Continuous Test-Time Adaptation via Dual Alignment

Boyuan Zhang , Jie Pen* , Shuai Yang , Lichuan Gu
School of Artificial Intelligence, Anhui Agricultural University, Hefei 230036, China
23115885@stu.ahau.edu.cn, {panjie, yangs, glc}@ahau.edu.cn

Abstract

Continual test-time adaptation (CTTA) aims to adapt a source-pretrained model to continuously shifting target distributions. Recent advances in this field either minimize prediction entropy on streaming target samples, which is efficient but prone to error accumulation, or rely on teacher–student pseudo-labeling, which provides more stable predictions at the cost of high computational overhead. To overcome this stability–efficiency trade-off, we propose Dual Alignment (D-Align), a stable and efficient CTTA method that synergistically combines correlation alignment and masked consistency alignment. Specifically, correlation alignment constructs a reliable pseudo-source domain by selecting resilient target samples, maintaining their diversity via a dynamic memory bank, and aligning nonlinear feature relations through image-level visual domain prompt–driven alignment, enabling effective feature alignment without costly backpropagation. Masked consistency alignment operates on domain-prompted target samples, enforcing prediction consistency across multiple masked views to promote robust and discriminative representations. Extensive experimental results validate the effectiveness of D-Align under dynamic domain shifts compared with state-of-the-art methods.

1 Introduction

Deep neural networks achieve strong performance under matched training and test distributions, however, their robustness degrades under distribution shifts [Zhou *et al.*, 2023]. Test-time adaptation (TTA) mitigates this issue by updating models online with unlabeled test data [Wang *et al.*, 2021], but suffers from error accumulation in continually evolving domains. Continuous test-time adaptation (CTTA) extends TTA to streaming scenarios by enabling continual adaptation across evolving target distributions [Wang *et al.*, 2022].

Existing CTTA methods can be broadly categorized into two paradigms: entropy minimization and teacher–student schemes. Entropy minimization updates either Batch Normalization statistics [Niu *et al.*, 2022] or a small subset of model parameters [Liu *et al.*, 2024b], yielding immediate performance gains with minimal computational cost. However, under continual distribution shifts, this strategy is highly sensitive to miscalibrated predictions, which leads to uncontrolled error accumulation over time. In contrast, teacher–student frameworks generate pseudo labels for target samples and can provide more stable supervision. Nevertheless, conventional mean-teacher approaches [Tarvainen and Valpola, 2017] become increasingly noisy in dynamic environments, amplifying distributional drift. Although recent works [Döbler *et al.*, 2023; Park *et al.*, 2025] employ test-time augmentation to improve pseudo-label quality, the required multiple forward passes incur substantial latency and computational overhead, limiting their practicality in CTTA.

As a result, existing methods face an inherent trade-off between computational efficiency and adaptation stability [Wang *et al.*, 2022; Liu *et al.*, 2024a]. This trade-off arises because current approaches lack reliable and adaptive mechanisms to model continuously evolving target distributions in online streaming settings, where non-stationary shifts progressively undermine both entropy-based updates and pseudo-label supervision. Consequently, achieving stable and efficient adaptation under continuously shifting target distributions remains a fundamental challenge in CTTA.

Motivated by the aforementioned issue, we propose Dual Alignment (D-Align), a stable and efficient CTTA method that synergistically combines correlation alignment and masked consistency alignment. Specifically, Correlation Alignment (CorA) performs feature distribution matching by constructing a pseudo-source domain. However, in the CTTA setting, where target data arrive sequentially and exhibit non-stationary shifts, constructing a reliable pseudo-source domain is non-trivial. To tackle this, we first introduce a Feature Resilience (FR) metric to select reliable samples that possess globally consistent features. Building upon this, we design a continual pseudo-source memory bank, which maintains the diversity and reliability of the pseudo-source distribution through a progressive dynamic sample update strategy. Furthermore, to compensate for the limitations of linear alignment in correlation alignment, we incorporate image-level visual domain prompting techniques to capture complex nonlinear relationships among features. Moreover, we intro-

*Corresponding author.

duce masked consistency alignment, where each target sample is first transformed by a domain-specific visual prompt and then subjected to different masking operations. Based on these prompted and masked views, entropy consistency and entropy ranking losses enforce prediction consistency, stabilizing the learning of target domain–specific knowledge while preserving discriminative power under partial feature occlusion. Extensive experiments on multiple standard benchmarks demonstrate that D-Align consistently achieves state-of-the-art performance under continuously evolving target distributions. In summary, our contributions can be summarized as:

- We make the first attempt to introduce correlation alignment into CTTA and propose CorA, which achieves feature alignment by applying a linear transformation to the output features without updating model parameters, thereby significantly reducing computational overhead.
- Subsequently, we propose a D-Align method that synergistically optimizes correlation and masked consistency alignment. By integrating a continual pseudo-source memory bank for feature rectification and masked views for structural learning, our approach ensures stable adaptation across continually shifting distributions.
- Extensive experiments on four benchmarks demonstrate that D-Align outperforms state-of-the-art methods, achieving error rates of 21.8% on CIFAR100-C and 37.9% on ImageNet-C.

2 Related Work

Continual Test-Time Adaptation refers to scenarios where the target domain is non-static, highlighting the need to adapt to continuous domain shifts and thereby extending the practicality of TTA. This paradigm promotes a rethinking of conventional protocols by emphasizing a stronger focus on stability. CoTTA was among the first to formalize the more challenging Continual Test-Time Adaptation setting and employs a mean teacher model to generate more stable pseudo-labels and introduces a stochastic weight restoration technique to mitigate catastrophic forgetting during long-term adaptation [Wang *et al.*, 2022]. NOTE proposes Instance-Aware Batch Normalization (IABN) to mitigate distribution shifts caused by temporally correlated and non-i.i.d. data streams [Gong *et al.*, 2022]. RoTTA further advances the paradigm by introducing Practical TTA (PTTA), simulating real-world dynamics and employing Robust Batch Normalization (RBN) with a memory bank [Yuan *et al.*, 2023]. In a different approach. ViDA designs a dual-branch adapter that explicitly separates domain-invariant and domain-specific features, and dynamically weights domain knowledge based on uncertainty, thereby effectively balancing the learning of new knowledge with the preservation of source domain knowledge during continual adaptation [Liu *et al.*, 2024b]. Continual-MAE enhances domain-invariant representations using a distribution-aware masked autoencoder and leverages Monte Carlo dropout to filter valid target regions based on uncertainty [Liu *et al.*, 2024a]. REM proposes Ranked Entropy Minimization to mitigate model collapse, stabilizing adaptation by explicitly enforcing the rank order of prediction entropy across progressively masked samples [Han *et al.*, 2025]. Recently, PAID employs Householder reflections to update model weights while strictly preserving pairwise angular structures to maintain semantic consistency [Wang *et al.*, 2025].

3 Method

3.1 Problem Formulation and Overview

Problem Formulation. Let $D_s = \{(x_i, y_i)\}_{i=1}^{n}$ be the source domain consisting of n labeled samples, where each pair (x_i, y_i) corresponds to an input sample and its ground-truth label. A model $f(\cdot;\theta_s)$ is first trained on D_s, with θ_s denoting the source model parameters. During deployment, the model encounters an unlabeled target data stream organized as a sequence of target domains $\{D_t^\tau\}_{\tau=1}^{m}$, where $D_t^\tau = \{x_{t,i}^\tau\}_{i=1}^{n_\tau}$ contains n_τ samples drawn from the target distribution P_t^τ, and m is the total number of domains (or time steps). The goal of continuous test-time adaptation is to update the source-trained model $f(\cdot;\theta_s)$ to handle the continuously evolving target data $\{D_t^\tau\}_{\tau=1}^{m}$ in an online manner, without accessing the original source data D_s.

Overall Framework. We propose a dual-alignment method (D-Align) for continuous test-time adaptation, integrating correlation alignment with masked consistency alignment to achieve stable online adaptation. The correlation alignment branch aligns target features with a reliable pseudo-source formed from resilient samples in a class-wise memory bank, using prompt-driven nonlinear alignment to capture complex dependencies. Masked consistency alignment enforces prediction agreement between original and masked inputs, updating only normalization layers and prompts for efficient target adaptation. The overall framework of D-Align is illustrated in Fig. 1.

3.2 Correlation Alignment

The correlation alignment branch selects resilient pseudo-source samples, maintains them in a continual memory bank, and applies image-level prompt-driven nonlinear alignment to capture complex feature relationships.

Feature Resilience Criterion for Robust Pseudo-Source Sample Selection. In CTTA, selecting highly reliable samples from the data stream is essential for constructing a stable pseudo-source distribution. Here, we propose a new metric, Feature Resilience (FR). The key idea of FR is that a truly reliable sample should not only yield confident predictions but also maintain robustness under partial feature occlusion. By quantifying the stability of model predictions when salient features are masked, FR provides a more comprehensive measure of sample reliability, enabling more effective selection of highly dependable samples. Specifically, let the original image be divided into patches $x = (p_1, p_2, \ldots, p_{N_p})$, where N_p is the total number of patches, and the most salient patches are subsequently masked according to the self-attention scores of a ViT [Dosovitskiy *et al.*, 2021]:

$$\mathcal{A} = \sum_{h=1}^{H} \text{Softmax}\left(\frac{Q_{h,cls} {K_{h,img}}^\top}{\sqrt{d}}\right), \tag{1}$$

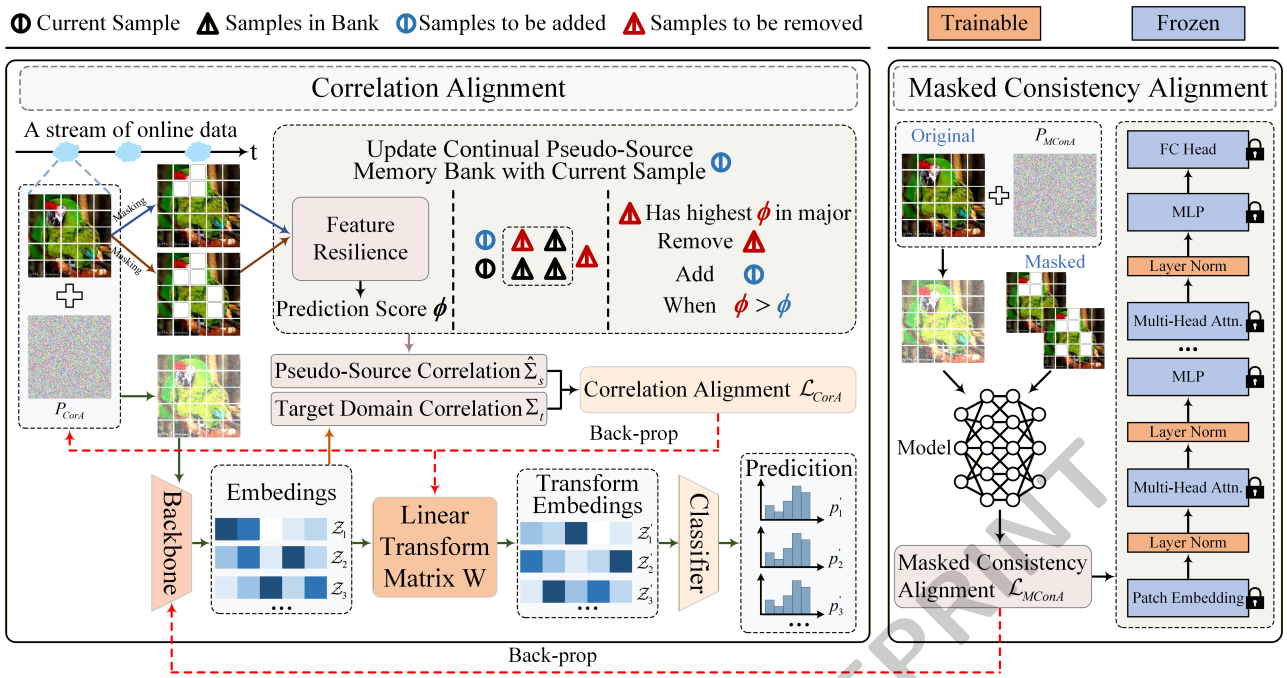

Figure 1: The framework of D-Align.

where $\mathcal{A} \in \mathbb{R}^{N_p}$ contains the score for each of the image patches, H is the number of heads in the multi-head attention, $Q_{h,cls}$ and $K_{h,img}$ are query for class token and key for image token of the h-th attention head, respectively, d denotes the dimension of each attention head. This step can identify the image regions that the model focuses on most when making predictions. Subsequently, we implement a progressive feature masking strategy guided by the score vector $\mathcal{A}$. The mask index set $\mathcal{I}_m$ is dynamically constructed as follows:

$$\mathcal{I}_m = \pi(j) \mid j = 1, 2, \ldots, \lfloor r_m \cdot N_p \rfloor, \tag{2}$$

where r_m denotes the masking ratio, π denotes a permutation of the image patch index set $\{1, \ldots, N_p\}$, which arranges the elements in descending order according to the values of $\mathcal{A}$, and $\lfloor \cdot \rfloor$ denotes the floor operation. This design ensures the prioritized masking of the image patches that the model attends to most, thereby systematically evaluating the model's robustness when critical discriminative features are absent. Next, we construct a binary masking vector $M \in \{0, 1\}^{N_p}$, where the i-th element M_i is defined as:

$$M_i = \begin{cases} 0, & \text{if } i \in \mathcal{I}_m \\ 1, & \text{otherwise} \end{cases}, \tag{3}$$

where 0 indicates that the image patch is removed and 1 indicates it is retained. The masked image x_m constructed using the mask M is given by:

$$x_m = (p_i \mid i \in 1, \ldots, N_p, M_i = 1), \tag{4}$$

we employed a multi-ratio masking strategy to obtain a more comprehensive stability assessment. A set of masked images $\{x_{m_1}, x_{m_2}, \ldots, x_{m_N}\}$ is obtained using different masking ratios. In the implementation, we construct masked images $\{x_{m_1}, x_{m_2}\}$ using only two masking ratios, $m_1 = 0.1$ and $m_2 = 0.2$. Based on the constructed masked images $\{x_{m_1}, x_{m_2}\}$, FR is calculated as follows:

$$FR = \frac{1}{K} \sum_{j=0}^{K} \left(\text{Ent}(f(x^i_{m_j})) - \bar{E}_i \right)^2, \tag{5}$$

where $x^i_{m_j}$ is the i-th sample of x_{m_j}. K is the number of masking ratios. $\text{Ent}(x) = -\sum_{c=1}^{C} \hat{p}_c(x) \log \hat{p}_c(x)$ denotes the entropy of the model's predictive distribution $\hat{p}(x)$ over C classes. $\bar{E}_i$ denotes the average entropy of the i-th sample under different masking ratios. A lower FR indicates that the model's uncertainty remains relatively stable even when certain key features are occluded, suggesting that the model has captured global and holistic representations rather than relying on localized features.

Continual Pseudo-Source Memory Bank. In the CTTA task, data arrives sequentially in an online manner as a stream of batches, requiring the model to generate predictions immediately for each batch without the opportunity to revisit previous samples. This inherent constraint precludes global access to the complete dataset for constructing a stable pseudo-source distribution, a fundamental prerequisite for the effective feature alignment demonstrated in TCA [You *et al.*, 2025]. While it is possible to construct local pseudo-source correlations within individual batches, such an approach is prone to unstable feature alignment and potential performance degradation, primarily due to statistical biases and limited sample diversity inherent in small batch sizes. To address this issue, we propose a Continually Pseudo-Source Memory Bank $\mathcal{S}_C$, which accumulates and retrieves reliable pseudo-source correlations across batches through a dynamic update mechanism, thereby facilitating robust feature alignment on

streaming data.

$$\mathcal{S}_c \leftarrow \begin{cases} \mathcal{S}_c \cup \{(z_i^c, \phi_i^c)\}, & \text{if } |\mathcal{S}_c| < V, \\ \mathcal{S}_c \setminus \{(z_{\text{worst}}^c, \phi_{\text{worst}}^c)\} \cup \{(z_i^c, \phi_i^c)\}, & \text{if } |\mathcal{S}_c| = V \\ & \text{and } \phi_i^c < \phi_{\text{worst}}^c, \\ \mathcal{S}_c, & \text{otherwise.} \end{cases} \tag{6}$$

$$\phi_i^c = \text{Ent}(f_t(x_t^i)) + FR, \tag{7}$$

where the output of the model $f_t(x_t^i)$ corresponds to class c. $\mathcal{S}_C = \{\mathcal{S}_0, \mathcal{S}_1, ..., \mathcal{S}_c, ..., \mathcal{S}_{C-1}\}$ denotes the partition of $\mathcal{S}_C$ into C mutually exclusive subsets. Each subset $\mathcal{S}_c$ independently stores and updates the pseudo-source samples of class c, maintaining class balance and diversity within the memory bank, thereby mitigating potential class bias during feature correlation alignment. Given the i-th sample, we first use the current model $f_t(\cdot)$ to obtain its pseudo label. If the predicted label corresponds to class c, the embedding of this sample is denoted as $z_i^c = \psi(x_t^i)$, and the associated prediction score is ϕ_i^c. $(z_{worst}^c, \phi_{worst}^c)$ denotes the sample with the highest score (i.e., lowest reliability) in $\mathcal{S}_c$. V denotes the capacity of $\mathcal{S}_c$. It is worth noting that the number of samples in $\mathcal{S}_C$ is kept very small. For instance, in the CIFAR-10-C task, with $V = 10$ per class, the memory bank holds only 100 samples in total, accounting for merely 1% of the dataset, which ensures low computational and storage overhead during CTTA.

At the initial stage, when a subset $\mathcal{S}_c \subset \mathcal{S}_C$ is not yet full, incoming samples are directly admitted until the capacity is reached, and correlation alignment is not applied. Once $\mathcal{S}_c$ becomes full, the update strategy switches to a replacement mechanism: each new sample (z_i^c, ϕ_i^c) is compared with the stored sample $(z_{\text{worst}}^c, s_{\text{worst}}^c)$ that has the highest score, and replacement occurs only if the new sample yields a lower score. By iteratively replacing suboptimal samples with higher-quality ones, the memory bank $\mathcal{S}_c$ progressively concentrates on reliable target-domain representations, allowing its feature distribution to converge toward the high-confidence region of class c. Correlation alignment is then performed once $\mathcal{S}_C$ satisfies the following condition:

$$\sum_{c=0}^{C-1} \mathbb{I}(|\mathcal{S}_c| = V) > \lfloor C/\tau \rfloor, \tag{8}$$

where $\mathbb{I}$ denotes the indicator function, which equals 1 if the condition is true and 0 otherwise, and τ represents the threshold. Once $\mathcal{S}_C$ is ready, its covariance matrix $\hat{\Sigma}_s$ is computed as the pseudo-source correlation. Correlation alignment is then achieved through a linear transformation matrix W, with the optimization objective defined as follows:

$$\min_W \|W^T \Sigma_t W - \hat{\Sigma}_s\|_F^2, \tag{9}$$

where Σ_t denotes the covariance matrix of target-domain samples in the current batch, and $\|\cdot\|_F$ represents the Frobenius norm.

Unfortunately, linear transformations exhibit inherent limitations in capturing the complex nonlinear relationships underlying real-world distribution shifts [You *et al.*, 2025]. To alleviate this, we propose a novel approach that achieves nonlinear alignment through image-level visual domain prompting, as detailed below.

Prompt-driven Nonlinear Alignment. The key insight is to learn differentiable perturbations directly in the input space, exploiting the powerful intrinsic nonlinear mapping capacity of the pre-trained backbone network to convert simple input-level signals into complex and expressive nonlinear transformations within the feature space. This enables effective nonlinear alignment without additional heavy training overhead. The parameters of these perturbations are optimized according to Eq. (9), as illustrated below:

$$\begin{cases} P_{\text{CorA}} = \mathbf{A}_C \cdot \mathbf{B}_C \\ x_{\text{CorA}}^t = x_t + P_{\text{CorA}} \end{cases}, \tag{10}$$

where $\mathbf{A}_C \in \mathbb{R}^{c \times L \times r}, \mathbf{B}_C \in \mathbb{R}^{c \times r \times L}$ are two low-rank matrices following the LoR-VP design proposed in [Jin *et al.*, 2025]. We adopt this formulation to further reduce the number of parameters in visual domain prompting and lower computational overhead. Here, c denotes the number of image channels, L represents the image size, and r is a hyperparameter satisfying $r << L$. Our visual domain prompting method learns perturbations in the input space and leverages the nonlinear mapping capability of the pre-trained backbone network to transform simple input perturbations into complex and highly nonlinear feature-space transformations. This effectively compensates for the limited representational capacity of linear transformations when handling complex nonlinear relationships, while introducing a small amount of additional computational overhead.

Building on the prompt-driven nonlinear alignment, we can learn a transformation W that effectively performs nonlinear alignment in the feature space. However, directly using the transformed features as the final representations may lead to the loss of target domain–specific information. To strike an appropriate balance between source-domain knowledge and target-domain adaptiveness, we introduce a homeostatic feature fusion strategy, which enhances adaptation robustness by integrating the transformed features with their original counterparts, as described below:

$$Z_t' = \alpha \cdot ((Z_t - \mu_t)W + \hat{\mu}_s) + (1 - \alpha) \cdot Z_t, \tag{11}$$

where Z_t' denotes the final set of output feature vectors, while $\hat{Z}_s = \{z_i | z_i \in \mathcal{S}_C\}$ and $Z_t = \{z_i | z_i \in B\}$ represent the sets of feature vectors from $\mathcal{S}_C$ and from a batch B, respectively. μ_t and $\hat{\mu}_s$ denote the mean embeddings of Z_t and $\hat{Z}_s$, respectively. α is a hyperparameter. Finally, the prediction results after performing correlation alignment through the Continual Pseudo-Source Memory Bank are given by $P_t' = g(Z_t')$, where $g(\cdot)$ denotes the classification head.

3.3 Masked Consistency Alignment

While correlation alignment aligns second-order statistics to correct distribution shifts, it does not explicitly encode target domain–specific information. We therefore incorporate domain prompt learning, in which a learnable visual prompt is applied to each target image x_t prior to alignment. By

parameterizing the prompt in a lightweight low-rank form, the model captures structured domain variations through conditional feature modulation, enabling more effective target-domain adaptation. Formally, the prompt is computed as:

$$\begin{cases} P_{\mathrm{MConA}} = \mathbf{A}_M \cdot \mathbf{B}_M \\ x^t_{\mathrm{MConA}} = x_t + P_{\mathrm{MConA}} \end{cases}, \tag{12}$$

where $\mathbf{A}_M \in \mathbb{R}^{c\times L\times r}$ and $\mathbf{B}_M \in \mathbb{R}^{c\times r\times L}$ are low-rank matrices defined analogously to those in Eq. (10). This learnable prompt functions as a compact memory module that captures and injects target-domain patterns into the model.

Building on the prompted image x^t_{MConA}, we introduce a masked consistency alignment mechanism to further enhance model adaptability. While the correlation alignment branch efficiently corrects feature statistics using a pseudo-source anchor, it lacks the ability to incorporate target-specific knowledge. To address this limitation, we adapt the normalization layers to the target-domain distribution through entropy minimization. Inspired by [Han *et al.*, 2025], prediction consistency is enforced between the prompted image and its masked variants. Given x^t_{MConA} and its corresponding masked samples $\{x_{m_1}, x_{m_2}\}$, the alignment mechanism is built upon two complementary components: the entropy consistency loss (ECL) and the entropy ranking loss (ERL), defined as follows:

$$\mathcal{L}_{ECL} = \sum_{i=1}^{2} \mathcal{H}(f(x_{m_i}), \mathbf{sg}(f(x_t))), \tag{13}$$

$$\mathcal{L}_{ERL} = \sum_{i<j}^{M_N} \max(0, \mathrm{Ent}(f(x_{m_i})) - \mathbf{sg}(\mathrm{Ent}(f(x_{m_j})))), \tag{14}$$

where $\mathcal{H}(p, q)$ denotes the cross-entropy between two probability distributions p and q, $f(x)$ represents the output of the model, $\mathbf{sg}(\cdot)$ is the stop-gradient operation, and $M_N = \{0, m_1, m_2\}$ denotes the set of mask ratios. ECL aligns the predictions of masked and original images, thereby implicitly reducing prediction entropy. ERL enforces that predictions on images with higher masking ratios must exhibit greater entropy than those with lower masking ratios. Together, ECL and ERL mitigate abrupt entropy collapse and alleviate overconfident predictions caused by unconstrained entropy minimization, while jointly optimizing the visual prompt to encode reliable target-domain knowledge. Through optimization via consistency alignment, D-Align effectively distills and injects target-domain knowledge into the model, thereby enabling adaptive performance under continuously evolving target distributions.

3.4 Optimization Objective

In D-Align, correlation alignment and masked consistency alignment are integrated in a sequential manner to form a unified adaptation framework. This phased design is motivated by the complementary properties of the two mechanisms: masked consistency alignment enables rapid adaptation to the target distribution by updating normalization parameters, while correlation alignment preserves model stability through feature distribution matching. Specifically, the optimization proceeds in two stages. In the first stage, the normalization parameters are updated via masked consistency alignment, with the objective

$$\mathcal{L}_{\mathrm{MConA}} = \mathcal{L}_{\mathrm{ECL}} + \mathcal{L}_{\mathrm{ERL}}, \tag{15}$$

which enforces prediction consistency between masked and original samples. In the second stage, correlation alignment is applied to the output embeddings by minimizing

$$\mathcal{L}_{\mathrm{CorA}} = \min_{W} \|W^\top \Sigma_t W - \hat{\Sigma}_s\|_F^2. \tag{16}$$

Together, this two-stage strategy achieves efficient online adaptation by balancing rapid target-domain adjustment with source-domain knowledge preservation, thereby improving robustness under continual distribution shifts.

4 Experiments

4.1 Experimental Setups

Datasets and Implementations. We evaluate our method on three standard classification CTTA benchmarks, including CIFAR10-to-CIFAR10C, CIFAR100-to-CIFAR100C, and ImageNet-to-ImageNet-C. We further test CLIP on VisDA-C, PACS, VLCS, and OfficeHome. Implementation details are in the supplementary material.

Comparison Methods. For the CTTA setting, we compare our proposed D-Align against ten competitive methods: Tent [Wang *et al.*, 2021], CoTTA [Wang *et al.*, 2022], VDP [Gan *et al.*, 2023], SAR [Niu *et al.*, 2023], PETAL [Brahma and Rai, 2023], ViDA [Liu *et al.*, 2024b], Continual-MAE [Liu *et al.*, 2024a], DPCore [Zhang *et al.*, 2025], REM [Han *et al.*, 2025], and PAID [Wang *et al.*, 2025]. For the TTA setting using CLIP, we evaluate D-Align against four representative methods: Tent [Wang *et al.*, 2021], TPT [Shu *et al.*, 2022], CLIPArTT [Hakim *et al.*, 2025], and WATT (including both WATT-P and WATT-S variants) [Osowiechi *et al.*, 2024].

4.2 Performance of CTTA under Standard Corruption

ImageNet-to-ImageNet-C. Using a source model pre-trained on ImageNet, we conduct CTTA by sequentially evaluating it on the 15 corruption types in ImageNet-C. As reported in Tab. 1, D-Align achieves state-of-the-art performance, obtaining an average error rate of 37.9%. This corresponds to a 17.9% improvement over the source model and a 1.3% gain over the previous best method, REM. To further assess the upper bound of adaptation capability, we compare against an ideal supervised oracle trained with target labels. The performance gap is only 2.2%, highlighting that our unsupervised framework is highly effective at leveraging unlabeled target data and achieves performance close to fully supervised adaptation. Furthermore, to evaluate our method on a benchmark that better reflects real-world physical properties, we conducted experiments on ImageNet-3DCC. The detailed results are provided in the supplementary material.

CIFAR10-to-CIFAR10C & CIFAR100-to-CIFAR100C. We further evaluate our method on two classic CIFAR10-to-CIFAR10C and CIFAR100-to-CIFAR100C scenarios. As

Time		$t \longrightarrow$																
Method	REF	gaussian	shot	impulse	defocus	glass	motion	zoom	snow	frost	fog	brightness	contrast	elastic_trans	pixelate	jpeg	Mean↓	Gain
Source [Dosovitskiy *et al.*, 2021]	ICLR'21	53.0	51.8	52.1	68.5	78.8	58.5	63.3	49.9	54.2	57.7	26.4	91.4	57.5	38.0	36.2	55.8	0.0
Tent [Wang *et al.*, 2021]	ICLR'21	52.2	48.9	49.2	65.8	73.0	54.5	58.4	44.0	47.7	50.3	23.9	72.8	55.7	34.4	33.9	51.0	+4.8
CoTTA [Wang *et al.*, 2022]	CVPR'22	52.9	51.6	51.4	68.3	78.1	57.1	62.0	48.2	52.7	55.3	25.9	90.0	56.4	36.4	35.2	54.8	+1.0
VDP [Gan *et al.*, 2023]	AAAI'23	52.7	51.6	50.1	58.1	70.2	56.1	58.1	42.1	46.1	45.8	23.6	70.4	54.9	34.5	36.1	50.0	+5.8
SAR [Niu *et al.*, 2023]	ICLR'23	49.3	43.8	44.9	58.2	60.9	46.1	51.8	41.3	44.1	41.8	23.8	57.2	49.9	32.9	32.7	45.2	+10.6
PETAL [Brahma and Rai, 2023]	CVPR'23	52.1	48.2	47.5	66.8	74.0	56.7	59.7	46.8	47.2	52.7	26.4	91.3	57.7	32.3	32.0	52.3	+3.5
ViDA [Liu *et al.*, 2024b]	ICLR'24	47.7	42.5	42.9	52.2	56.9	45.5	48.9	38.9	42.7	40.7	24.3	52.8	49.1	33.5	33.1	43.4	+12.4
Continual-MAE [Liu *et al.*, 2024a]	CVPR'24	46.3	41.9	42.5	51.4	54.9	43.3	**40.7**	**34.2**	35.8	64.3	23.4	60.3	**37.5**	29.2	31.4	42.5	+13.3
DPCore [Zhang *et al.*, 2025]	ICML'25	**42.2**	38.7	39.3	**47.2**	51.4	47.7	46.9	39.3	36.9	37.4	**22.0**	**44.4**	45.1	30.9	**29.6**	39.9	+15.9
REM [Han *et al.*, 2025]	ICML'25	43.5	38.1	39.2	53.2	49.0	43.5	42.6	37.5	**35.2**	35.4	23.2	46.8	41.6	28.9	30.2	39.2	+16.6
PAID [Wang *et al.*, 2025]	NeurIPS'25	48.8	43.7	44.4	49.4	49.6	47.3	44.2	37.5	39.4	42.1	25.2	50.0	39.3	35.5	36.5	42.2	+13.6
D-Align (Ours)	-	42.7	**38.0**	**38.7**	51.4	**45.9**	**41.3**	41.6	35.3	35.6	**33.9**	22.9	45.8	38.1	**27.6**	30.1	**37.9**	**+17.9**
Supervised	-	42.5	36.9	37.1	46.4	44.0	37.4	38.4	34.2	33.1	32.5	21.5	43.3	34.6	26.1	27.5	35.7	+20.1

Table 1: Classification error rate (%) for ImageNet-to-ImageNet-C online CTTA task. Mean (%) denotes the average error rate across 15 target domains. Gain (%) is the percentage of improvement in model accuracy compared with the source method.

Method	CIFAR10-to-CIFAR10C		CIFAR100-to-CIFAR100C	
	Mean↓	Gain	Mean↓	Gain
Source (ICLR'21)	28.2	0.0	35.4	0.0
Tent (ICLR'21)	23.5	+4.7	32.1	+3.3
CoTTA (CVPR'22)	24.6	+3.6	34.8	+0.6
VDP (AAAI'23)	24.1	+4.1	32.0	+3.4
SAR (ICLR'23)	26.6	+1.6	26.2	+9.2
PETAL (CVPR'23)	24.4	+3.8	28.0	+7.4
ViDA (ICLR'24)	20.7	+7.5	27.3	+8.1
Continual-MAE (CVPR'24)	12.6	+15.6	26.4	+9.0
DPCore (ICML'25)	15.4	+12.8	25.1	+10.3
REM (ICML'25)	9.4	+18.8	23.4	+12.0
PAID (NeurIPS'25)	11.0	+17.1	24.9	+10.5
D-Align (Ours)	**8.6**	**+19.6**	**21.8**	**+13.6**
Supervised	6.4	+21.8	16.1	+19.3

Table 2: Average error rate (%) for online CTTA on CIFAR10-C and CIFAR100-C.

Dataset	Domain	CLIP	Tents	TPT	CLIPArTT	WATT-P	WATT-S	D-Align
CIFAR	CIFAR-10	88.74	91.69	88.06	90.04	**91.41**	91.05	91.24±0.14
	CIFAR-10.1	83.25	87.60	81.80	86.35	**87.78**	86.98	86.65±0.04
	CIFAR-100	61.68	69.74	63.78	69.79	70.38	**70.74**	70.05±0.08
	Mean	77.89	83.01	77.65	82.06	**83.19**	82.92	82.65
VisDA-C	3D (trainset)	84.43	84.86	79.35	85.09	85.42	85.36	**87.19**±0.06
	YT (valset)	84.45	84.68	83.57	84.40	84.57	84.69	**84.87**±0.16
	Mean	84.44	84.77	81.46	84.75	85.00	85.03	**86.03**
Office-Home	Art	73.75	74.03	75.76	73.84	75.65	75.76	**78.90**±0.01
	Clipart	63.33	63.42	63.08	63.54	66.23	65.77	**66.58**±0.09
	Product	85.32	**85.51**	84.07	85.23	85.41	85.41	85.20±0.06
	RealWorld	87.71	87.74	85.89	87.61	88.22	**88.37**	88.18±0.04
	Mean	77.53	77.68	77.20	77.56	78.88	78.83	**79.71**
PACS	Art	96.34	96.65	95.52	96.57	96.31	96.39	**96.78**±0.03
	Cartoon	96.08	96.22	94.77	96.00	96.52	**96.62**	96.59±0.08
	Photo	99.34	99.40	99.42	99.28	99.48	**99.52**	**99.52**±0.12
	Sketch	82.85	82.96	83.22	83.93	**86.92**	86.65	86.18±0.05
	Mean	93.65	93.81	93.23	93.95	**94.81**	94.80	94.76
VLCS	Caltech101	**99.51**	**99.51**	99.36	**99.51**	99.43	**99.51**	**99.51**±0.09
	LabelMe	68.15	67.89	54.88	67.96	66.67	68.49	**69.95**±0.02
	SUN09	68.85	69.27	67.30	68.68	72.61	**73.13**	72.91±0.19
	VOC2007	84.13	84.42	76.74	84.09	83.30	83.41	**83.98**±0.05
	Mean	80.16	80.27	74.57	80.06	80.25	81.14	**81.47**

Table 3: Classification accuracy (%) comparison on vision-language model using CLIP ViT/B-32 across different datasets and domains.

shown in Tab. 2, D-Align achieves an average error rate of 8.6%, corresponding to a 19.6% improvement over the source model and a 0.8% gain over the previous best method, REM. Detailed fine-grained results are provided in the supplementary material. On the more challenging CIFAR100-to-CIFAR100-C benchmark, D-Align achieves a mean error of 21.8%. This represents a 13.6% reduction compared to the source model's 35.4% error. Moreover, our method achieves the best performance when compared against strong CTTA baselines, including Continual-MAE, PAID, and REM. Detailed results are provided in the supplementary material.

4.3 Performance of CLIP on Cross-Domain Benchmarks

Natural Domain Shift. As shown in Tab. 3, D-Align improves the average accuracy on CIFAR-10, CIFAR-10.1, and CIFAR-100 by 4.76% over the original model. Compared with the prior state-of-the-art method WATT, our approach achieves competitive results, demonstrating that D-Align is effective not only on ViT-Base architectures but also on vision–language models such as CLIP, thereby broadening its applicability.

Simulated and Video Shifts. To assess performance under large domain shifts from simulated environments to real-world video streams, we evaluate on the VisDA-C benchmark. As shown in Tab. 3, D-Align achieves notable gains. On the highly divergent synthetic 3D domain, it attains an accuracy of 87.19%, surpassing WATT-S (85.36%) by 1.83%. On the realistic YT (YouTube) video domain, D-Align maintains superior performance with a top accuracy of 84.87%. Overall, it achieves an average accuracy of 86.03%, outperforming WATT-S by 1.0%, highlighting robustness under complex synthetic-to-real domain shifts.

Texture and Style Shifts. We further evaluate adaptability under texture, artistic-style, and photographic shifts on Office-Home, PACS, and VLCS. On Office-Home, D-Align achieves 79.71% average accuracy, surpassing WATT-P (78.88%) with a notable gain on Art (78.90%). On PACS, it obtains 94.76%, comparable to WATT-P (94.81%). On VLCS, D-Align reaches 81.47%, improving over WATT-S (81.14%) by 0.33%. These results show that D-Align achieves leading or comparable performance across domain shift benchmarks, validating its generalization and stable CLIP adaptation.

Method	Parameters↓	GPU memory↓	Total Time↓	Number of FP	Total Models	Error
Source [Dosovitskiy *et al.*, 2021]	0	3012MB	19min39s	1	M_{src}	28.2
Tent [Wang *et al.*, 2021]	0.04M	12458MB	53min44s	1	M_{src}	23.5
CoTTA [Wang *et al.*, 2022]	86.4M	23024MB	241min58s	3	$M_{test} + M_{ema} + M_{src}$	24.6
ViDA [Liu *et al.*, 2024b]	93.7M	12870MB	521min56s	12	$M_{test} + M_{ema}$	20.7
Continual-MAE [Liu *et al.*, 2024a]	86.5M	14018MB	440min20s	12	$M_{test} + M_{ema}$	12.6
REM [Han *et al.*, 2025]	0.03M	24082MB	152min37s	3	M_{test}	9.4
CorA (ours)	0.60M	3286MB	32min27s	1	M_{src}	25.1
D-Align (ours)	0.63M	23982MB	159min48s	4	M_{test}	**8.6**

Table 4: Efficiency comparison. We report the number of trainable parameters, GPU memory usage, total runtime, number of forward passes (FP), total model count, and error rate (%) for CTTA on the CIFAR10-C dataset. The total runtime is measured on an NVIDIA RTX 3090 GPU across all 15 corruption types, with the batch size fixed at 12.

	Dataset	Source	CoTTA	CorA
CIFAR100-C	Gaussian Noise	55.0	55.0	**52.8**
	Shot noise	51.5	51.3	**49.6**
	Impulse Noise	26.9	**25.8**	26.3
	Defocus blur	24.0	24.1	**23.2**
	Glass blur	60.5	59.2	**57.3**
	Motion blur	29.0	28.9	**28.2**
	Zoom blur	21.4	21.4	**21.1**
	Snow	21.1	21.0	**20.9**
	Frost	25.0	24.7	**24.0**
	Fog	35.2	34.9	**33.4**
	Brightness	11.8	**11.7**	12.0
	Contrast	34.8	**31.7**	34.0
	Elastic transform	43.2	**40.4**	41.7
	Pixelate	56.0	55.7	**51.9**
	JPEG compression	35.9	35.6	**35.4**
Mean		35.4	34.8	**34.1**

Table 5: Classification error rate (%) of CorA on CIFAR100-C online CTTA tasks.

4.4 Analysis Experiment

Efficiency Analysis. To evaluate the computational efficiency of our method, we compare it with Tent, CoTTA, and recent advanced approaches such as ViDA and Continual-MAE, as shown in Tab. 4. First, regarding the CorA module, it completes the full CTTA pipeline in 32m27s, significantly outperforming CoTTA (241m58s) while maintaining comparable accuracy. In terms of memory usage, the additional overhead introduced by CorA is almost negligible (approximately 270 MB). This efficiency comes from our lightweight Continual Pseudo-Source Memory Bank, which stores only a very small subset of samples, and from the fact that feature alignment is performed solely through a linear transformation, without requiring costly backpropagation. These results demonstrate that CorA offers high computational efficiency. Secondly, D-Align offers clear advantages in terms of training time, the number of trainable parameters, and the amount of model state that must be stored. Unlike conventional approaches that rely on costly multiple forward passes for uncertainty estimation, D-Align strikes a better balance between speed and accuracy. It outperforms Continual-MAE by a margin of 31.75%, while requiring only 30% of the inference time and 0.7% of the trainable parameters. Overall, D-Align effectively avoids the burdens of full-model adaptation, offering a practical solution that balances high accuracy with minimal computational cost.

Ablation Study. To validate the effectiveness of our proposed correlation alignment, we conduct ablation experiments by freezing all model parameters and performing evaluations under the online CTTA setting without backpropagation on CIFAR100-C. As reported in Tab. 5, utilizing CorA alone yields a 1.3% performance improvement over the source model on CIFAR100-C. Notably, it surpasses CoTTA [Wang *et al.*, 2022] by 0.7%, thereby validating the efficacy of our proposed CorA module. Additionally, we provide results on CIFAR10-C, ImageNet-C, and ImageNet-3DCC in the supplementary material. Subsequently, we conducted a more detailed ablation study on the ImageNet-C dataset under the CTTA setting, as shown in Table 6. Specifically, we remove P_{CorA} (Ex_1) to evaluate the contribution of the Prompt-driven Nonlinear Alignment strategy within CorA. Compared with D-Align (Ex_0), Ex_1 exhibits a 0.3% performance drop, indicating that P_{CorA} effectively compensates for the limitations of purely linear alignment. Besides, we examine the effectiveness of P_{MConA} in Ex_2. The results show that removing P_{MConA} incurs a 0.4% performance drop, verifying its contribution. Subsequently, in Ex_3, we ablate the Feature Resilience (FR) criterion. This leads to a 0.7% decrease in accuracy, demonstrating that FR enhances the reliability of the pseudo-source memory and thereby facilitates better correlation alignment. Finally, disabling CorA in Ex_4 results in a substantial 1.2% performance degradation, further confirming the effectiveness of CorA.

	P_{CorA}	P_{MConA}	FR	CorA	Mean↓	Gain
Ex_0	✓	✓	✓	✓	**37.9**	0.0
Ex_1	-	✓	✓	✓	38.2	-0.3
Ex_2	✓	-	✓	✓	38.3	-0.4
Ex_3	✓	✓	-	✓	38.6	-0.7
Ex_4	✓	✓	✓	-	39.1	-1.2

Table 6: Ablation Study. Average error rate(%) for the ImageNet-to-ImageNet-C. Gain(%) represents the percentage of improvement in model accuracy compared with the source method.

5 Conclusion

In this work, we propose D-Align, a continual test-time adaptation framework that combines correlation alignment and consistency-based entropy minimization for efficient and stable adaptation. D-Align integrates feature resilience, a continual pseudo-source memory bank, and image-level visual prompts to construct robust pseudo-source correlations, and learns target-domain knowledge through stabilized consistency alignment. Extensive experiments on standard benchmarks demonstrate the effectiveness of D-Align under dynamic domain shifts.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China under Grants 62306008, 32501770 and 32472007, the Key Project of Anhui Province's Science and Technology Innovation Tackle Plan under Grant 202423k09020040, and the Natural Science Youth Project of the Education Department of Anhui Province under Grant 2025AHGXZK40156.

References

[Brahma and Rai, 2023] Dhanajit Brahma and Piyush Rai. A probabilistic framework for lifelong test-time adaptation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, pages 3582–3591. IEEE, 2023.

[Döbler *et al.*, 2023] 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.

[Dosovitskiy *et al.*, 2021] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021.

[Gan *et al.*, 2023] 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 Brian Williams, Yiling Chen, and Jennifer Neville, editors, *Thirty-Seventh AAAI Conference on Artificial Intelligence, Washington, DC, USA, February 7-14*, pages 7595–7603, 2023.

[Gong *et al.*, 2022] Taesik Gong, Jongheon Jeong, Taewon Kim, Yewon Kim, Jinwoo Shin, and Sung-Ju Lee. NOTE: robust continual test-time adaptation against temporal correlation. In Sanmi Koyejo, S. Mohamed, A. Agarwal, Danielle Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*, 2022.

[Hakim *et al.*, 2025] Gustavo A Vargas Hakim, David Osowiechi, Mehrdad Noori, Milad Cheraghalikhani, Ali Bahri, Moslem Yazdanpanah, Ismail Ben Ayed, and Christian Desrosiers. Clipartt: Adaptation of clip to new domains at test time. In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 7092–7101. IEEE, 2025.

[Han *et al.*, 2025] Jisu Han, Jaemin Na, and Wonjun Hwang. Ranked entropy minimization for continual test-time adaptation. *CoRR*, abs/2505.16441, 2025.

[Jin *et al.*, 2025] Can Jin, Ying Li, Mingyu Zhao, Shiyu Zhao, Zhenting Wang, Xiaoxiao He, Ligong Han, Tong Che, and Dimitris N Metaxas. Lor-vp: Low-rank visual prompting for efficient vision model adaptation. *arXiv preprint arXiv:2502.00896*, 2025.

[Liu *et al.*, 2024a] Jiaming Liu, Ran Xu, Senqiao Yang, Renrui Zhang, Qizhe Zhang, Zehui Chen, Yandong Guo, and Shanghang Zhang. Continual-mae: Adaptive distribution masked autoencoders for continual test-time adaptation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pages 28653–28663. IEEE, 2024.

[Liu *et al.*, 2024b] Jiaming Liu, Senqiao Yang, Peidong Jia, Renrui Zhang, Ming Lu, Yandong Guo, Wei Xue, and Shanghang Zhang. Vida: Homeostatic visual domain adapter for continual test time adaptation. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024.

[Niu *et al.*, 2022] 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.

[Niu *et al.*, 2023] Shuaicheng Niu, Jiaxiang Wu, Yifan Zhang, Zhiquan Wen, Yaofo Chen, Peilin Zhao, and Mingkui Tan. Towards stable test-time adaptation in dynamic wild world. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023.

[Osowiechi *et al.*, 2024] David Osowiechi, Mehrdad Noori, Gustavo Vargas Hakim, Moslem Yazdanpanah, Ali Bahri, Milad Cheraghalikhani, Sahar Dastani, Farzad Beizaee, Ismail Ayed, and Christian Desrosiers. Watt: Weight average test time adaptation of clip. *Advances in neural information processing systems*, 37:48015–48044, 2024.

[Park *et al.*, 2025] Hyewon Park, Hyejin Park, Jueun Ko, and Dongbo Min. Hybrid-tta: Continual test-time adaptation via dynamic domain shift detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2877–2886, 2025.

[Shu *et al.*, 2022] Manli Shu, Weili Nie, De-An Huang, Zhiding Yu, Tom Goldstein, Anima Anandkumar, and Chaowei Xiao. Test-time prompt tuning for zero-shot generalization in vision-language models. *Advances in Neural Information Processing Systems*, 35:14274–14289, 2022.

[Tarvainen and Valpola, 2017] Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. *Advances in neural information processing systems*, 30, 2017.

[Wang *et al.*, 2021] Dequan Wang, Evan Shelhamer, Shaoteng Liu, Bruno A. Olshausen, and Trevor Darrell. Tent: Fully test-time adaptation by entropy minimization. In *9th International Conference on Learning Representa-*

tions, ICLR 2021, Virtual Event, Austria, May 3-7, 2021. OpenReview.net, 2021.

[Wang *et al.*, 2022] Qin Wang, Olga Fink, Luc Van Gool, and Dengxin Dai. Continual test-time domain adaptation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 7191–7201. IEEE, 2022.

[Wang *et al.*, 2025] Kunyu Wang, Xueyang Fu, Yuanfei Bao, Chengjie Ge, Chengzhi Cao, Wei Zhai, and Zheng-Jun Zha. Paid: Pairwise angular-invariant decomposition for continual test-time adaptation. In *Advances in Neural Information Processing Systems, San Diego, CA, December 2-7*, 2025.

[You *et al.*, 2025] Linjing You, Jiabao Lu, and Xiayuan Huang. Test-time correlation alignment. In *Forty-second International Conference on Machine Learning, Vancouver, BC, Canada, July 13-19*, 2025.

[Yuan *et al.*, 2023] Longhui Yuan, Binhui Xie, and Shuang Li. Robust test-time adaptation in dynamic scenarios. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, pages 15922–15932. IEEE, 2023.

[Zhang *et al.*, 2025] Yunbei Zhang, Akshay Mehra, Shuaicheng Niu, and Jihun Hamm. Dpcore: Dynamic prompt coreset for continual test-time adaptation. In *Forty-second International Conference on Machine Learning, Vancouver, BC, Canada, July 13-19*, 2025.

[Zhou *et al.*, 2023] Kaiyang Zhou, Ziwei Liu, Yu Qiao, Tao Xiang, and Chen Change Loy. Domain generalization: A survey. *IEEE Trans. Pattern Anal. Mach. Intell.*, 45(4):4396–4415, 2023.