. In pre-trained language models, these prompts typically appear as completion sentences or masked sentences, such as "This is [MASK]", with the model being trained to predict suitable words for the masked positions [20–22]. In VLMs, these predictions often involve image classification tasks [3].

To avoid the tedious task of manually setting prompts, current research employs learnable prompts, allowing the model to autonomously learn appropriate prompts. For example, the COOP [23] method introduced learnable prompts, primarily focusing on specific tasks, which may lead to model overfitting. Building on this, the CoCoOP [24] method incorporates meta-learning, enhancing CLIP's generalization by integrating image feature biases. However, these methods involve retraining CLIP, not fully exploiting its zero-shot learning potential. Therefore, our strategy is to avoid retraining and directly enhance CLIP's zero-shot generalization capabilities during the testing phase.

B. Continual Test-Time Domain Adaptation

Test-time adaptation involves updating the model during the testing phase without accessing source domain data, setting it apart from domain adaptation and generalization [9, 25]. Existing test-time adaptation methods include entropy minimization [11, 26, 27] (enhancing model certainty by reducing test entropy), batch normalization [28] (adjusting normalized data during testing for model stability), and pseudo-labeling [10, 29] (updating the model by assigning labels to unlabeled test data). However, these techniques are mainly suitable for static target domains. When the target domain

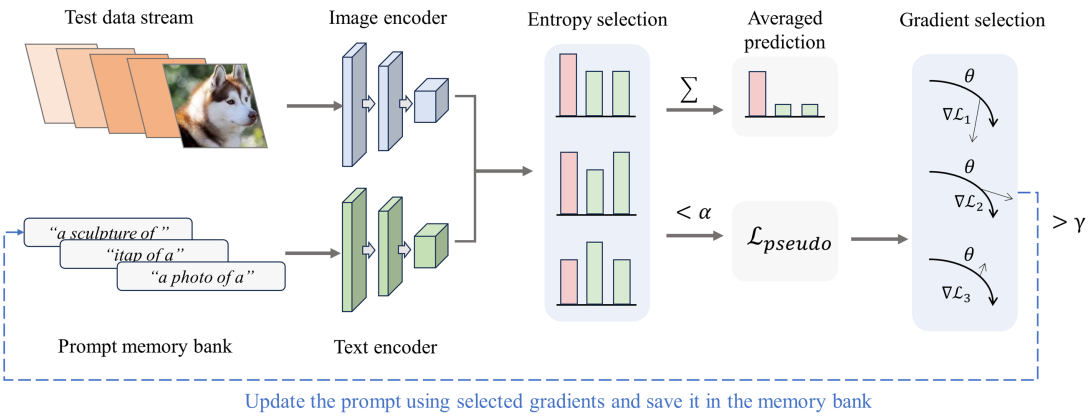

Fig. 1. **Overview of our PCoTTA framework.**

changes dynamically over time, they may bias towards extreme samples, leading to error accumulation and reduced model prediction capabilities, or focus excessively on new samples, causing catastrophic forgetting and model overfitting, decreasing its generalization ability [12, 30].

For continual test-time adaptation, CoTTA [12] first introduced a dynamic teacher-student model updating approach. It generates more reliable pseudo-labels through weighted teacher model predictions, smoothly updates the teacher model using student model parameters over a period, and randomly restores weights to the original model. This allows for continuous adaptation and stability across various environments. Similarly, NOTE [31] introduced instance-aware batch normalization to enhance the performance of test samples and address out-of-distribution sample issues. The RMT [13] method employs an average teacher-student model for robust predictions. However, these methods update both models as teacher and student models, which is not feasible for large-scale vision-language models due to increased memory usage and computational complexity. Therefore, we maintain a dynamic prompt-based memory bank, ensuring model stability without increasing memory usage.

III. Preliminaries

In this section, we first introduce the structure of CLIP and then the concept of learnable prompts.

Contrastive Language-Image Pre-training (CLIP), integrates an image encoder based on ResNet [32] or ViT [33] and a text encoder using a transformer architecture [34]. It aligns images and texts closely using contrastive loss. For specific image classification tasks, CLIP employs textual prompts, such as "a photo of a [CLASS]," to guide the text encoder towards generating semantically relevant representations corresponding to input images. The inference in CLIP is based on the probability of an input image belonging to a certain class, calculated by the similarity between the class embedding and the image feature. The process can be summarized as:

$$P(y=i|\mathbf{x}) = \frac{\exp(\text{similarity}(w_i, f(\theta))/\tau)}{\sum_{j=1}^{K} \exp(\text{similarity}(w_j, f(\theta))/\tau)} \quad (1)$$

where w_i is class text embedding for class i, $f(\theta)$ is the image feature extracted by the image encoder and K is the total number of classes.

The CoOp [23] framework, enhancing CLIP, introduces learnable prompts while keeping other encoder parameters static. It designs the prompts as a sequence of learnable vectors $\mathbf{t} = [\mathbf{V}_1][\mathbf{V}_2]\dots[\mathbf{V}_M][\text{CLASS}]$, each aligning with the word embedding dimensionality in CLIP. The text encoder then processes these prompts to produce a classification vector, which is used to compute the likelihood of each class prediction:

$$P(y=i|\mathbf{x}) = \frac{\exp(\text{similarity}(g(t_i), f(\theta))/\tau)}{\sum_{j=1}^{K} \exp(\text{similarity}(g(t_j), f(\theta))/\tau)} \quad (2)$$

where $g(t_i)$ is the text feature of the text encoder when processing the learnable prompt t_i. In this formula, each class-specific prompt t_i is incorporated with the respective word embedding for the i-th class, facilitating the alignment of text and image representations.

IV. Methodology

In this section, we first introduce the problem definition of continual test-time adaption in CLIP and provide an overview of PCoTTA. Then, we detail the concepts of maximum gradient search and the dynamic prompt-based memory bank.

A. Problem Definition

In this study, we focus on the CLIP model for continual test-time adaptation (continual TTA). In our approach, each text prompt t_i is treated as a set of trainable parameters, optimized to enhance prediction accuracy for a specific test image $\mathbf{x}$. The CLIP model M, with its image encoder parameters θ frozen, computes the feature representation $f(\mathbf{x}.\theta)$ of the image, while the embeddings for each trainable text prompt t_i are generated by the model's frozen text encoder as $g(t_i)$.

The core challenge of continual TTA in CLIP involves adapting a pre-trained model M, equipped with learnable prompts t_i, to continuously incoming streams of unlabeled data from a target domain (denoted as x^T) without revisiting the original training data or undergoing extensive retraining.

This target data, derived in batches from the evolving dataset S^T, is processed sequentially ($x_t^T \rightarrow x_{t+1}^T$), reflecting the temporal progression of the dataset ($S_t^T \rightarrow S_{t+1}^T$). At each time step t, the prompt t_i is adapted to make predictions on the current batch x^T, progressively tuning the model to the nuances of the data stream until it achieves optimal adaptation to the continually evolving domain.

The overview of our approach is specified in the Figure 1. Initially, a test data stream with continually changing distributions is sequentially inputted into the CLIP image encoder, yielding image features. Concurrently, prompts from the memory bank are used in the text encoder to obtain text features. These text features, along with the image features, are then utilized to generate multiple entropy values. Subsequently, those with entropy values below a threshold α are selected as pseudo-labels. This process is followed by maximal gradient selection, where prompts with gradient values over a threshold γ are chosen to update the prompts, which are then stored in the memory bank. Finally, the output is derived from a weighted prediction based on all the prompts through averaged prediction. Leveraging a memory bank and maximum gradient search, PCoTTA maintains stability in dynamic environments and rapidly acquires new knowledge to enhance its generalization capabilities.

Next, we will detail the maximum gradient search and the dynamic prompt-based memory bank. The maximum gradient search serves to make the model learn new samples that are more distinct from old knowledge, quickly adapting to new distribution changes. The dynamic prompt-based memory bank is designed to maintain both old and new prompts, enabling the model to mitigate catastrophic forgetting and the accumulation of errors.

B. Maximum Gradient Search

In online continual learning, there is a method aimed at utilizing past task samples stored in a memory bank to reduce catastrophic forgetting. Inspired by this, to save memory usage, instead of storing samples, we store prompts. Additionally, Gu et al. [35] has shown that samples which generate gradients in the network that are most likely to be interfered with by incoming new samples are beneficial for updating the neural network based on backpropagation of gradients. Therefore, in our prompt-based memory bank, in order to obtain the new knowledge effectively, we select prompts with the largest gradients for prediction.

We first filter the entropy values below the threshold as pseudo-labels for simulated updates. For each prompt t_i in the memory bank $T = \{t_1, t_2, \ldots, t_N\}$, the CLIP model predicts the conditional probability for each category i using the similarity between the text prompt embedding and the image feature representation. This process is formulated:

$$\mathbf{P} = \{P(y = i|\mathbf{x})\}_{i\in\{1,\ldots,N\}} \tag{3}$$

Subsequently, the entropy of the prediction distribution for each prompt t_i is calculated, and the prompt prediction yielding an entropy lower than the threshold is chosen as the pseudo-label:

$$\hat{t} = \arg\min_i H(P(y|\mathbf{x}, t_i)) < \alpha \tag{4}$$

where $\mathcal{H}$ denotes the entropy function, and α is the predefined entropy threshold.

For each text prompt t_i, we use an cross-entropy loss function $\mathcal{L}$ that measures the discrepancy between the predictions and the pseudo-label, and calculate the gradient of the loss with respect to the embedding vector of t_i, G_{t_i}:

$$\mathbf{G}_{t_i} = \nabla_{t_i}\mathcal{L}(p(y = i|\mathbf{x}), y_{\text{pseudo}}) \tag{5}$$

We select the text prompt t_i that causes the maximum gradient change, and update t_i using gradient ascent:

$$\hat{t} = t_i \,|\, i = \arg\max_i \|\mathbf{G}_{t_i}\|_2 \tag{6}$$

$$t_i \leftarrow t_i + \eta \cdot \mathbf{G}_{t_{\hat{i}}} \tag{7}$$

where $\mathbf{G}_{t_{\hat{i}}}$ is the gradient corresponding to the selected text prompt $\hat{t}$, and η is the learning rate. This approach allows us to update the text prompts directly to maximize the gradient response of the model output, enhancing the predictive performance for specific tasks. By doing so, we can adjust the text prompts to better align with the image features, thus improving classification accuracy.

The final prediction $\mathbf{P}_{\text{final}}$ is then determined by combining the predictions from all prompts in the memory bank:

$$\mathbf{P}_{\text{final}} = \sum_{i=1}^{N} w_i \cdot \mathbf{P}_i \tag{8}$$

The final prediction is a synthesized outcome that combines diverse viewpoints and prioritizes them based on their assessed confidence. This methodology enhances the robustness and accuracy of the prediction by leveraging the collective intelligence embedded in the memory bank's prompts.

C. Dynamic Prompt-based Memory Bank

In order to make the prompt-based memory bank learn new knowledge while preserving old knowledge, we fix the initial prompt and dynamically update the memory bank for later prompts. We rank all prompts according to the L2 norm of their gradients $\|\mathbf{G}_{t_i}\|_2$ to identify the prompts that pose the greatest challenge to the model's current parameters. Based on the gradient magnitudes, we select the top K prompts with the largest $\|\mathbf{G}_{t_i}\|_2$. We set a threshold γ, such that a prompt t_i remains in the bank only if $\|\mathbf{G}_{t_i}\|_2 > \gamma$. This process can be formalized as follows:

$$T_{\text{updated}} = \{t_i | t_i \in T, \|\mathbf{G}_{t_i}\|_2 > \gamma\}_{\text{top } K} \tag{9}$$

After each adaption, we replace the existing memory bank with the newly selected set of prompts T_{updated}. This ensures that the memory bank always contains prompts that can guide the model to learn and adapt. Through this method of dynamically maintaining the memory bank, we ensure that the model is continually presented with new challenges, and by

Algorithm 1 Prompt-based Memory Bank for Continual Test-Time Domain Adaptation

Input: Memory bank T of text prompts $\{t_1, t_2, \ldots, t_N\}$, CLIP model M, image feature representation function $f(\cdot)$, text prompt embedding function $g(\cdot)$, threshold entropy α, gradient norm threshold γ.

Initialize the memory bank T with a set of text prompts.
for each batch of target domain images x_t^T **do**
  for each text prompt $t_i \in T$ **do**
    Compute the conditional probabilities $\mathbf{p}_i$ using $f(\mathbf{x}.\theta)$ and $g(t_i)$.
    Calculate entropy $\mathcal{H}(\mathbf{p}_i)$ for each prompt.
    if $\mathcal{H}(\mathbf{p}_i) < \alpha$ **then**
      Select t_i as pseudo-label $\hat{t}$. eq.(4)
      Compute the gradient $\mathbf{G}_{t_i}$ with respect to t_i. eq.(5)
    end if
  end for
  Select top K prompts with the largest gradient norms that exceed γ. Replace existing T with the updated set of prompts T_{updated}. eq.(9)
  Combine all prompt predictions $\mathbf{P}_i$ to obtain the final prediction $\mathbf{P}_{\text{final}}$. eq.(8)
end for

learning from these challenges, it enhances its generalization capabilities and adaptability. This process helps the model perform well not only on the current task but also better transfer to new tasks. We provide a detailed pseudo-code of PCoTTA in Algorithm 1.

V. Experiments and Results

In this section, we first introduce the experimental settings, which includes detailed descriptions of benchmarks, baselines, datasets, and implementation details. This is followed by an analysis of the experimental results and ablation studies.

A. Experiment Settings

Benchmarks. We conduct the assessments across three benchmarks: test-time adaptation [9], continual test-time adaptation [12], and long-term adaptation [12]. In all benchmarks, we update the model only during the testing phase without retraining or accessing source data. For the test-time adaptation setting, we evaluate each model on each individual distribution separately, resetting after each evaluation without involving continuing. For the continual test-time adaptation setting, we continuously test across 15 scenarios from the first Gaussian to the last Pixel, without resets, continuously adapting to the entire distribution. For the long-term adaptation, we continuously test 3 conditions for 5 rounds. Additionally, we present results from ablation studies that analyze the impact of each component on the model.

Baselines. We compare the PCoTTA against classic test-time adaptation methods in test-time adaptation, continual test-time adaptation and long-term adaptation settings. The methods include TENT [9], which employs test entropy minimization. Pseudo-Label [10], which updates the model using generated pseudo-labels. and TPT [8], the first method to integrate CLIP prompt tuning for test-time adaptation, minimizing marginal entropy through image augmentation. Each method is followed by the notation "Cont. TTA" to denote the continual test-time adaptation benchmark, while those without it refer to the standard TTA benchmark. The methodologies for TTA and Cont. TTA are consistent, differing only in whether the adaptation is continuous. TPT, due to its reset after adapting to each image, cannot be compared in the Continual TTA setting. This comparison allows us to observe potential catastrophic forgetting or error accumulation during continual test-time adaptation. This also helps to confirm that PCoTTA, in the process of long-term adaptation, not only effectively reduces these issues, but also achieves significant improvement through the preserving of both new and old knowledge.

Datasets. In the context of test time adaptation and its continuous variant, CIFAR10-C [6], CIFAR100-C [12] and ImageNet-C [7] serve as the primary datasets for TTA and Continual TTA, particularly for assessing corruption robustness [9, 12, 26]. CIFAR10-C and CIFAR100-C each comprise 10,000 test images spanning 10 and 100 categories respectively, while ImageNet-C encompasses 50,000 test images across 1,000 categories. These datasets feature 15 types of corruptions, categorized into noise (like gaussian, shot, impulse), blur (including defocus, glass, motion, zoom), weather (such as snow, frost, fog), and digital (covering brightness, contrast, elastic trans, pixelate, jpeg). Our evaluation employs these datasets at damage level 1.

Implementation Details. In our experiments, we adhere to the settings of TTA and Continual TTA and the implementation details of CLIP, TENT, Pseudo-label, and TPT for the evaluation of these baselines. For our method, PCoTTA, we initiate with the 4-token prompt "a photo of a" and selected samples with entropy less than 0.2 for pseudo-label, setting $\gamma = 10$. The batch size is set to 100. The length of memory bank is 5 prompts. Results are evaluated on CLIP with ViT-B/16. And in the final prediction we use the average prediction with the same weights. The learning rate is 0.0025.

B. Results

All tests are conducted using the ViT-B/16 model. To facilitate the observation of catastrophic forgetting and error accumulation, we place the results of TTA and continual TTA in the same Table I. Here, "(Cont. TTA)" represents the continual adaptation from Gaussian to Pixelate. All methods, except for CLIP and TPT, have been compared for both continual and non-continual scenarios. If the accuracy of continual TTA falls more than 2% below that of TTA, it is considered indicative of potential catastrophic forgetting or error accumulation, and is marked with an asterisk (*). Since previous methods could not self-correct for error accumulation or catastrophic forgetting, if the accuracy of continual TTA drops below 2%, we reset the model and continue with continual adaptation, denoted with two asterisks (**).

TABLE I

ACCURACY(%) FOR THREE STANDARD ONLINE CONTINUAL TEST-TIME ADAPTATION TASKS WHICH ARE **CIFAR10-C**,**CIFAR100-C**,**IMAGENET-C**. EACH METHOD IS EVALUATED ON BOTH TTA AND CONTINUAL TTA TESTING. THE BOLD NUMBER INDICATES **BEST** RESULT.

			$t \longrightarrow$														
Dataset	Method	Continual?	Gauss.	Shot	Impul.	Defo.	Glass	Motion	Zoom	Snow	Frost	Fog	Bright.	Contr.	Elas.	Jpeg	Pixel.	Average
CIFAR10-C	CLIP	✗	70.3	72.1	68.5	74.1	48.1	68.2	66.3	71.8	72.7	72.8	75.3	73.1	65.2	76.4	71.1	69.7
	TPT	✗	69.7	71.7	69.4	71.1	47.7	62.2	60.7	71.1	70.3	70.6	72.7	69.8	62.7	74.6	70.3	67.6
	TENT	✗	69.9	71.8	68.2	73.5	48	67.8	65.2	71.4	71.9	72.4	74.9	72.5	64.7	75.9	70.4	69.2
	TENT (Cont. TTA)	✓	69.9	70.1*	67.6	72.9	46.8	65.7*	64.4	69.9*	70.9	71.1	74	70.8*	64.7	75.3	70	68.3
	Pseudolabel	✗	70.4	72.1	68	73.8	48	68.4	65.7	71.5	72.2	72.7	75.2	72.8	65.1	76.1	70.8	69.5
	Pseudolabel (Cont. TTA)	✓	70.4	70.7*	67.9	73	46.1*	68	64.8	71.4	72	71.9	75	71.5	64.5	75.6	70	68.9
	PCoTTA	✗	71.8	73.3	70.2	74.4	48.3	69.1	66.2	72.2	72.8	73.3	76.2	73.5	65.7	77.3	71.5	70.4
	PCoTTA (Cont. TTA)	✓	**71.9**	**74.7**	**71.5**	**75.7**	**49.1**	**71.2**	**68.2**	**73**	**73.6**	**73.9**	**77.1**	**73.9**	**66.3**	**78.8**	**72.3**	**71.4**
CIFAR100-C	CLIP	✗	43	44.8	40.2	46	24.3	40.6	39.5	43.3	44.5	44.8	47.7	44.1	37.4	48.9	43.6	42.2
	TPT	✗	41.3	42.8	39.4	42.3	22.3	37.1	35.7	41.8	41.8	42.6	43.7	40.9	37.4	45.1	41.5	39.7
	TENT	✗	43.4	45	40.8	46.4	24.2	40.9	39.8	43.8	45	45	48.2	44.1	37.5	49.4	43.9	42.5
	TENT (Cont. TTA)	✓	43.4	41.8*	37.9*	40.4*	14.6*	29.9*	39.8**	40.2	41.6	36.6	25.5	44.1**	34.2	45.7	36.9	36.8
	Pseudolabel	✗	43.3	45.1	40.6	46.5	24.6	40.9	39.8	43.6	44.9	45.1	48.1	44.4	37.7	49.3	44.1	42.5
	Pseudolabel (Cont. TTA)	✓	43.3	39.8*	36*	41.3*	20.3*	37*	36.1*	39.3	40.9	41.5	44.9	41.2	34.9	45.9	40.8	38.9
	PCoTTA	✗	43.6	45.5	41.2	46.6	24.7	41	40.3	44.1	45.5	45.2	48.6	44.4	37.7	49.9	44.3	42.8
	PCoTTA (Cont. TTA)	✓	**44**	**45.8**	**41.5**	**47.3**	**25.4**	**41.3**	**40.4**	**44.6**	**45.8**	**45.2**	**49.1**	**45**	**38.1**	**50.4**	**44.5**	**43.2**
ImageNet-C	CLIP	✗	57.6	57.2	51.2	61.6	53.9	60.2	48.9	53.4	55.3	58.7	64.5	61.7	58.5	57.6	59	57.3
	TPT	✗	57.3	57.1	51.7	56.5	54.5	60.5	49.4	55.2	56.8	59	65.3	62.3	59.8	58.8	60.1	57.6
	TENT	✗	58.6	58.2	51.7	41.2	54.7	61.1	49.2	54.5	56.2	34.1	65.7	42.1	59.3	59.1	60.1	53.7
	TENT (Cont. TTA)	✓	58.6	58.7	37*	42.1**	52.4	60.5	24.2*	50.1*	37.8*	34.1**	63.9	42.1	42.3*	58.8	38.7*	46.8
	Pseudolabel	✗	58	57.4	51.4	61.9	54.2	60.5	49.1	53.9	55.5	58.9	64.8	61.9	58.9	58	59.4	57.6
	Pseudolabell (Cont. TTA)	✓	58	53.7*	49.1	59.8*	52.5*	59.2	48.4	53.5	55.5	57.8	63.9	60.6	57.1*	57.4	58.1	56.3
	PCoTTA	✗	58.8	58.4	52.1	62.4	54.6	61.4	49.4	54.9	56.5	59.2	66	62.4	59.6	59.3	60.5	58.4
	PCoTTA (Cont. TTA)	✓	**59.1**	**59.4**	**52.6**	**63.1**	**55.7**	**61.8**	**50.7**	**55.7**	**57.9**	**60.1**	**66.9**	**63.1**	**60.6**	**60.5**	**61.4**	**59.2**

* indicates that a difference in accuracy $> 2\%$ may lead to catastrophic forgetting or error accumulation

** indicates that the accuracy $< 2\%$ and resets the model to re-adapt continuously

As shown in Table I, starting with the CIFAR10-C dataset, TPT shows poor generalization performance. This may be due to its inability to utilize inter-image relational information, as each reset is akin to starting anew without accumulating knowledge from previous scenarios. The TENT method, employing entropy minimization strategy, already exhibits error accumulation during single TTA tests, leading to reduced accuracy. This accumulation effect is more pronounced in Continual TTA testing, particularly in scenarios like shot and motion, where accuracy drops by more than 2%. This could be due to entropy minimization causing the model to be overly confident in its current predictions, which may lead to ignoring the true labels and accumulating more errors when faced with continuously changing scenarios. The Pseudo-labeling method, while stable in TTA testing, encounters error accumulation issues in Continual TTA as well. For example, in the glass scenario of CIFAR10-C, the accuracy for pseudo-labeling is 48%, but it drops to 46.1% in Continual TTA testing, suggesting that the model's reliance on its predictions may lead to accumulating errors. In contrast, our PCoTTA method exhibits higher robustness and stability.

On the more complex CIFAR100-C dataset, TENT experiences more severe error accumulation and catastrophic forgetting, with accuracy dropping to less than 2% in the zoom scenario, indicating serious catastrophic forgetting. This may be due to the increased vulnerability of the entropy minimization strategy without proper regularization measures as the types of disturbances increase, leading the model to discard knowledge of old scenes when adapting to new ones. Moreover, the negative impact of error accumulation is more pronounced due to the larger number of categories.

On the most challenging ImageNet-C dataset, despite TPT's ability to handle more complex situations through image enhancement techniques, the performance of TENT and pseudo-labeling methods further declines due to accumulated errors in continual adaptation. This underscores the importance of continual adaptation ability when dealing with large-scale and complex data. In contrast, our PCoTTA method maintains a high accuracy of 59.2% in the Continual TTA testing of ImageNet-C, showing high stability and robustness.

Long-Term Adaptation In Table II, we conducted five continual rounds of testing for the three conditions of weather (snow,frost,fog), in order to simulate continuous prolonged adaptation in realistic scenarios. The results show that the

TABLE II

WE PERFORM FIVE ROUNDS CONTINUAL EVALUATIONS OF THREE TEST CONDITIONS FOR WEATHER IN IMAGENET-C TO EVALUATE LONG-TERM ADAPTIVE PERFORMANCE. NUMBERS ARE ACCUARCY(%). THE BOLD NUMBER INDICATES **BEST** RESULT.

t ——————————————————————————→

Round	1			2			3			4			5			Mean
Condition	Snow	Frost	Fog	Snow	Frost	Fog	Snow	Fog	Frost	Snow	Frost	Fog	Snow	Frost	Fog	
CLIP	53.4	55.3	58.7	53.4	55.3	58.7	53.4	58.7	55.3	53.4	55.3	58.7	53.4	55.3	58.7	55.8
TENT (Cont. TTA)	54.5	55.2	58.2	52.4	54.4	56	48.8	53.5	54.4	44.5	47.4	49.6	32.3	42.5	45.4	49.9
Pseudolabel (Cont. TTA)	53.9	52.6	56.2	52.5	55	57.7	49.5	54.1	55	45.6	47.7	49.8	35.5	43.4	49.8	50.6
TPT	**55.2**	56.8	59	55.2	56.8	59	55.2	59	56.8	55.2	56.8	59	55.2	56.8	59	57
PCoTTA (Cont. TTA)	54.9	**56.9**	**60.1**	**55.9**	**57.1**	**60.4**	**56.1**	**57.4**	**60.8**	**56.1**	**57.5**	**61**	**56.3**	**58**	**61.4**	**58**

TABLE III

EFFECTS OF THE PCOTTA COMPONENTS. RESULTS ARE AVERAGED OVER THREE DATASETS

Method	CIFAR10-C	CIFAR100-C	ImageNet-C	Average
CLIP	69.7	42.2	57.3	56.4
+Maximum Gradient Search (MGS)	70.1	42.6	58.1	56.9
+Prompt Memory Bank (PMB)	70.8	42.8	58.5	57.4
+MGS +PMB	**71.4**	**43.2**	**59.2**	**57.9**

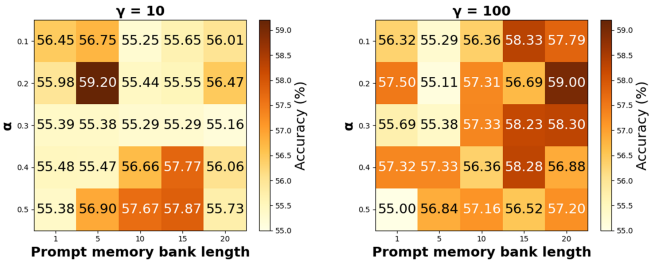

Fig. 2. Evaluation of PCoTTA selecting hyperparameters which includes entropy thershold α, gradient thershold γ and prompt-based memory bank length on ImageNet-C. Numbers are accuarcy(%)

TENT and pseudo-labeling methods can be stable initially, but at the beginning of 3 rounds, the prolonged adaptation leads to the accumulation of errors and reduces the accuracy. Instead, PCoTTA learns new knowledge during long time adaptation.

Overall, our PCoTTA method enhances CLIP's generalization capabilities across multiple scenarios and maintains stability during continual adaptation. By utilizing a memory bank for averaging predictions, PCoTTA retains knowledge from previous scenarios and accumulates new knowledge while adapting to new ones, avoiding catastrophic forgetting.

C. Ablation Study

In Table III, we compare the effects of Maximum Gradient Search (MGS) and Prompt Memory Bank (PMB) within the PCoTTA framework. The MGS likely enhances the model's focus on critical features that are essential during distribution shifts, while the PMB diversifies the model's memory, enabling adaptation to evolving data patterns. Together, they significantly increase the model's robustness to the challenges of continual distribution shifts.

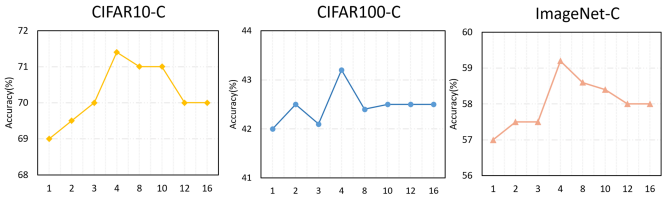

Fig. 3. Ablation experiment for prompt length of 1, 2, 3, 4, 8, 10, 12, 16 on three datasets CIFAR10-C, CIFAR100C and ImageNet-C. Numbers are accuarcy(%)

In our ablation study illustrated in Figure 2, we investigate the impact of various hyperparameters. These parameters include the entropy threshold α, the gradient threshold γ, and the length of the prompt-based memory bank. In the heatmap, the accuracy of the model peaks when the γ value is 10, with a prompt-based memory bank length of 5 and an entropy of 0.2. However, when the length of the prompt-based memory bank is increased to 20, the accuracy decreases significantly, suggesting that an increase in memory span does not imply an increase in performance. Increasing the γ to 100 slightly decreases the overall accuracy, demonstrating that the gradient threshold has an impact on performance.

The line graph in Figure 3 compares the effect of prompt length, with the highest accuracy when prompt length = 4, while longer prompt lengths instead decrease performance, suggesting the potential for overfitting when extending the length of the memory bank.

VI. CONCLUSION AND FUTURE WORK

In this study, we introduce a new Prompt-based Memory Bank method (PCoTTA) that requires no additional memory, specifically designed for large-scale vision-language models. The primary goal is to enhance the model's robustness against catastrophic forgetting and error accumulation during continual test-time updates, especially when dealing with changes in data distribution. This memory bank approach allows for smooth updates without the need to load complex teacher-student models and can rapidly learn new knowledge through maximized gradient searching. Our PCoTTA method significantly improves the CLIP model's zero-shot generalization performance in the face of real-world distribution shifts and

effectively mitigates issues of catastrophic forgetting and error accumulation. However, the current research is only focused on independent distribution shifts. In the future, we plan to explore the effects of continuous distribution changes and mixed distributions to further enhance and stabilize CLIP's zero-shot generalization capabilities.

REFERENCES

[1] C. Jia, Y. Yang, Y. Xia, Y. Chen, Z. Parekh, H. Pham, Q. V. Le, Y. Sung, Z. Li, and T. Duerig, "Scaling up visual and vision-language representation learning with noisy text supervision," in *ICML*, vol. 139. PMLR, 2021, pp. 4904–4916.

[2] A. Fürst, E. Rumetshofer, J. Lehner, V. T. Tran, F. Tang, H. Ramsauer, D. P. Kreil, M. Kopp, G. Klambauer, A. Bitto, and S. Hochreiter, "CLOOB: modern hopfield networks with infoloob outperform CLIP," in *NeurIPS*, 2022.

[3] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, "Learning transferable visual models from natural language supervision," in *ICML*, vol. 139. PMLR, 2021, pp. 8748–8763.

[4] J. Ding, N. Xue, G. Xia, and D. Dai, "Decoupling zero-shot semantic segmentation," in *CVPR*. IEEE, 2022, pp. 11 573–11 582.

[5] P. Liu, W. Yuan, J. Fu, Z. Jiang, H. Hayashi, and G. Neubig, "Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing," *ACM Comput. Surv.*, vol. 55, no. 9, pp. 195:1–195:35, 2023.

[6] B. Recht, R. Roelofs, L. Schmidt, and V. Shankar, "Do CIFAR-10 classifiers generalize to cifar-10?" *CoRR*, vol. abs/1806.00451, 2018.

[7] D. Hendrycks and T. G. Dietterich, "Benchmarking neural network robustness to common corruptions and perturbations," in *ICLR*. OpenReview.net, 2019.

[8] M. Shu, W. Nie, D.-A. Huang, Z. Yu, T. Goldstein, A. Anandkumar, and C. Xiao, "Test-time prompt tuning for zero-shot generalization in vision-language models," 2022.

[9] D. Wang, E. Shelhamer, S. Liu, B. A. Olshausen, and T. Darrell, "Tent: Fully test-time adaptation by entropy minimization," in *ICLR*, 2021.

[10] P. Morerio, R. Volpi, R. Ragonesi, and V. Murino, "Generative pseudo-label refinement for unsupervised domain adaptation," in *WACV*, 2020, pp. 3119–3128.

[11] P. T. Sivaprasad and F. Fleuret, "Uncertainty reduction for model adaptation in semantic segmentation," in *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual*, 2021, pp. 9613–9623.

[12] Q. Wang, O. Fink, L. V. Gool, and D. Dai, "Continual test-time domain adaptation," in *CVPR*. IEEE, 2022, pp. 7191–7201.

[13] M. Döbler, R. A. Marsden, and B. Yang, "Robust mean teacher for continual and gradual test-time adaptation," in *CVPR*. IEEE, 2023, pp. 7704–7714.

[14] J. Wang, P. Zhou, M. Z. Shou, and S. Yan, "Position-guided text prompt for vision-language pre-training," in *CVPR*. IEEE, 2023, pp. 23 242–23 251.

[15] X. Gu, T. Lin, W. Kuo, and Y. Cui, "Open-vocabulary object detection via vision and language knowledge distillation," in *ICLR*. OpenReview.net, 2022.

[16] B. Li, K. Q. Weinberger, S. J. Belongie, V. Koltun, and R. Ranftl, "Language-driven semantic segmentation," in *ICLR*. OpenReview.net, 2022.

[17] H. A. Rasheed, M. Maaz, M. U. Khattak, S. H. Khan, and F. S. Khan, "Bridging the gap between object and image-level representations for open-vocabulary detection," in *NeurIPS*, 2022.

[18] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever, "Language models are unsupervised multitask learners," 2019.

[19] N. Ding, S. Hu, W. Zhao, Y. Chen, Z. Liu, H. Zheng, and M. Sun, "Openprompt: An open-source framework for prompt-learning," in *ACL*. Association for Computational Linguistics, 2022, pp. 105–113.

[20] J. Devlin, M. Chang, K. Lee, and K. Toutanova, "BERT: pre-training of deep bidirectional transformers for language understanding," in *NAACL-HLT*. Association for Computational Linguistics, 2019, pp. 4171–4186.

[21] F. Petroni, T. Rocktäschel, S. Riedel, P. S. H. Lewis, A. Bakhtin, Y. Wu, and A. H. Miller, "Language models as knowledge bases?" in *EMNLP-IJCNLP*. Association for Computational Linguistics, 2019, pp. 2463–2473.

[22] J. Luo, Y. Li, Y. Pan, T. Yao, H. Chao, and T. Mei, "Coco-bert: Improving video-language pre-training with contrastive cross-modal matching and denoising," in *ACM*. ACM, 2021, pp. 5600–5608.

[23] K. Zhou, J. Yang, C. C. Loy, and Z. Liu, "Learning to prompt for vision-language models," *CoRR*, vol. abs/2109.01134, 2021.

[24] ——, "Conditional prompt learning for vision-language models," *CoRR*, vol. abs/2203.05557, 2022.

[25] J. Liang, D. Hu, and J. Feng, "Do we really need to access the source data? source hypothesis transfer for unsupervised domain adaptation," in *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, ser. Proceedings of Machine Learning Research, vol. 119. PMLR, 2020, pp. 6028–6039. [Online]. Available: http://proceedings.mlr.press/v119/liang20a.html

[26] M. Zhang, S. Levine, and C. Finn, "MEMO: test time robustness via adaptation and augmentation," in *NeurIPS*, 2022.

[27] S. Goyal, M. Sun, A. Raghunathan, and J. Z. Kolter, "Test time adaptation via conjugate pseudo-labels," in *NeurIPS*, 2022.

[28] S. Niu, J. Wu, Y. Zhang, Z. Wen, Y. Chen, P. Zhao, and M. Tan, "Towards stable test-time adaptation in dynamic wild world," in *ICLR*. OpenReview.net, 2023.

[29] P. T. Sivaprasad and F. Fleuret, "Uncertainty reduction for model adaptation in semantic segmentation," in *CVPR*. Computer Vision Foundation / IEEE, 2021, pp. 9613–9623.

[30] R. Wang, H. Zuo, Z. Fang, and J. Lu, "Multiple teacher model for continual test-time domain adaptation," in *AI 2023: Advances in Artificial Intelligence - 36th Australasian Joint Conference on Artificial Intelligence*, ser. Lecture Notes in Computer Science, vol. 14471. Springer, 2023, pp. 304–314.

[31] T. Gong, J. Jeong, T. Kim, Y. Kim, J. Shin, and S. Lee, "NOTE: robust continual test-time adaptation against temporal correlation," in *NeurIPS*, 2022.

[32] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *CoRR*, vol. abs/1512.03385, 2015.

[33] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," *CoRR*, vol. abs/1706.03762, 2017.

[34] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," in *ICLR*. OpenReview.net, 2021.

[35] Y. Gu, X. Yang, K. Wei, and C. Deng, "Not just selection, but exploration: Online class-incremental continual learning via dual view consistency," in *CVPR*. IEEE, 2022, pp. 7432–7441.