Disentangle Source and Target Knowledge for Continual Test-Time Adaptation

Tianyi Ma
University of Technology Sydney
Sydney, Australia
tianyi.ma@student.uts.edu.au

Maoying Qiao
University of Technology Sydney
Sydney, Australia
maoying.qiao@uts.edu.au

Abstract

***Co**ntinual **T**est-**T**ime **A**daptation (CoTTA) task is proposed to tackle the challenges of constant domain shifts during testing. The goals are twofold: 1) to preserve the knowledge from the source domain without source data and 2) to effectively extract target knowledge using unlabeled target domain data. Existing works primarily focus on either source or target knowledge, attempting to learn both in a mixed manner. This may harm the source knowledge preservation and target knowledge extraction. To this end, this paper proposes a **So**urce and **Ta**rget knowledge **Di**sentangle **T**ransformer (SoTa-DiT) with the prompting mechanism. Specifically, in a vision transformer (ViT), we employ source and target prompts, supervised by two groups of deliberately designed loss functions, to learn source and target knowledge separately. The source prompt focuses on anti-source-forgetting by extracting and preserving knowledge from the source model, while the target prompt focuses on pro-target-extracting using target data contrastive learning. With comprehensive evaluations across various datasets using different ViT backbones, we demonstrate that this dual-prompt architecture of SoTa-DiT is effective and that disentangling knowledge with the prompts benefits CoTTA. As a result, SoTa-DiT significantly improves image classification accuracy under the CoTTA setting.*

1. Introduction

Test-time adaptation (TTA) [6,17,21,23,25,32,42,46,47] addresses distribution shifts between the source training domain and the target test domain. Standard source-free TTA methods involve tuning a model with unlabeled test-time data to mitigate the domain shift problem. While effectively bridging the domain gap between the source and a single fixed target domain, they struggle with the constant target domain shifts in the test time. Such constant distribution shifts of the test domain are common in real-world applications. For instance, an auto-driving model may encounter distribution shifts due to changes in weather, light conditions, and surrounding environments. These domain shifts hinder the model from achieving optimum performance.

Continual **T**est-**T**ime **A**daptation (CoTTA) [37, 43] is proposed to address the challenge of constant target domain distribution shifts. In CoTTA, an off-the-shelf pre-trained source model is provided to be tuned and tested with unlabeled target data from various target domains across different time steps. The objectives for better CoTTA are twofold: 1) to preserve the source domain knowledge over time without access to source data and 2) to efficiently extract the target domain knowledge with only unlabeled test-time target domain data. Recent works mainly focus on preserving the source knowledge. For instance, tuning a more robust teacher model [28, 31, 43], training a stabilized model insensitive to the domain shifts [13,32], or learning a group of prototype based on the source model [4]. For target knowledge extracting, recent work facilitates efficient sample selection strategy for negative and positive learning [33, 44]. However, recent works do not address both objectives simultaneously. They mainly focus on refining only one type of knowledge or extracting source and target knowledge in a mixed manner. We argue that this may harm knowledge preserving and extracting. Thus, we need to consider both types of knowledge and extract them separately.

To this end, we proposed a **So**urce and **Ta**rget knowledge **Di**stentangle **T**ransformer (SoTa-DiT) to explore knowledge disentangle for CoTTA utilizing the prompting mechanism in a vision transformer (ViT) backbone, as illustrated in Fig.1. In SoTa-DiT, we adopt the ViT [12] backbone because it enables the parallel use of multiple tiny, easy-to-plug-in vectors named visual prompts to extract knowledge separately, yielding a disentangle effect. We then adapt the standard student and teacher distillation architecture following [43] for the ViT model as our baseline. Based on that, a dual-prompt architecture is further designed to extract the source and target knowledge separately. More specifically, our SoTa-DiT has three key points:

First, SoTa-DiT employs a visual prompt named **S**ource **P**rompt (SP) to extract and preserve the source knowledge. SP is concatenated with the patched images to be fed into

the transformer model. A group of loss functions are deliberately designed to tune SP and integrate the source knowledge into the model. We demonstrate that incorporating a contrastive loss between the augmented image feature embedded by the source model and the original image feature embedded by SP with the adapted model, along with a similarity loss between SP and the source model classification token, helps extract and preserve the source knowledge within SP. Additionally, implementing the symmetric cross-entropy loss [11] between SP output and model output while keeping SP fixed during back-propagation effectively integrates the preserved source knowledge into the model.

Second, SoTa-DiT employs another visual prompt called **T**arget **P**rompt (TP) to extract target-domain knowledge. TP is also concatenated with patched images and fed into the transformer. A mask is applied to TP to prevent interactions between SP and TP. Following [11], we train TP with a contrastive loss conducted between the original and augmented image features. Additionally, a pseudo-label loss is conducted between TP prediction and combined TP+SP prediction to provide TP with necessary label information.

Third, SoTa-DiT combines the source and target knowledge from SP and TP by simply averaging the SP and TP predictions after the softmax layers. Our evaluation demonstrates that this straightforward operation effectively integrates the source and target knowledge disentangled by visual prompts, resulting in a performance boost. By incorporating the three key points mentioned above, SoTa-DiT effectively preserves the source knowledge while extracting and integrating novel target domain knowledge. Consequently, SoTa-DiT adapts to the continually changing target domains effectively and significantly improves the classification accuracy on multiple datasets with different ViT backbones. For instance, SoTa-DiT achieves 61.2% average accuracy on ImageNet-C with ViT-B-16 backbone, $+4.5\%$ compared with the state-of-the-art method.

Our key contributions are summarized as follows:

• We proposed a novel **So**urce and **Ta**rget knowledge **Di**stentangle **T**ransformer (SoTa-DiT) method for continual test-time adaptation (CoTTA). SoTa-DiT disentangles the source and target knowledge through prompt learning.

• For source knowledge, we designed a source prompt supervised by the source model using source contrastive and similarity loss. For target knowledge, we designed a target prompt supervised by the target contrastive and cross-entropy loss. The two prompts extract two types of knowledge separately without direct interaction.

• We conduct comprehensive experiments to evaluate the effectiveness of source and target prompts in SoTa-DiT with different ViT backbones across various datasets. The results demonstrate the effectiveness of SoTa-DiT.

2. Related Works

2.1. Continual Test-Time Adaptation

Continual test-time adaptation (CoTTA) [43] addresses tha challenge of consecutive domain shifts during testing. Unlike standard test-time adaptation [16, 29, 46, 48], which deals with a single novel test domain, CoTTA tests the model against constantly changing domains. Recent research primarily focuses on preventing the adapted model from forgetting the pre-trained source model knowledge. For instance, [43] uses exponential update to learn a teacher model that preserves the source knowledge, while [11] explores the symmetrical cross-entropy loss for a more robust teacher. [32] designed an anti-forgetting regulation to supervise the adapted model with the source model. [4] supervises the model with prototypes generated from the source model. [38] adopts a self-distillation model that tunes only the additional adapter with the BN layer to retain the source knowledge. [13] applies a decorative prompt for the convolution network to extract domain-agnostic knowledge, and [33] dynamically reweights the predictions of source and adapted models. Other works focus on extracting high-quality target domain knowledge for adaptation. For example, [44] detects hard samples to conduct negative learning and reweighs the influence of samples to extract unbiased target knowledge. [11] explores contrastive learning for target domain knowledge learning. This paper argues that both source knowledge preservation and target knowledge extraction are critical. Jointly learning two types of knowledge may hinder either the source knowledge preservation or target knowledge absorption. Hence, we propose disentangling the source and target knowledge to enhance source preservation and target extraction.

2.2. Visual Prompting Mechanism

Vision transformer (ViT) [12] is an attention-based architecture for vision recognition and has been applied for multiple computer vision tasks like image classification [2, 3, 5], object detection [1, 49, 52], and image segmentation [9, 20, 41]. Prompting mechanism [26, 50, 51], initially proposed for the natural language processing, modifies the transformer model for downstream tasks with prompt tuning. Visual prompt learning serves a similar purpose [7, 22, 24, 30, 51]. Recent works have leveraged the prompt learning mechanism for cross-domain learning. For instance, [36] uses a style-conditioned visual prompt for general knowledge adaptation. [15] applies the visual prompt to learning target-domain knowledge for test-time adaptation. [27] facilitates domain-general and domain-specific prompts to disentangle general and specific knowledge for cross-domain few-shot learning. Inspired by these works, we proposed SoTa-DiT, which employs source prompt and target prompt to extract source-related

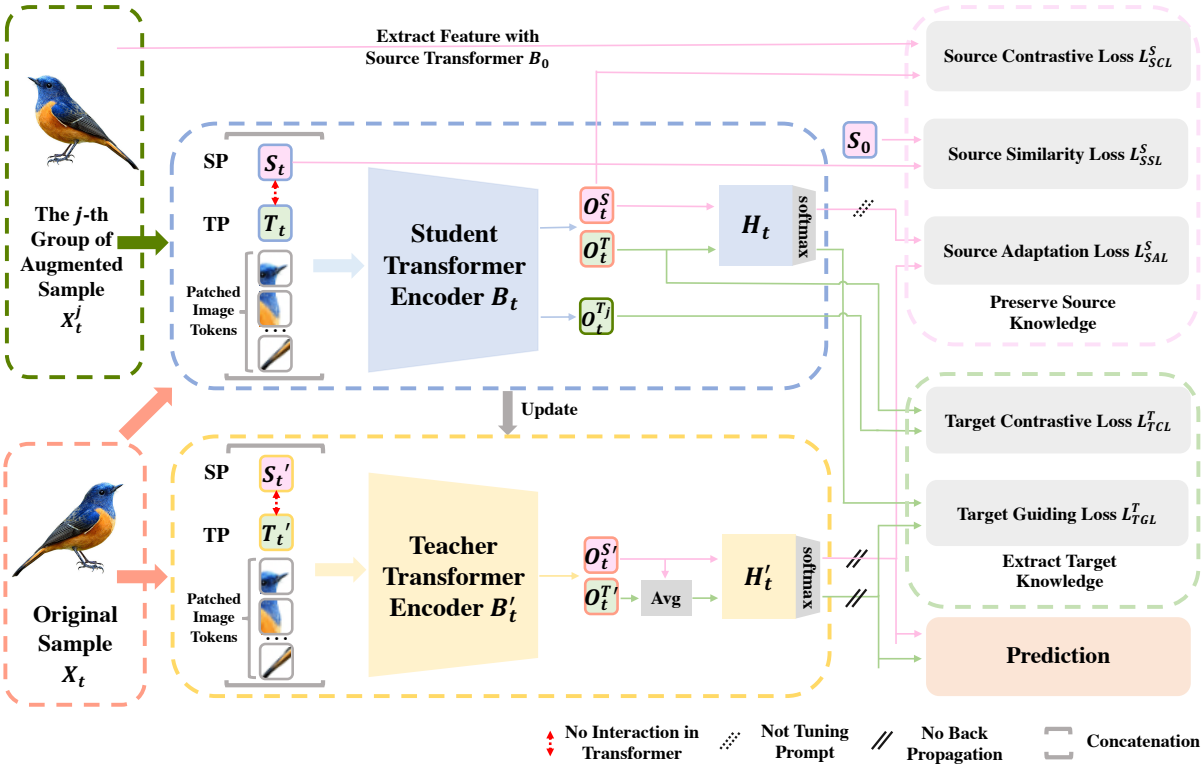


Figure 1. The overall architecture of SoTa-DiT involves two prompts, named source prompt (SP) and target prompt (TP), within a vision transformer (ViT) backbone using a teacher-student distillation architecture. At the time step 0, both teacher and student networks are initialized with the source model. Meanwhile, the prompts from both networks copy the source classification token. At time step t, we feed the original and augmented images to get prompt outputs O_t. The augmented images are also fed into the source model to get the source token output. These outputs are then utilized to tune 1) SP by calculating a source contrastive loss $\mathcal{L}_{SCL}^S$ and a source similarity loss $\mathcal{L}_{SSL}^S$ to preserve source knowledge, and 2) TP by calculating a target contrastive loss $\mathcal{L}_{TCL}^T$ to extract target knowledge and a target guiding loss $\mathcal{L}_{TGL}^T$ to learn label information. Furthermore, a source adaptation loss $\mathcal{L}_{SAL}^S$ is calculated to adapt the preserved source knowledge to the model. Finally, we average the predictions of SP and TP from the teacher model to obtain the final prediction.

and target-related knowledge separately, yielding a disentangle effect. While [13] also explores the visual prompt for CoTTA, there are three notable distinctions: 1) The purpose differs: we utilize prompts to disentangle source and target knowledge while their work applies prompts to retain domain-agnostic knowledge. 2) The prompt form differs: SoTa-DiT applies prompts in ViT to extract knowledge from the entire image, whereas their approach decorates only part of the feature map in a standard CNN backbone. 3) The training strategy differs: SoTa-DiT trains the prompts separately with different types of learnings and different loss functions for distinct purposes, while their work trains the two prompts jointly with the same loss function.

3. Methodology

3.1. Problem Formulation

This paper explores image classification task under the source-free CoTTA setting. We start with a source model $f_{\theta_s}(x)$ pre-trained on the source domain Φ_S with data $(\mathcal{X}^S, \mathcal{Y}^S)$. Under the source-free setting, $(\mathcal{X}^S, \mathcal{Y}^S)$ is unavailable for fine-tuning. Given $f_{\theta_s}(x)$, our objective is to achieve a higher image classification accuracy when testing on continually changing target domains. At test time, we have unlabelled target data batches $\mathcal{X}^T = (X_1, X_2, ...)$ from test domains different from Φ_S. These target data are presented to the model sequentially, following a time-step sequence. At a time step t, f_{θ_t} classify target data batch X_t. Meanwhile, X_t is used to tune $f_{\theta_t}(x)$ for the next time step $t+1$. Note that the distribution of X_t is constantly changing, and the model is evaluated in an online manner when adapting dynamically to each new data batch.

3.2. Overall Architecture

The overall architecture of **So**urce and **Ta**rget knowledge **Di**stentangle **T**ransformer (SoTa-DiT) is illustrated in Fig.1. In SoTa-DiT, we utilize a student model, denoted as $f_\theta = (B, \mathbf{S}, \mathbf{T}, H)$ and a teacher model, denoted as

$f_{\theta'} = (B', \mathbf{S}', \mathbf{T}', H')$. Each model consists of a transformer, denoted as B, a **S**ource **P**rompt (SP), denoted as $\mathbf{S}$, a **T**arget **P**rompt (TP), denoted as $\mathbf{T}$, and a classification head, denoted as H. At the time step 0, f_{θ_0} and $f_{\theta_0'}$ are both initialized with the source model f_{θ_s}. The visual prompts, specifically S_0, T_0, $S_0{}'$ and $T_0{}'$, are initialized with the classification token from the source model f_{θ_s}.

At time step t, we present a test data batch $X_t = \{x_1, x_2, ...x_{N_t}\}$, comprising N_t images from C categories. The data batch is sampled from an unknown domain with a distribution different from the source dataset. Each image x_i is first processed into patched image tokens $G_i \in \mathbb{R}^{K \times D}$, where K represents the number of tokens and D denotes the dispatch embedding dimension. For $\mathbf{S}$ and $\mathbf{T}$, we set $\mathbf{S}, \mathbf{T} \in \mathbb{R}^{1 \times D}$ to ensure their compatibility with the embedding dimension of the patched image tokens. Then, a mask is applied to $\mathbf{S}$ and $\mathbf{T}$ to prevent the direct interaction between them within B_t and $B_t{}'$. Finally, we concatenate $\mathbf{S}$ and $\mathbf{T}$ to G and feed them into f_{θ_t} and $f_{\theta_t'}$, following:

$$\begin{aligned}(O_i^S, O_i^T, E_i) &= B_t([\mathbf{S}_t, \mathbf{T}_t, G_i]) \\ (O_i^{S'}, O_i^{T'}, E_i{}') &= B_t{}'([\mathbf{S}_t{}', \mathbf{T}_t{}', G_i])\end{aligned} \tag{1}$$

where $[\ \ ]$, O, and E represent concatenation operation, prompt output, and the patched image embedding after the transformer blocks, respectively.

Following the standard prompt learning scheme, we discard E and use prompt output O for subsequent operations. We denote the output of $\mathbf{S}$ and $\mathbf{T}$ for X_t^T from f_{θ_t} and $f_{\theta_t'}$ as $\mathbf{O}_t^S = (O_1^S, O_2^S, ...O_{N_t}^S)$, $\mathbf{O}_t^T = (O_1^T, O_2^T, ...O_{N_t}^T)$, $\mathbf{O}_t^{S'} = (O_1^{S'}, O_2^{S'}, ...O_{N_t}^{S}{}')$ and $\mathbf{O}_t^{T'} = (O_1^{T'}, O_2^{T'}, ...O_{N_t}^{T}{}')$, respectively. Then, $\mathbf{O}_t^{S'}$ and $\mathbf{O}_t^{T'}$ are fed into the teacher classification head $H_t{}'$ and a softmax layer to get the prediction $\mathbf{P}_t^{S'} = (P_1^{S'}, P_2^{S'}, ...P_{N_t}^{S}{}')$ and $\mathbf{P}_t^{T'} = (P_1^{T'}, P_2^{T'}, ...P_{N_t}^{T}{}')$. Subsequently, we average $\mathbf{P}_t^{S'}$ and $\mathbf{P}_t^{T'}$ to obtain the final inference prediction $\mathbf{P}_t$ for X_t and calculate the classification accuracy.

Meanwhile, the student network f_{θ_t} is tuned for the next time step $t+1$ using the prompt output $\mathbf{O}$ from the time step t. The specific procedure of loss calculation for both prompts is explained in the following subsections. During the back-propagating procedure, B_t and H_t are updated with learning rate δ, while $\mathbf{S}$ and $\mathbf{T}$ are updated with learning rate $\mu \times \delta$. Here, μ is introduced to provide prompts an extra learning rate, enhancing their ability to extract the knowledge. After updating the student network $f_{\theta_{t+1}}$, we update teacher network $f_{\theta_t'}$ using the exponential moving averages (EMA):

$$f_{\theta_{t+1}{}'} = \gamma f_{\theta_t'} + (1-\gamma) f_{\theta_{t+1}}, \tag{2}$$

where hyper-parameter γ to controls the updating speed.

3.3. Preserve Source Knowledge

We utilize source prompt, denoted as $\mathbf{S}$, to preserve the source knowledge stored within the source model and adapt it to other parts of the model. $\mathbf{S}$ is trained with a source contrastive loss $\mathcal{L}_{SCL}^S$ and a source similarity loss $\mathcal{L}_{SSL}^S$ to preserve the source knowledge. Then, we apply a symmetric cross-entropy loss called source adaptation loss $\mathcal{L}_{SAL}^S$ to adapt the source knowledge to the other parts of the model. The loss calculation details are introduced as follows.

3.3.1 Source Contrastive Loss

At the time step t, we obtain the SP output of the original test image batch from the student network, denoted as $\mathbf{O}_t^S = (O_1^S, O_2^S, ...O_{N_t}^S)$. Meanwhile, each image from the test data batch X_t is randomly augmented M times, generating M groups of augmented data $X_t^1, X_t^2, ...X_t^M$. For each group of augmented data X_t^j, we have $X_t^j = (x_1^j, x_2^j, ...x_{N_t}^j)$. Then, the augmented data are processed into dispatched image tokens and fed into the source transformer B_0 with $\mathbf{S}_t$ being concatenated, following Eq.1. Finally, we obtain the output of the augmented data X_t^j from the source model, denoted as $\mathbf{O}_t^{S_j} = (O_1^{S_j}, O_2^{S_j}, ...O_{N_t}^{S_j})$. Finally, the source contrastive loss $\mathcal{L}_{SCL}^S$ is calculated as:

$$\mathcal{L}_{SCL}^S = -\frac{1}{M}\sum_{j=1}^{M}\sum_{i=1}^{N_t}\frac{\exp\left(\mathrm{sim}(O_i^S, O_i^{S_j})/\tau_s\right)}{\sum_{k=1}^{N_t}\exp\left(\mathrm{sim}(O_i^S, O_k^{S_j})/\tau_s\right)}, \tag{3}$$

where $\exp(x) = e^x$, and $\mathrm{sim}(z_i, z_j) = \frac{z_i \cdot z_j}{\|z_i\| \times \|z_j\|}$.

With $\mathcal{L}_{SCL}^S$, $\mathbf{S}$ extracts and preserves the source knowledge from the source model. However, $\mathbf{S}$ can still become unstable as the time step increases. To further stabilize the tuning process of $\mathbf{S}$, we apply a simple yet effective source similarity loss to restrict $\mathbf{S}$.

3.3.2 Source Similarity Loss

To stabilize the tuning procedure of $\mathbf{S}$, we constrain the source prompt S_t at each time step t with the source prompt S_0 from the initial time step using a similarity loss. Notice that the initial source prompt is the same as the classification token from the source model. The similarity loss $\mathcal{L}_{SSL}^S$ is calculated as followed:

$$\mathcal{L}_{SSL}^S = 1 - |\frac{S_t \cdot S_0}{\|S_t\| \times \|S_0\|}|, \tag{4}$$

.

The source similarity loss ensures that the shape of $\mathbf{S}_t$ remains similar to the original source model classification token. This restriction stabilizes the tuning of $\mathbf{S}$. Further, we allow $\mathbf{S}_t$ to be reasonably elastic by adding a weight parameter.

3.3.3 Source Adaptation Loss

With $\mathcal{L}_{SCL}^{S}$ and $\mathcal{L}_{SSL}^{S}$, the source prompt $\mathbf{S}$ extracts the source knowledge from the source model and preserves it effectively. To further adapt the preserved source knowledge to other parts of the model, we apply a source adaptation loss, denoted as $\mathcal{L}_{SAL}^{S}$. During the backpropagation, $\mathcal{L}_{SAL}^{S}$ updates all model parameters except for the source prompt $\mathbf{S}$. This process ensures that $\mathbf{S}$ adapts the source knowledge to other network parts without losing the preserved knowledge.

At the time step t, we first obtain the SP output of the image batch from both the student and teacher networks, denoted as $\mathbf{O}_t^S = (O_1^S, O_2^S, ...O_{N_t}^S)$ and $\mathbf{O}_t^{S'} = (O_1^{S'}, O_2^{S'}, ...O_{N_t}^{S}{}')$, respectively. We then fed $\mathbf{O}_t^S$ and $\mathbf{O}_t^{S'}$ into the student classification H_t and teacher classification head $H_t{}'$ along with softmax layers. This process generates the SP predictions of the student and teacher models for the C image categories, denoted as $\mathbf{P}_t^S = (P_1^S, P_2^S, ...P_{N_t}^S)$ and $\mathbf{P}_t^{S'} = (P_1^{S'}, P_2^{S'}, ...P_{N_t}^{S}{}')$, respectively. Here, $P_i^S = (p_1^S, p_2^S, ...p_C^S)$, and $P_i^{S'} = (p_1^{S'}, p_2^{S'}, ...p_C^{S}{}')$, where p_i^S and $p_i^{S'}$ represent the possibilities that a test sample belongs to each category. Finally, the source adaptation loss $\mathcal{L}_{SAL}^{S}$ is calculated between $\mathbf{P}_t^S$ and $\mathbf{P}_t^{S'}$, given by:

$$\mathcal{L}_{SAL}^{S} = -\frac{1}{N_t C}\sum_{i=1}^{N_t}(\frac{1}{2}P_i^{S'}\cdot\log P_i^S + \frac{1}{2}P_i^S\cdot\log P_i^{S'}), \quad (5)$$

.

The overall loss for $\mathbf{S}$ is formulated as:

$$\mathcal{L}^S = \mathcal{L}_{SAL}^{S} + \alpha\mathcal{L}_{SCL}^{S} + \eta\mathcal{L}_{SSL}^{S}, \quad (6)$$

where hyper-parameter α and η controls the loss weights.

3.4. Extract Target Knowledge

We utilize the target prompt, denoted as $\mathbf{T}$, to extract target knowledge from the unlabelled test data efficiently. Specifically, $\mathbf{T}$ is trained with a target contrastive loss $\mathcal{L}_{TCL}^{T}$ and a target guiding loss $\mathcal{L}_{TGL}^{T}$, as described in the following two subsections.

3.4.1 Target Contrastive Loss

We first obtain the TP output of the original image from the student network, denoted as $\mathbf{O}_t^T = (O_1^T, O_2^T, ...O_{N_t}^T)$. Similar to the source contrastive loss calculation, we randomly augment the test data batch X_t, generating M groups of augmented data $X_t^1, X_t^2, ...X_t^M$. The augmented data are then processed and fed into the student transformer B_t with $\mathbf{T}_t$ being concatenated, following Eq.1. Then, we obtain the TP output of the augmented data X_t^j, denoted as $\mathbf{O}_t^{T_j} = (O_1^{T_j}, O_2^{T_j}, ...O_{N_t}^{T_j})$. Finally, the target contrastive loss $\mathcal{L}_{TCL}^{T}$ is calculated as:

$$\mathcal{L}_{TCL}^{T} = -\frac{1}{M}\sum_{j=1}^{M}\sum_{i=1}^{N_t}\frac{\exp\left(\mathrm{sim}(O_i^T, O_i^{T_j})/\tau_t\right)}{\sum_{k=1}^{N_t}\exp\left(\mathrm{sim}(O_i^T, O_k^{T_j})/\tau_t\right)}, \quad (7)$$

.

With the contrastive loss $\mathcal{L}_{TCL}^{T}$, $\mathbf{T}_t$ extracts target knowledge by clustering the test images based on their similarities. However, without further label information, $\mathbf{T}_t$ may struggle to predict the actual image label. To address this, we design a guiding loss to provide $\mathbf{T}_t$ with label information.

3.4.2 Target Guiding Loss

We utilize both $\mathbf{S}_t$ and $\mathbf{T}_t$ to generate a label prediction to guild $\mathbf{T}_t$. The TP and SP outputs from the teacher, denoted as $\mathbf{O}_t^{T'}$ and $\mathbf{O}_t^{S'}$, are fed into teacher classification head $H_t{}'$ to generate the label prediction $P_i^{T+S'} = (p_1^{T+S'}, p_2^{T+S'}, ...p_C^{T+S'})$, given by:

$$P_i^{T+S'} = \mathrm{softmax}\big(H_t{}'(\frac{\mathbf{O}_t^{T'} + \mathbf{O}_t^{S'}}{2})\big). \quad (8)$$

Then, we feed the TP output $\mathbf{O}_t^T = (O_1^T, O_2^T, ...O_{N_t}^T)$ into the student classification head H_t and a softmax layer to obtain the TP prediction $\mathbf{P}_t^T = (P_1^T, P_2^T, ...P_{N_t}^T)$, with $P_i^T = (p_1^T, p_2^T, ...p_C^T)$. Finally, we calculate the target guilding loss $\mathcal{L}_{TGL}^{T}$ using $\mathbf{P}_t^T$ and $P_i^{T+S'}$, given by:

$$\mathcal{L}_{TGL}^{T} = -\frac{1}{N_t C}\sum_{i=1}^{N_t}(\frac{1}{2}P_i^{T+S'}\cdot\log P_i^T + \frac{1}{2}P_i^T\cdot\log P_i^{T+S'}). \quad (9)$$

The target guiding loss enables $\mathbf{T}$ to produce more accurate label prediction with the target knowledge. We formulate the overall loss for the target prompt $\mathbf{T}$ as:

$$\mathcal{L}^T = \mathcal{L}_{TCL}^{T} + \beta\mathcal{L}_{TGL}^{T}, \quad (10)$$

where hyper-parameter β controls the loss weights.

3.4.3 Overall Loss

Finally, we formulate the overall loss function $\mathcal{L}$ as:

$$\mathcal{L} = \delta L^S + L^T. \quad (11)$$

, where hyper-parameter δ controls the overall loss weights.

Table 1. The averaged image classification accuracy (%) of different methods with various transformer backbone on multiple datasets. IN-C, IN-R, IN-D109 represents ImageNet-C, ImageNet-R, and ImageNet-D109, respectively. 'Source' represents the backbone model without adaptation.

Dataset	Backbone	SOURCE	MEMO [47]	DDA [14]	TENT [42]	ETEA [32]	COTTA [43]	RMT [11]	SAR [33]	ROID [28] (sota)	SoTa-DiT
IN-C	ViT-S-16	28.2	39.8	41.0	43.1	43.9	43.8	44.1	44.5	45.1	**48.1**
	ViT-B-16	42.8	50.9	52.2	54.0	56.0	54.6	54.7	56.2	56.7	**61.2**
	ViT-L-16	46.2	52.9	51.6	59.3	61.2	59.9	62.1	62.4	63.9	**70.4**
IN-R	ViT-S-16	32.4	33.1	32.0	36.7	45.0	41.2	46.3	45.1	47.8	**51.2**
	ViT-B-16	44.0	45.4	45.6	46.7	51.0	30.4	31.2	51.4	55.8	**60.2**
	ViT-L-16	51.2	50.2	52.3	53.6	56.7	30.2	34.8	55.9	62.8	**69.4**
IN-D109	ViT-S-16	39.8	41.4	42.6	9.6	44.9	20.0	21.2	38.7	47.8	**50.2**
	ViT-B-16	46.4	40.2	47.0	16.0	52.6	26.6	25.8	42.6	55.0	**58.2**
	ViT-L-16	57.4	39.8	37.6	17.2	61.9	25.1	30.2	53.3	62.0	**68.2**

Table 2. The image classification accuracy (%) of different methods using Vit-B-16 backbone on the ImageNet-C dataset for 15 different types of corruption. SoTa-DiT achieves state-of-the-art performance, exceeding other methods by clear margins.

Method	gaussian	shot	impulse	defocus	glass	motion	zoom	snow	frost	fog	bright	contrast	elsatic	pixelate	jpeg	**Avg.**
SOURCE	46.2	30.5	33.3	31.2	27.1	44.3	30.3	53.3	48.2	45.3	75.2	8.9	44.0	60.8	62.7	42.8
MEMO	53.4	31.2	35.4	52.2	44.7	50.6	42.8	37.5	48.8	60.1	68.2	63.2	51	62.4	62.1	50.9
DDA	52.2	30.9	38.4	50.0	45.2	50.1	46.1	38.8	49.2	66.1	70.2	65.4	56.0	62.3	61.5	52.2
TENT	51.3	34.9	39.9	56.2	43.2	54.8	50.1	40.1	52.9	60.0	74.5	68.1	53.4	62.7	67.7	54.0
ETEA	51.2	35.5	40.2	**58.7**	44.5	56.1	49.2	42.8	53.8	68.9	**78.4**	**68.4**	58.0	65.5	68.2	56.0
COTTA	52.7	34.2	40.8	53.4	45.2	55.1	49.8	40.2	50.6	**69.1**	76.6	66.4	55.2	64.1	65.1	54.6
RMT	55.4	63.9	56.6	50.6	54.5	57.1	44.7	50.6	51.8	47.8	76.2	31.0	56.5	65.5	58.4	54.7
SAR	51.8	36.4	41.5	53.7	46.7	57.5	**52.8**	50.1	50.7	68.1	74.6	65.7	57.9	68.9	65.9	56.2
ROID (sota)	54.8	36.2	58.2	55.4	46.2	57.1	50.2	42.9	57.2	62.3	77.6	67.2	53.4	64.2	67.9	56.7
SoTa-DiT (ours)	**63.8**	**66.5**	**65.0**	57.6	**59.9**	**61.0**	51.0	**61.8**	**62.6**	53.8	76.5	33.9	**65.7**	**70.5**	**69.0**	**61.2**

4. Experiments

4.1. Settings

Datasets. SoTa-DiT is evaluated on three CoTTA datasets: ImageNet-C [19], ImageNet-R [18], and ImageNet-D109 [33]. ImageNet-C includes 15 types of image corruption with 5 severity levels. We conduct the evaluations under the level 5 corruption. ImageNet-R includes 30000 images with different renditions from 200 categories. ImageNet-D109, a subset of ImageNet-D [34, 35], contains 109 classes with 6 types of domain shifts. SoTa-DiT is compared with the other methods on these three datasets, and ablation studies are conducted on the ImageNet-C dataset.

Implementation Details. We evaluate SoTa-DiT with three different ViT backbones: Vit-B-16, ViT-S-16, and ViT-L-16 [12]. The off-the-shelf source model are pretrained on ImageNet [10], following [8, 39, 40, 45, 53]. The ablation studies are conducted on the Vit-B-16 backbone. During the test-time tuning, the image batches are provided to the source network in an online manner, following [43]. The network predicts the image category and adapts to the test domain, which constantly changes over time. We set the learning rate to 0.001 and the tuning batch size to 8. γ (Eq.2) and the training step are set to 0.999 and 1. μ, τ_s (Eq.3), and τ_t (Eq.7) are set to 20, 0.05 and 0.05 by default. δ (Eq.11) is set to 1.1. α, η (Eq.6) and β (Eq.10) are set to 1.0, 0.9 and 1.0 by default. M (Eq.7) is set to 2 by default.

4.2. Effectiveness of SoTa-DiT

We compare SoTa-DiT with the baseline and the state-of-the-art methods in Tab.1. We draw three observations.

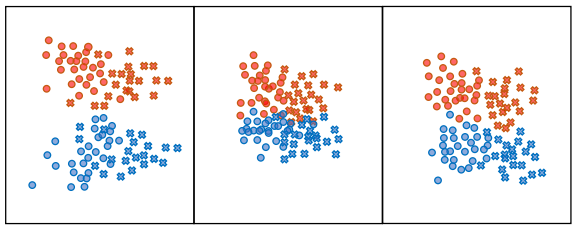
First, SoTa-DiT outperforms the state-of-the-art method across multiple datasets when evaluating with three common ViT backbones. Specifically, on the ViT-B-16 backbone, SoTa-DiT outperforms the state-of-the-art on ImageNet-C, ImageNet-R, and ImageNet-D109 by $+4.5\%$, $+4.4\%$, and $+3.2\%$, respectively.

Second, SoTa-DiT demonstrates more substantial performance advantages with larger ViT backbones. For instance, on ImageNet-C, SoTa-DiT with ViT-S-16, ViT-B-16, and ViT-L-16 outperforms the state-of-the-art by $+3.0\%$, $+4.5\%$, and $+6.5\%$, respectively.

Third, SoTa-DiT consistently outperforms other methods across various types of corruption on the ImageNet-C dataset. As detailed in Tab.2, SoTa-DiT outperforms other methods on 10 out of 15 types of corruption.

Table 3. Evaluation of the key components, including the Source Prompt (SP), Target Prompt (TP), and loss functions, on ImageNet-C. We list the averaged classification accuracy (%).

Method	$\mathcal{L}^S_{SCL}$	$\mathcal{L}^S_{SSL}$	$\mathcal{L}^S_{SAL}$	$\mathcal{L}^T_{TCL}$	$\mathcal{L}^T_{TGL}$	Average Acc.
Baseline (CoTTA)	✗	✗	✗	✗	✗	54.6
SP only	✓	✓	✓	✗	✗	55.1
TP only	✗	✗	✗	✓	✓	58.9
TP only+$\mathcal{L}^S_{SCL}$	✓	✗	✗	✓	✓	59.2
SoTa-DiT*	✓	✓	✓	✗	✗	60.4
SoTa-DiT	✓	✓	✓	✓	✓	**61.2**



(a) TP Only **(b) SP Only** **(c) SoTa-DiT**

Figure 2. Visualization of image embeddings from two different categories, represented by crosses and circles, with two different types of corruption, colored in red and blue. Samples are selected from the ImageNet-C dataset.

4.3. Ablation of Key Components

4.3.1 Source and Target Prompts

We examine the effectiveness of source prompt (SP), target prompt (TP), and their associated loss functions on the ImageNet-C dataset, detailed in Tab.3. Additional results are shown in the supplemental material. The 'Baseline' refers to the ViT adapted with the method from [43]. In the 'SoTa-DiT*', we let $\mathcal{L}^S_{SAL}$ tune SP during the backpropagation. We draw four observations based on the results:

First, training with only SP or TP improves performance. Adding SP to the baseline increases the average accuracy by $+0.5\%$, demonstating that preserving source knowledge with SP alone benefits CoTTA. Adding TP to the baseline increases the average accuracy by $+4.3\%$, indicating that extracting target knowledge with TP boosts performance.

Second, comparing 'TP only+$\mathcal{L}^S_{SCL}$' with 'TP only,' we see that directly adding a source contrastive loss to the TP benefits CoTTA, with the accuracy increased by $+0.3\%$.

Third, comparing 'SoTa-DiT*' with 'SoTa-DiT' reveals that letting $\mathcal{L}^S_{SAL}$ tune the SP harms CoTTA, with a performance decrease of -0.8%. We infer that the preserved source knowledge within SP is harmed by $\mathcal{L}^S_{SAL}$.

Finally, jointly using SP and TP brings a significant performance boost compared to using SP ($+6.1\%$) or TP alone ($+2.3\%$). The result indicates that extracting source and target knowledge in a disentangling manner is beneficial.

Additionally, we visualized the feature embedding of samples from two image categories with two different types of corruption in ImageNet-C, as shown in Fig.2. The figure illustrates that using TP alone separates the image features from each other based on domain and class, which aids classification. However, several samples are misclassified and appear distant from their true class centers, possibly due to overfitting the extreme cases with TP alone and forgetting source knowledge. Conversely, using SP alone does not push the domain and class centers as far as using TP alone. However, the sample misclassification issue is slightly alleviated. When SP and TP are used together in SoTa-DiT, the final embeddings show clearer boundaries between categories and domains, with fewer samples misplaced in other clusters.

Table 4. Evaluation of different ways to combine source and target knowledge on ImageNet-C. We list the classification accuracy (%).

Method	Average Acc.
Baseline (CoTTA)	54.6
SP only	56.2
TP only	59.9
Avg. before H	60.4
SoTa-DiT	**61.2**

4.3.2 Knowledge Combining

We investigate different ways of combining source and target knowledge, as detailed in Tab.4. The table denotes different methods: 1) 'SP only': Utilize SP predictions alone. 2) TP only: Utilize TP predictions alone. 3) 'Avg. before H': Average SP and TP outputs before the classification head H. 4) 'SoTa-DiT': Average the SP and TP predictions after the softmax layer. We draw three observations:

Firstly, both SP and TP enhance the prediction accuracy after disentangling the source and target knowledge compared to the baseline. For instance, comparing 'SP only' from Tab.4 and 'SP only' from Tab.3, we observe an improvement of $+1.1\%$. Similarly, the 'TP only' also shows an increase in accuracy by $+1.0\%$. This proves that disentangled knowledge is of a higher quality.

Secondly, comparing 'Avg. before H' and 'SoTa-DiT', we observe that averaging the predictions after the softmax layers brings higher accuracy ($+0.8\%$). We infer that the target and source knowledge may provide better predictions for different samples. After the softmax layer, the advantages of both types of knowledge are further amplified.

Third, comparing 'SoTa-DiT' with 'SP only' and 'TP only', we find that jointly utilizing the disentangled source and target knowledge brings further improvement, with accuracy increasing by $+5.0\%$ and $+1.3\%$. This highlights the advantages of combining knowledge for CoTTA.

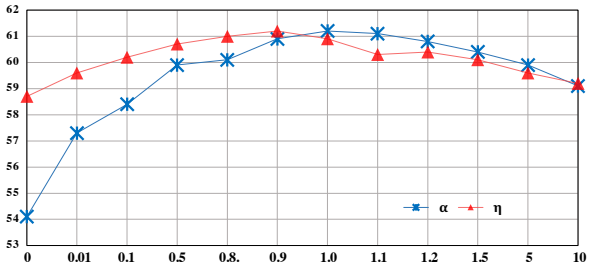


Figure 3. Averaged classification accuracy of SoTa-DiT with different source contrastive loss and source similarity loss weight, named 'α' and 'η', for the source prompt on ImageNet-C.

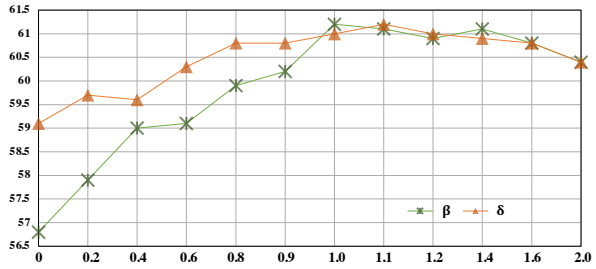


Figure 4. Averaged classification accuracy of SoTa-DiT with different target guiding loss weight 'β' for the target prompt and overall loss trade-off weight 'δ' on ImageNet-C.

4.4. Ablation of Hyper-parameters

4.4.1 Source Contrastive Loss Weight α

We investigate the influence of the source contrastive loss weight α in Eq.6 for SP. The results are illustrated in Fig.3 with a blue line. We draw the following observation:

As α increases, the accuracy increases and then decreases. We infer the reason as two-fold. First, when α is below the optimum value of 1.0, the weight of source contrastive loss is low, and the ability of SP to extract and preserve source knowledge decreases. Conversely, as α exceeds 1.0, the emphasis on the source contrastive loss increases, potentially overshadowing the cross-entropy loss. Consequently, the preserved source knowledge may not be adapted to the other parts of the model effectively.

4.4.2 Source Similarity Loss Weight η

We investigate the influence of the source similarity loss weight η in Eq.6 for SP. The results are illustrated in Fig.3 with a red line. We draw the following observation:

Similar to the effect observed with α, as η increases, the accuracy increases and then decreases. We infer the reason as two-fold. When η is below 0.9, similarity loss cannot effectively constrain the shape of SP, leading to forgotten source knowledge. When η exceeds 0.9, similarity loss dominates. As a result, SP loses its elasticity and cannot extract additional source knowledge.

4.4.3 Target Guiding Loss Weight β

We investigate the influence of the target guiding loss weight β in Eq.10 for TP. The results are depicted in Fig.4 with a green line. We draw the following observation:

As β increases, the accuracy increases significantly and then decreases. We infer that two factors contribute to this. Firstly, the guiding loss is inactive when β is less than 1.0. Consequently, TP is unable to learn label information effectively. As the target contrastive loss only clusters the samples without providing label information, TP may match sample clusters with random labels, leading to a severe mismatch between samples and their labels. When β exceeds 1.0, the contrastive loss is partially deactivated, hindering the effective extraction of target knowledge.

4.4.4 Overall Loss Trading-off Weight δ

We investigate the influence of the overall loss trade-off weight δ in Eq.10. The results are depicted in Fig.4 with an orange line. We draw the observation:

As δ increases, the accuracy increases gradually and then decreases. We infer that two factors contribute to this. When δ is smaller than 1.1, the weight of source prompt loss is low. SP is not properly tuned and thus cannot preserve enough source knowledge. Moreover, the preserved source knowledge is not effectively adapted to other parts of the model. Hence, the model is subject to source knowledge forgetting, leading to decreased performance. On the other hand, when δ exceeds 1.1, the weight of target prompt loss is relatively low. TP is not tuned properly and thus cannot extract enough target knowledge. Thus, the model cannot effectively adapt to the incoming novel domains.

5. Conclusion

This paper proposed a **S**ource and **T**arget knowledge **D**istentangle **T**ransformer (SoTa-DiT) for the continual test-time adaptation (CoTTA) task. SoTa-DiT utilizes two visual prompts, named source prompt and target prompt, within a vision transformer backbone to extract source knowledge and target knowledge in a disentangling manner. Supervised by two groups of deliberately designed loss, the source and target prompts enable the preservation of the source knowledge and effective learning of target knowledge. Our experiment results demonstrate the effectiveness of this dual-prompt architecture with the knowledge disentangling mechanism. SoTa-DiT significantly improves the image classification accuracy under the CoTTA setting with different transformer backbones across multiple datasets.

References

[1] Ershat Arkin, Nurbiya Yadikar, Xuebin Xu, Alimjan Aysa, and Kurban Ubul. A survey: object detection methods from cnn to transformer. *Multimedia Tools and Applications*, 82(14):21353–21383, 2023. 2

[2] Yakoub Bazi, Laila Bashmal, Mohamad M Al Rahhal, Reham Al Dayil, and Naif Al Ajlan. Vision transformers for remote sensing image classification. *Remote Sensing*, 13(3):516, 2021. 2

[3] Srinadh Bhojanapalli, Ayan Chakrabarti, Daniel Glasner, Daliang Li, Thomas Unterthiner, and Andreas Veit. Understanding robustness of transformers for image classification. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10231–10241, 2021. 2

[4] Goirik Chakrabarty, Manogna Sreenivas, and Soma Biswas. Santa: Source anchoring network and target alignment for continual test time adaptation. *Transactions on Machine Learning Research*, 2023. 1, 2

[5] Chun-Fu Richard Chen, Quanfu Fan, and Rameswar Panda. Crossvit: Cross-attention multi-scale vision transformer for image classification. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 357–366, 2021. 2

[6] Dian Chen, Dequan Wang, Trevor Darrell, and Sayna Ebrahimi. Contrastive test-time adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 295–305, 2022. 1

[7] Keyan Chen, Chenyang Liu, Hao Chen, Haotian Zhang, Wenyuan Li, Zhengxia Zou, and Zhenwei Shi. Rsprompter: Learning to prompt for remote sensing instance segmentation based on visual foundation model. *IEEE Transactions on Geoscience and Remote Sensing*, 2024. 2

[8] Xiangning Chen, Cho-Jui Hsieh, and Boqing Gong. When vision transformers outperform resnets without pretraining or strong data augmentations. *arXiv preprint arXiv:2106.01548*, 2021. 6

[9] Bowen Cheng, Ishan Misra, Alexander G Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1290–1299, 2022. 2

[10] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 6

[11] Mario Döbler, Robert A Marsden, and Bin Yang. Robust mean teacher for continual and gradual test-time adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7704–7714, 2023. 2, 6

[12] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 1, 2, 6

[13] Yulu Gan, Yan Bai, Yihang Lou, Xianzheng Ma, Renrui Zhang, Nian Shi, and Lin Luo. Decorate the newcomers: Visual domain prompt for continual test time adaptation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 7595–7603, 2023. 1, 2, 3

[14] Jin Gao, Jialing Zhang, Xihui Liu, Trevor Darrell, Evan Shelhamer, and Dequan Wang. Back to the source: Diffusion-driven adaptation to test-time corruption. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11786–11796, 2023. 6

[15] Yunhe Gao, Xingjian Shi, Yi Zhu, Hao Wang, Zhiqiang Tang, Xiong Zhou, Mu Li, and Dimitris N Metaxas. Visual prompt tuning for test-time domain adaptation. *arXiv preprint arXiv:2210.04831*, 2022. 2

[16] Taesik Gong, Jongheon Jeong, Taewon Kim, Yewon Kim, Jinwoo Shin, and Sung-Ju Lee. Note: Robust continual test-time adaptation against temporal correlation. *Advances in Neural Information Processing Systems*, 35:27253–27266, 2022. 2

[17] Xuehai He, Chunyuan Li, Pengchuan Zhang, Jianwei Yang, and Xin Eric Wang. Parameter-efficient model adaptation for vision transformers. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 817–825, 2023. 1

[18] Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, et al. The many faces of robustness: A critical analysis of out-of-distribution generalization. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 8340–8349, 2021. 6

[19] Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. *Proceedings of the International Conference on Learning Representations*, 2019. 6

[20] Jitesh Jain, Jiachen Li, Mang Tik Chiu, Ali Hassani, Nikita Orlov, and Humphrey Shi. Oneformer: One transformer to rule universal image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2989–2998, 2023. 2

[21] Minguk Jang, Sae-Young Chung, and Hye Won Chung. Test-time adaptation via self-training with nearest neighbor information. *arXiv preprint arXiv:2207.10792*, 2022. 1

[22] Menglin Jia, Luming Tang, Bor-Chun Chen, Claire Cardie, Serge Belongie, Bharath Hariharan, and Ser-Nam Lim. Visual prompt tuning. In *European Conference on Computer Vision*, pages 709–727. Springer, 2022. 2

[23] Shibo Jie and Zhi-Hong Deng. Fact: Factor-tuning for lightweight adaptation on vision transformer. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 1060–1068, 2023. 1

[24] Muhammad Uzair Khattak, Hanoona Rasheed, Muhammad Maaz, Salman Khan, and Fahad Shahbaz Khan. Maple: Multi-modal prompt learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19113–19122, 2023. 2

[25] Takeshi Kojima, Yusuke Iwasawa, and Yutaka Matsuo. Robustifying vision transformer without retraining from scratch

using attention-based test-time adaptation. *New Generation Computing*, 41(1):5–24, 2023. 1

[26] Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Computing Surveys*, 55(9):1–35, 2023. 2

[27] Tianyi Ma, Yifan Sun, Zongxin Yang, and Yi Yang. Prod: Prompting-to-disentangle domain knowledge for cross-domain few-shot image classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19754–19763, 2023. 2

[28] Robert A Marsden, Mario Döbler, and Bin Yang. Universal test-time adaptation through weight ensembling, diversity weighting, and prior correction. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2555–2565, 2024. 1, 6

[29] Chaithanya Kumar Mummadi, Robin Hutmacher, Kilian Rambach, Evgeny Levinkov, Thomas Brox, and Jan Hendrik Metzen. Test-time adaptation to distribution shift by confidence maximization and input transformation. *arXiv preprint arXiv:2106.14999*, 2021. 2

[30] Xing Nie, Bolin Ni, Jianlong Chang, Gaofeng Meng, Chunlei Huo, Shiming Xiang, and Qi Tian. Pro-tuning: Unified prompt tuning for vision tasks. *IEEE Transactions on Circuits and Systems for Video Technology*, 2023. 2

[31] Fahim Faisal Niloy, Sk Miraj Ahmed, Dripta S Raychaudhuri, Samet Oymak, and Amit K Roy-Chowdhury. Effective restoration of source knowledge in continual test time adaptation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2091–2100, 2024. 1

[32] Shuaicheng Niu, Jiaxiang Wu, Yifan Zhang, Yaofo Chen, Shijian Zheng, Peilin Zhao, and Mingkui Tan. Efficient test-time model adaptation without forgetting. In *International conference on machine learning*, pages 16888–16905. PMLR, 2022. 1, 2, 6

[33] Shuaicheng Niu, Jiaxiang Wu, Yifan Zhang, Zhiquan Wen, Yaofo Chen, Peilin Zhao, and Mingkui Tan. Towards stable test-time adaptation in dynamic wild world. *arXiv preprint arXiv:2302.12400*, 2023. 1, 2, 6

[34] Xingchao Peng, Qinxun Bai, Xide Xia, Zijun Huang, Kate Saenko, and Bo Wang. Moment matching for multi-source domain adaptation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1406–1415, 2019. 6

[35] Evgenia Rusak, Steffen Schneider, Peter Vincent Gehler, Oliver Bringmann, Wieland Brendel, and Matthias Bethge. Imagenet-d: A new challenging robustness dataset inspired by domain adaptation. In *ICML 2022 Shift Happens Workshop*, 2022. 6

[36] Kihyuk Sohn, Huiwen Chang, José Lezama, Luisa Polania, Han Zhang, Yuan Hao, Irfan Essa, and Lu Jiang. Visual prompt tuning for generative transfer learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19840–19851, 2023. 2

[37] Damian Sójka, Sebastian Cygert, Bartłomiej Twardowski, and Tomasz Trzciński. Ar-tta: A simple method for real-world continual test-time adaptation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3491–3495, 2023. 1

[38] Junha Song, Jungsoo Lee, In So Kweon, and Sungha Choi. Ecotta: Memory-efficient continual test-time adaptation via self-distilled regularization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11920–11929, 2023. 2

[39] Andreas Steiner, Alexander Kolesnikov, , Xiaohua Zhai, Ross Wightman, Jakob Uszkoreit, and Lucas Beyer. How to train your vit? data, augmentation, and regularization in vision transformers. *arXiv preprint arXiv:2106.10270*, 2021. 6

[40] Ilya Tolstikhin, Neil Houlsby, Alexander Kolesnikov, Lucas Beyer, Xiaohua Zhai, Thomas Unterthiner, Jessica Yung, Andreas Steiner, Daniel Keysers, Jakob Uszkoreit, Mario Lucic, and Alexey Dosovitskiy. Mlp-mixer: An all-mlp architecture for vision. *arXiv preprint arXiv:2105.01601*, 2021. 6

[41] Jeya Maria Jose Valanarasu, Poojan Oza, Ilker Hacihaliloglu, and Vishal M Patel. Medical transformer: Gated axial-attention for medical image segmentation. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part I 24*, pages 36–46. Springer, 2021. 2

[42] Dequan Wang, Evan Shelhamer, Shaoteng Liu, Bruno Olshausen, and Trevor Darrell. Tent: Fully test-time adaptation by entropy minimization. *arXiv preprint arXiv:2006.10726*, 2020. 1, 6

[43] 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. 1, 2, 6, 7

[44] Yanshuo Wang, Jie Hong, Ali Cheraghian, Shafin Rahman, David Ahmedt-Aristizabal, Lars Petersson, and Mehrtash Harandi. Continual test-time domain adaptation via dynamic sample selection. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1701–1710, 2024. 1, 2

[45] Xiaohua Zhai, Xiao Wang, Basil Mustafa, Andreas Steiner, Daniel Keysers, Alexander Kolesnikov, and Lucas Beyer. Lit: Zero-shot transfer with locked-image text tuning. *CVPR*, 2022. 6

[46] Jian Zhang, Lei Qi, Yinghuan Shi, and Yang Gao. Domainadaptor: A novel approach to test-time adaptation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 18971–18981, 2023. 1, 2

[47] Marvin Zhang, Sergey Levine, and Chelsea Finn. Memo: Test time robustness via adaptation and augmentation. *Advances in Neural Information Processing Systems*, 35:38629–38642, 2022. 1, 6

[48] Yifan Zhang, Xue Wang, Kexin Jin, Kun Yuan, Zhang Zhang, Liang Wang, Rong Jin, and Tieniu Tan. Adanpc: Exploring non-parametric classifier for test-time adaptation. In *International Conference on Machine Learning*, pages 41647–41676. PMLR, 2023. 2

[49] Zixiao Zhang, Xiaoqiang Lu, Guojin Cao, Yuting Yang, Licheng Jiao, and Fang Liu. Vit-yolo: Transformer-based yolo for object detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 2799–2808, 2021. 2

[50] Ruiqi Zhong, Kristy Lee, Zheng Zhang, and Dan Klein. Adapting language models for zero-shot learning by meta-tuning on dataset and prompt collections. *arXiv preprint arXiv:2104.04670*, 2021. 2

[51] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *International Journal of Computer Vision*, 130(9):2337–2348, 2022. 2

[52] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable detr: Deformable transformers for end-to-end object detection. *arXiv preprint arXiv:2010.04159*, 2020. 2

[53] Juntang Zhuang, Boqing Gong, Liangzhe Yuan, Yin Cui, Hartwig Adam, Nicha Dvornek, Sekhar Tatikonda, James Duncan, and Ting Liu. Surrogate gap minimization improves sharpness-aware training. *ICLR*, 2022. 6