Dual Strategies for Test-Time Adaptation

Nam Nguyen Phuong[1] Duc Nguyen The Minh[1] Phi Le Nguyen[1*] Ehsan Abbasnejad[2] Minh Hoai[3]

[1] Institute for AI Innovation and Societal Impact, Hanoi University of Science and Technology
[2] Monash University [3] Adelaide University [*] Corresponding author

Abstract

Conventional test-time adaptation (TTA) approaches typically adapt the model using only a small fraction of test samples, often those with low-entropy predictions, thereby failing to fully leverage the available information in the test distribution. This paper introduces DualTTA, a novel framework that improves performance under distribution shifts by utilizing a larger and more diverse set of test samples. DualTTA identifies two distinct groups: one where the model's predictions are likely consistent with the underlying semantics, and another where predictions are likely incorrect. For the first group, it minimizes prediction entropy to reinforce reliable decisions; for the second, it maximizes entropy to suppress overconfident errors and unlearn spurious behavior. These groups are adaptively selected using a new reliability criterion that measures prediction stability under both semantic-preserving and semantic-altering transformations, addressing the limitations of purely entropy-based selection. We further provide theoretical analysis and empirical justification showing that our approach enables a tighter separation between reliable and unreliable samples - in the context of their suitability for adaptation - leading to provably more effective model updates. The source code is available at https://github.com/namk65hust/DualTTA.

1. Introduction

Deep learning models can struggle when training and test distributions differ, a challenge known as distribution shift. For example, in many computer vision applications, this occurs when statistical properties of the source and target data diverge due to factors like image corruption, lighting changes, or adverse weather [8, 11]. As deep networks are highly sensitive to such shifts, adapting models to test scenarios is crucial for maintaining reliability in real-world deployments. To address this issue, various approaches have been proposed, including domain generalization [26, 37, 38], domain adaptation [4, 23], test-time training [18, 25], and Test-Time Adaptation (TTA) [1, 13, 16, 19, 20, 28, 35, 36]. This paper focuses on TTA, a practical setting in which a trained model must adapt to an unseen target domain during inference without access to test data labels or the original training data.

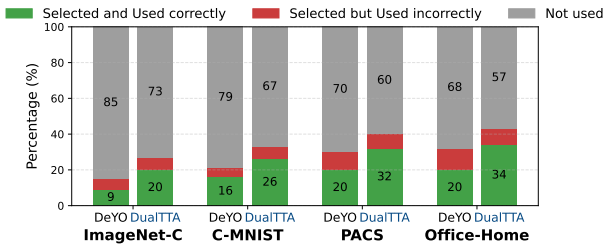

Figure 1. Comparison of sample utilization between DualTTA and the prior state-of-the-art method DeYO [13] across four datasets. DeYO adapts using only a small portion of test samples (green + red), with many of them (red) driving adaptation in the wrong direction. DualTTA, while not perfect, leverages a broader set of samples and misuses a much smaller fraction - achieving better data efficiency and more reliable adaptation.

Given the absence of labeled data, TTA typically updates the model's parameters by optimizing an unsupervised objective defined on selected test samples. The effectiveness of a TTA method depends critically on the choice of the objective function and the selection of samples for adaptation. One of the most widely adopted objectives in TTA is entropy minimization [22, 28], as low entropy correlates with the decisiveness of the model. A notable approach, TENT [28], adapts the model during testing by minimizing the entropy of its predictions on each test sample, encouraging the model to make more confident predictions. However, subsequent studies [13, 19] have shown that not all test samples are suitable for adaptation, and incorporating inappropriate samples can lead to model degradation. Motivated by this, several TTA methods have been proposed to focus on selecting the right samples for adaptation [13, 19, 20, 28]. For instance, EATA [19] and SAR [20] retain only test samples with low prediction entropy. More recently, DeYO [13] demonstrated that entropy alone is insufficient to filter out unsuitable samples and proposed an

additional criterion based on probability differences.

Existing TTA methods typically adapt using only a small subset of test samples deemed "reliable" based on low-entropy predictions, discarding most of the available data. This over-selective strategy suffers from two key failure modes: (i) *insufficient utilization*: the coverage is limited, as the pool of confident samples is small under significant distribution shifts; and (ii) *misguided adaptation*: confidence does not imply correctness, so high-confidence errors may be misclassified as reliable and, when used for adaptation, can reinforce spurious predictions. For example, DeYO [13], a state-of-the-art TTA method, adapts on only ≈14% of test samples (see Fig. 1), of which only 60-70% yield beneficial updates (about 9%). Relying solely on entropy minimization as the adaptation objective is thus both insufficient in scope and misguided in execution.

To address these limitations, we propose **DualTTA**, a new TTA framework that expands the effective adaptation set through a more principled partitioning of test samples into *likely-correct* and *likely-incorrect* groups. This partitioning is guided by a novel reliability criterion that combines prediction entropy with stability under both semantic-preserving and semantic-altering transformations. By going beyond raw confidence, this criterion better distinguishes informative samples from overconfident outliers, enabling more effective and robust adaptation.

Our paper introduces two core technical contributions. First, we propose a novel reliability criterion based on prediction stability under semantic-preserving and semantic-altering transformations. This criterion enables us to distinguish *likely-correct* samples - those with stable predictions under superficial changes but unstable under content changes - from *likely-incorrect* ones. Based on this classification, we maintain separate adaptation strategies for the two groups. Second, we introduce a dual-objective optimization for test-time adaptation: entropy minimization for *likely-correct* samples to reinforce accurate predictions, and entropy maximization for *likely-incorrect* ones to mitigate overconfidence and unlearn spurious patterns.

We further provide theoretical analysis showing that our stability-based criterion, combined with the dual-objective formulation, yields a sharper separation between reliable and unreliable samples and leads to better adaptation. Details of this proof will be presented fully in the Supplementary material section.

2. Related work

2.1. Test-Time Adaptation

TTA aims to enhance model performance in online settings by adapting to unlabeled test data without relying on original training data. Existing methods [12, 17, 20, 21, 28, 30, 34–36] mainly follow two approaches: (1) entropy minimization and (2) pseudo-label generation.

Entropy minimization. These methods assume that confident model predictions (low entropy) indicate better performance, and adapt the model using an unsupervised entropy-based loss on selected test samples. TENT [28] initiated this line, with follow-ups like EATA [19], which filtered high-entropy samples to avoid noisy gradients, and SAR [20], which minimized both entropy and loss sharpness using the SAM optimizer [5] to address small batch sizes and label imbalance. DeYO [13] further identified that low-entropy samples could still lack discriminability and proposed filtering such samples. While effective, these methods struggle to fully leverage the available target data.

Pseudo-label generation derives pseudo-labels for selected samples and uses them in adaptation loss. CoTTA [29] generated robust pseudo-labels via augmentation and multiple inferences. FATA [2] extended entropy-minimization methods by transforming feature-level representations of selected reliable samples. Though this approach improves sample utilization, it incurs high computational cost, limiting real-world applicability.

2.2. Prior use of transformation techniques

Several TTA methods incorporate semantic-altering perturbations, such as spatial transformations, into self-training frameworks [29], often assuming all augmented data is beneficial. This can lead to degraded performance when misleading samples are included. In contrast, our approach selectively identifies which perturbed samples to adapt and how. We use patch-shuffling from DeYO [13] to disrupt spatial structure while preserving texture, encouraging reliance on semantic features. Other spatial perturbations (e.g., cropping, jigsaw, deformation, adversarial augmentation [3, 33]) also help challenge spatial sensitivity.

Semantic-preserving transformations alter low-level features (e.g., color, texture) without changing class semantics, simulating domain shifts. Techniques such as AdaIN [9], FDA [32], and others [7, 10] support style adaptation. While past TTA work used such perturbations to enforce prediction consistency [24, 34], they often overlook harmful cases. Our method instead uses these perturbations to assess prediction stability, improving sample selection and adaptation.

Unlike prior work that uses transformations for regularization, we employ them in a dual-criterion framework for selecting and adapting samples, leading to more reliable and data-efficient test-time adaptation.

3. Proposed Method

In this section, we introduce DualTTA, our proposed method. We first outline the preliminaries of entropy minimization and discuss the limitations of existing sample se-

lection strategies used in entropy-based TTA methods.

3.1. Limitations of existing entropy-based methods

TTA considers the setting where a model f_θ, parameterized by θ, has been pre-trained on a source dataset $\mathcal{D}^{\text{tr}} = \{(\mathbf{x}_i^{\text{tr}}, y_i^{\text{tr}})\}_{i=1}^{N^{\text{tr}}}$ but both the training dataset $\mathcal{D}^{\text{tr}}$ and the labels y^{tst} of the test data $\mathcal{D}^{\text{tst}} = \{\mathbf{x}_i^{\text{tst}}\}_{i=1}^{N^{\text{tst}}}$ are unavailable during inference. Thus TTA methods must rely on unsupervised learning signals, with Shannon entropy minimization [22] being a popular approach. This approach optimizes the model to produce low-entropy predictions on selected test samples. Let $\hat{y}_c = f_\theta(c|\mathbf{x})$ be the probability of a sample $\mathbf{x}$ belonging to class c. The model's prediction uncertainty for $\mathbf{x}$ can be measured based on entropy: $\text{Ent}_\theta(\mathbf{x}) = -\sum_{c\in\mathcal{C}} \hat{y}_c \log \hat{y}_c$, where $\mathcal{C}$ denotes the set of class labels. Entropy-based TTA methods perform adaptation by optimizing:

$$\min_\theta \sum_{\mathbf{x}\in\mathcal{S}} \text{Ent}_\theta(\mathbf{x}), \tag{1}$$

where $\mathcal{S}$ is a subset of data encountered during test time. The performance of existing entropy-based TTA methods often depends on the choice of $\mathcal{S}$, and these methods have several limitations, as outlined below.

Suboptimality of the entropy criterion. Recent methods such as EATA [19], SAR [20] aim to improve adaptation quality by selecting only confident samples with low entropy: $\mathcal{S} = \{\mathbf{x} \in \mathcal{D}^{tst} \mid \text{Ent}_\theta(\mathbf{x}) < \tau_{\text{ent}}\}$, where τ_{ent} is a predefined entropy threshold. DeYO [13] further refines sample selection by incorporating structural cues. DeYO employs a sample selection strategy that reduces sensitivity to background information by applying content modifications through the patch-shuffling method. However, some samples depend on fine details or texture-related features for prediction (e.g., in domain-shift datasets like PACS or Office-Home). As a result, this approach may fail to filter out unreliable samples.

To empirically analyze this issue, we conducted an experiment to evaluate the quality of the samples selected for adaptation by comparing their ground-truth labels with the model's predicted labels, i.e., the "assumed" labels used by entropy-minimization methods during adaptation. The results in Fig. 1 show that only about 60-70% of the selected samples have predictions that match the ground truth. For example, in the PACS dataset, although 28% of test samples were used for adaptation, only 15% were correctly predicted. Including incorrectly predicted samples in the adaptation process, and further minimizing prediction entropy on them, can degrade performance.

Dependence on Entropy Minimization. Based on the idea of entropy minimization, current TTA methods devote significant effort to select a small "high-quality" subset of test samples and then apply entropy minimization exclusively on this set. While low-entropy samples often constitute around 25-30% of ImageNet-C samples, methods like DeYO utilize only 14%, discarding the remaining potentially useful data (see Fig. 1). Such aggressive filtering, though intended to ensure reliable updates, drastically reduces the number of samples available for adaptation. This sparse coverage restricts the model's exposure to diverse target-domain features particularly detrimental for underrepresented classes, and ultimately undermines the statistical strength and robustness of the adaptation process.

Addressing these limitations, we introduce a double-criterion selection mechanism based on consistency under semantic-preserving and semantic-altering transformations. This approach filters out samples whose predictions remain stable under content changes but vary with superficial changes. By combining these two transformation-based criteria, our strategy improves the reliability of selected samples beyond what raw entropy scores provide, while also recovering a significant portion of previously discarded data, leading to higher data efficiency.

3.2. Overview of DualTTA

Fig. 2 illustrates the overall pipeline of DualTTA. It uses two transformation functions, simulating **semantic-altering** and **semantic-preserving**, to generate pseudo-labels, which are then compared with the original predictions to guide sample selection. Semantic-altering, inspired by [13], involves dividing images into patches and shuffling them to disrupt content structure, while semantic-preserving operates in the latent space by only changing the mean and standard deviation of the distribution without affecting the semantic content. These transformations serve distinct purposes in generating pseudo-labels, which are subsequently compared with the original predictions. The differences between the predictions serve as a metric for selecting and processing samples during adaptation.

Samples whose prediction outputs remain stable under semantic-preserving transformations but exhibit significant variations under semantic-altering transformations are considered **likely-correct** and are denoted as $\mathcal{D}^+$. Conversely, samples whose prediction outputs change dramatically under semantic-preserving transformations while being less affected by semantic-altering transformations are considered **likely-incorrect** and are denoted as $\mathcal{D}^-$. DualTTA performs model adaptation by minimizing the entropy of the model's predictions for likely-correct samples $\mathcal{D}^+$ while maximizing the entropy of the model's predictions for likely-incorrect ones $\mathcal{D}^-$.

The rationale behind DualTTA is to adapt models more effectively by boosting confidence in likely-correct predictions while suppressing overconfidence in likely-incorrect ones. It distinguishes between these cases by analyz-

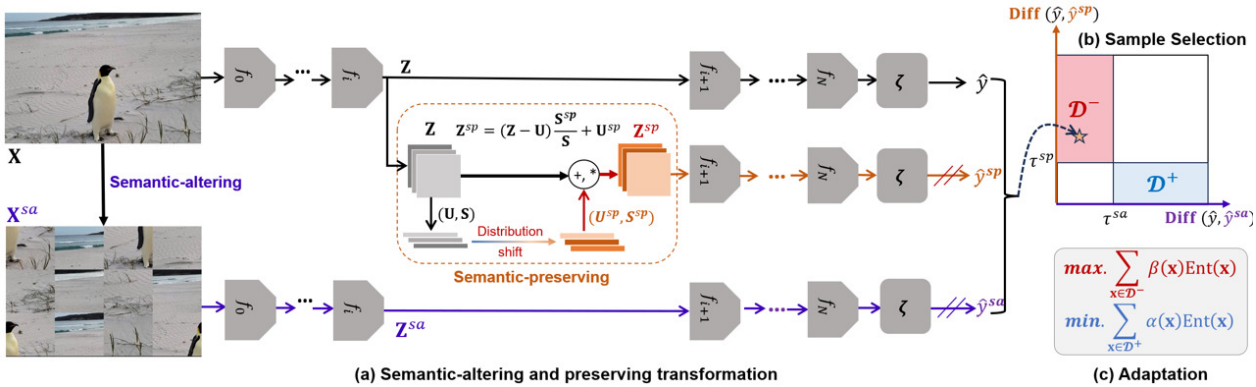

Figure 2. **Overview of DualTTA.** DualTTA apply 2 transformations: semantic-altering and semantic-preserving on each sample. The model's predictions for the original samples and its transformed variants are compared to determine reliability. Samples with stable predictions under semantic alteration but varying predictions under semantic preservation are classified as **likely-incorrect**, while those with unstable predictions under content-altering but stable under content-preserving are determined as **likely-correct**. likely-correct samples undergo entropy minimization, while likely-incorrect ones are penalized via entropy maximization to prevent adaptation degradation.

ing prediction behavior under semantic-preserving and semantic-altering transformations - reinforcing stable, semantically consistent predictions and penalizing unstable, potentially spurious ones. Unlike other TTA methods that rely solely on entropy minimization applied uniformly to low-entropy (i.e., confident) samples, DualTTA introduces a dual adaptation strategy: it explicitly selects separate sets of likely-correct and likely-incorrect test samples and treats them differently. This design not only moves beyond simplistic confidence-based selection by incorporating transformation-driven reliability but also mitigates the harm caused by overconfident mispredictions during adaptation.

3.3. Likely-correct and -incorrect Determination

We now describe our novel sample selection strategy, which evaluates prediction stability before and after applying input transformations. For a given input sample $\mathbf{x}$, we assume the existence of a *semantic-preserving* transformation yielding $\mathbf{x}^{sp}$ and a *semantic-altering* transformation yielding $\mathbf{x}^{sa}$. We will discuss such transformations in Sec. 3.5. Let $\hat{y}$, $\hat{y}^{sp}$, and $\hat{y}^{sa}$ denote the predicted probability vectors for the original input $\mathbf{x}$, its semantic-preserving variant, and its semantic-altering variant, respectively.

We define the difference between two probability vectors $\hat{y}$ and y as:

$$\text{Diff}(\hat{y}, y) = \hat{y}_k - y_k, \text{where } k = \underset{c}{\text{argmax}}\, \hat{y}_c. \tag{2}$$

From the test set $\mathcal{D}^{tst}$, we create two subsets $\mathcal{D}^+, \mathcal{D}^-$ for the likely-correct and -incorrect samples:

$$\mathcal{D}^+ = \{\mathbf{x} \in \mathcal{D}^{tst} | \text{Diff}(\hat{y}, \hat{y}^{sa}) > \tau^{sa}, \text{Diff}(\hat{y}, \hat{y}^{sp}) < \tau^{sp}\},$$
$$\mathcal{D}^- = \{\mathbf{x} \in \mathcal{D}^{tst} | \text{Diff}(\hat{y}, \hat{y}^{sa}) < \tau^{sa}, \text{Diff}(\hat{y}, \hat{y}^{sp}) > \tau^{sp}\},$$

where τ^{sa} and τ^{sp} are pre-defined thresholds. The subset $\mathcal{D}^+$ comprises samples with stable predictions under semantic-preservation but significant changes when structural content is perturbed. In contrast, the subset $\mathcal{D}^-$ includes samples that are highly sensitive to superficial changes while exhibiting minimal prediction variations when structural content is altered.

3.4. Loss function for adaptation

For TTA, we propose to minimize the following loss:

$$\mathcal{L}_{Dual} = \mathcal{L}^+ - \lambda \mathcal{L}^-, \text{where} \tag{3}$$
$$\mathcal{L}^+ = \sum_{\mathbf{x} \in \mathcal{D}^+} \alpha(\mathbf{x})\text{Ent}(\mathbf{x}), \text{and } \mathcal{L}^- = \sum_{\mathbf{x} \in \mathcal{D}^-} \beta(\mathbf{x})\text{Ent}(\mathbf{x}).$$

In the above, λ is the trade-off coefficient between the two loss terms, and $\alpha(\mathbf{x})$ and $\beta(\mathbf{x})$ are samples weight that reflects the model's confidence in the sample, as well as the degree of alignment/misalignment between the model and the sample $\mathbf{x}$ under semantic-altering and semantic-preserving transformations. Specifically:

$$\alpha(\mathbf{x}) = e^{\text{Ent}_0 - \text{Ent}(\mathbf{x})} + e^{\text{Diff}(\hat{y}, \hat{y}^{sa})} + e^{\text{Diff}_0 - \text{Diff}(\hat{y}, \hat{y}^{sp})}, \tag{4}$$
$$\beta(\mathbf{x}) = e^{\text{Ent}_0 - \text{Ent}(\mathbf{x})}, \tag{5}$$

where $\text{Ent}_0, \text{Diff}_0$ are pre-defined normalization factors. The sample weight $\alpha(\mathbf{x})$ is higher when the model is confident in its prediction (low entropy), the difference between the model's predictions for the original and content-altered samples is large, and the difference between the predictions for the original and content-preserved samples is small. Then, θ will be updated based on $\mathcal{L}_{Dual}$ for adaptation: $\theta = \theta - \nabla_\theta \mathcal{L}_{Dual}$.

3.5. Image Transformations

Our DualTTA framework relies on semantic-preserving and semantic-altering transformations, but it is not limited to any specific ones, such transformations are generally easy to construct. For image data, operations that significantly disrupt spatial arrangement (e.g., shuffling image patches) typically alter the object category and serve as semantic-altering transformations. In contrast, modifications that slightly adjust pixel intensities or make minor spatial changes without altering the overall structure are considered semantic-preserving.

Importantly, these transformations can be applied not only at the input level but also at intermediate stages of the model (e.g., feature maps at one of the first layers of backbone), offering flexibility in their design. In this paper, we adopt patch shuffling, following DeYO [13], as our semantic-altering transformation, which disrupts spatial structure while preserving texture, thereby encouraging reliance on high-level semantic features and allowing fair comparison with DeYO. In the remainder of this section, we describe the specific semantic-preserving transformation used in our experiments, while noting that other variants are equally applicable.

Given a batch of test samples $\mathbf{X} \in \mathbb{R}^{B\times C\times H\times W}$ and a pre-trained model f_θ composed of N layers $f_1, f_2, ..., f_N$. The feature map $\mathbf{Z}_i \in \mathbb{R}^{B\times C_i\times H_i\times W_i}$ at the output of the first i layers is obtained by sequentially applying these layers to $\mathbf{X}$: $\mathbf{Z}_i = f_i \circ f_{i-1} \circ \cdots \circ f_1(\mathbf{X})$. We denote $\mathbf{U} \in \mathbb{R}^{B\times C_i}$ and $\mathbf{S} \in \mathbb{R}^{B\times C_i}$ as the channel-wise mean and standard deviation of each instance in the batch, respectively, which is defined as follows:

$$\mathbf{U}(b,c) = \frac{1}{H_i W_i}\sum_{h=1}^{H_i}\sum_{w=1}^{W_i}\mathbf{Z}(b,c,h,w),$$

$$\mathbf{S}(b,c) = \sqrt{\frac{1}{H_i W_i}\sum_{h=1}^{H_i}\sum_{w=1}^{W_i}(\mathbf{Z}(b,c,h,w)-\mathbf{U}(b,c))^2}.$$

From an abstract perspective, feature statistics encapsulate key characteristics of a given domain, such as color, texture, and contrast, often referred to as style statistics in previous studies [9, 14]. In out-of-distribution scenarios, these statistics frequently diverge from those in the training data due to inherent domain differences [6, 31].

We simulate the distribution shift process while maintaining the semantic information to identify samples within the batch that are sensitive to domain shifts. Specifically, this is achieved by modifying their style statistics from $(\mathbf{U}, \mathbf{S})$ to $(\mathbf{U}^{sp}, \mathbf{S}^{sp})$ according to the following formula, where the subscript i is omitted for brevity.

$$\mathbf{U}^{sp} = \mathbf{U} + \epsilon_{\mathbf{U}}\mathbf{U}^{\sigma}, \text{and } \mathbf{S}^{sp} = \mathbf{S} + \epsilon_{\mathbf{S}}\mathbf{S}^{\sigma}, \tag{6}$$

where $\epsilon_{\mathbf{U}}, \epsilon_{\mathbf{S}} \in \mathbb{R}^{B\times 1}$ are random Gaussian noises, and $\mathbf{U}^{\sigma}, \mathbf{S}^{\sigma} \in \mathbb{R}^{1\times C_i}$ are the standard deviation for the entries in $\mathbf{U}$ and $\mathbf{S}$ across the batch dimension, respectively. More formally, $\mathbf{U}^{\sigma}$ which is defined as:

$$\mathbf{U}^{\sigma}(c) = \sqrt{\frac{1}{B}\sum_{b=1}^{B}\left(\mathbf{U}(b,c) - \frac{1}{B}\sum_{b=1}^{B}\mathbf{U}(b,c)\right)^2}. \tag{7}$$

The matrix $\mathbf{S}^{\sigma}$ is defined similarly. The semantic-preserving transformation is done as follows:

$$\mathbf{Z}^{sp}(b,c,h,w) = (\mathbf{Z}(b,c,h,w) - \mathbf{U}(b,c))\frac{\mathbf{S}^{sp}(b,c)}{\mathbf{S}(b,c)} + \mathbf{U}^{sp}(b,c).$$

After transformation, the new feature $\mathbf{Z}^{sp}$ is fed into the next layer of the network for making probability predictions $\hat{y}^{sp} \in \mathbb{R}^{B*|\mathcal{C}|}$, where $|\mathcal{C}|$ denotes total number of classes, i.e., $\hat{y}^{sp} = \zeta \circ f_N \circ \cdots \circ f_{i+1}(\mathbf{Z}^{sp})$.

4. Experiments

We conducted extensive experiments across diverse benchmarks to evaluate the effectiveness, robustness, and data efficiency of DualTTA, comparing it with state-of-the-art test-time adaptation methods and analyzing the contribution of dual strategies and likely-incorrect sample.

4.1. Settings

Benchmark datasets. We conduct experiments on commonly used benchmarks covering three out-of-distribution scenarios: spurious correlation, domain shift, and data corruption. To evaluate model performance under extreme spurious correlation shifts, we use ColoredMNIST and Waterbirds [13]. For domain shift, we test on PACS [15] and Office-Home [27]. To measure robustness against data corruption, we employ ImageNet-C [8], containing 15 types of corruption with five levels of severity.

Backbone and normalization. Experiments are conducted using architectures that integrate three distinct normalization techniques: Batch Normalization (BN), Group Normalization (GN), and Layer Normalization (LN). Specifically, ResNet-18 and ResNet-50 are employed for BN and GN, while the ViT-Base model was utilized for LN. For details, we utilize ResNet18-BN for ColoredMNIST, ResNet50-BN for Waterbirds, PACS, Office-Home and ResNet50-GN, ViT-B-LN for ImageNet-C.

Test scenarios. For all datasets, we follow the mild scenario proposed by [28]. For ImageNet-C, we also experiment with two additional test scenarios suggested by [20] with imbalanced label shift and mixed distribution, under the highest level of corruption (Level 5).

Pretrained models. For ImageNet-C, all models are pretrained on the ImageNet dataset and subsequently adapted

during the testing phase. For the PACS, Office-Home datasets, the models are initially pre-trained on a source domain and later adapted to other target domains at test time.

Baseline and hyper-parameters. We evaluate the performance of DualTTA against state-of-the-art TTA methods, including TENT [28], SAR [20], EATA [19], and DeYO [13], using a consistent batch size of 64 across all experiments. For ResNet-18 and ResNet-50, we set the learning rate to 0.5×10^{-3}, while for ViT-B, we use 10^{-4}.

For DualTTA, we set the semantic-altering threshold to $\tau^{sa} = 0.4$ and define the semantic-preserving threshold as $\tau^{sp} = 0.7$, the difference normalization factor $\text{Diff}_0 = 0.7$, and the trade-off coefficient $\lambda = 0.5$. Additionally, we apply semantic-preserving transformations at layer $i = 1$ for ResNet and $i = 7$ for ViT-B. Further explanation will be presented in the Appendix.

4.2. Comparison with the state-of-the-art

4.2.1. Performance under spurious correlation shifts.

The proposed method, DualTTA, demonstrates superior performance on datasets exhibiting spurious correlations between objects and backgrounds. Since the spurious correlation dataset contains $\mathcal{C}$ labels based on both background and object, it is divided into $2\mathcal{C}$ groups. Avg Acc and Worst Acc represent the average accuracy and the lowest accuracy across these $2\mathcal{C}$ groups, respectively. Specifically, Table 1 shows that DualTTA achieves an average accuracy improvement of 4.54% and 1.37% over the second-best method, DeYO, on the ColoredMNIST and Waterbirds datasets.

4.2.2. Performance under domain shift.

Table 1 presents the average accuracy when the model is pre-trained on one of the four domains and then used for inference on the remaining three domains in the Office-Home and PACS dataset.

For the Office-Home dataset consisting of four domains: Art, Clipart, Product, and Real-world, performance improves by an average of 2.43% compared to DeYO. See Supplementary material for more details, where performance improvements reach up to 6.02% when the model is trained on Art and tested with TTA on Clipart.

On the PACS dataset comprising four domains: Art, Cartoon, Photo, and Sketch performance improves by an average of 0.96% compared to the second-best method, DeYO. Supplementary material provides more details, showing that performance gains can reach up to 7.02% when the model is trained on Photo and tested with TTA on Sketch.

4.2.3. Performance under corruption shift.

To ensure reproducibility and assess stability, we run all experiments three times using a fixed random seed of 2024 and report the mean and standard deviation of the resulting accuracies. Table 2 presents the accuracy when the model is pre-trained on ImageNet and tested on 15 types of level-5 corruptions using different backbone architectures: ResNet50-BN, ResNet50-GN and ViT-B-LN. DualTTA achieves improvements of up to **8.06%** and 2.65% in average over the second-best method, EATA, on ResNet-50-BN. On ViT-B-LN, DualTTA improves up to 7.52% and 0.56% in average over DeYO .

Table 1. Performance of DualTTA and baselines across several benchmarks: ColoredMNIST, Waterbirds, Office-Home, PACS.

Methods	ColoredMNIST		Waterbirds		Office-Home	PACS
	Avg Acc	Worst Acc	Avg Acc	Worst Acc		
No adapt	63.46	20.14	81.81	62.90	59.14	72.34
Tent	56.51	8.99	83.63	54.99	61.08	74.96
SAR	58.10	11.82	82.91	53.59	59.82	73.26
EATA	60.65	17.59	80.65	52.18	52.34	72.46
DeYO	77.98	65.59	87.17	71.98	59.08	75.16
DualTTA	**82.12**	**68.82**	**88.44**	**72.53**	**61.51**	**76.02**

To better reflect challenging test conditions in real-world deployment, SAR [20] introduces two more realistic test protocols: (i) dynamic changes in the ground-truth label distribution during testing, resulting in label imbalance across different corruptions; and (ii) the presence of concurrent distribution shifts. The results in Table 3 show that DualTTA consistently achieves strong performance across all settings and architectures. Under the imbalanced label condition, DualTTA outperforms the second-best baseline SAR by 1.10% on ViTBase-LN. In mixed shift setting, DualTTA significantly outperforms prior methods: 1.41% on ResNet-BN vs SAR, and 1.17% on ViTBase-LN vs DeYO.

4.3. Ablation Studies

4.3.1. Dual-transformation sample selection.

To understand the importance of combining semantic-altering and semantic-preserving transformations, we conduct an experiment comparing the performance of DualTTA with variations that exclude one of these transformations as a criterion for filtering samples. The experiment is conducted on three datasets: ColoredMNIST, PACS, and OfficeHome. As shown in Table 4, the semantic-altering condition plays a crucial role in ColoredMNIST, while the semantic-preserving transformation significantly enhances performance on PACS and OfficeHome. In more detail, applying the semantic-altering condition improves performance by 8.49% and 4.19% on ColoredMNIST and Waterbirds, respectively. Similarly, using semantic-preserving condition improves performance by 3.05% on ColoredMNIST, 1.64% on average across PACS, and 2.81% on average across OfficeHome.

We also provide an empirical justification for these 2 selection strategies, by evaluating their accuracy relative to the ground truth (% correct and % incorrect) across four quadrants on 50000 samples from ImageNet-C. As illustrated in

Table 2. Model performance on ImageNet-C with corruption level 5. The best results are colored **bold red**. DualTTA achieves the best performance on most corruption types, showing strong robustness under severe distribution shifts with different normalization layers.

Methods	Noise			Blur				Weather				Digital				Avg
	Gauss.	Shot	Impul.	Defoc.	Glass	Motion	Zoom	Snow	Frost	Fog	Bright.	Contr.	Elastic	Pixel	JPEG	
ResNet50-BN	0.30	0.37	0.35	0.24	0.21	0.43	0.66	1.06	1.30	1.67	2.29	0.86	0.66	0.73	0.84	0.81
+ Tent	18.32	21.70	18.54	18.71	18.57	33.58	44.91	45.74	46.09	61.78	70.54	31.67	48.35	51.39	51.55	38.76
+ SAR	16.60	20.56	19.93	13.74	12.86	34.01	45.33	46.11	45.66	61.78	70.34	32.36	48.43	52.58	48.92	37.95
+ EATA	**24.96**	**28.51**	26.43	20.28	21.84	36.95	46.84	48.27	47.91	63.18	70.47	35.74	51.12	53.47	52.09	41.87
+ DeYO	17.09	25.72	26.74	19.21	24.14	33.70	44.21	45.97	48.02	59.84	69.86	**41.25**	51.90	57.14	54.69	41.30
+ DualTTA	23.69	24.38	**29.59**	**20.30**	**25.97**	**42.33**	**51.13**	**54.03**	**50.76**	**65.77**	**71.90**	34.81	**56.64**	**58.79**	**57.77**	**44.52**
ResNet50-GN	22.09	23.03	22.04	19.79	11.40	21.46	25.04	40.28	46.97	34.02	68.81	36.25	18.51	29.24	52.60	31.44
+ Tent	8.91	9.57	8.71	9.22	7.24	12.08	14.75	12.80	13.36	1.31	69.90	40.27	3.37	49.51	52.45	20.90
+ SAR	**39.96**	42.08	41.35	19.35	22.01	37.65	39.12	24.89	46.87	54.32	72.37	49.37	5.82	54.89	57.31	40.49
+ EATA	37.42	41.49	39.64	**29.70**	**27.01**	37.90	**41.82**	**51.54**	**47.76**	**58.30**	71.64	51.66	26.76	59.18	58.88	**45.38**
+ DeYO	36.00	46.13	**46.99**	14.73	20.42	12.58	17.41	27.09	26.98	33.42	66.33	41.98	**22.56**	44.01	50.70	33.82
+ DualTTA	25.96	**46.31**	44.43	24.81	26.12	**42.06**	18.92	23.27	25.75	20.73	**72.96**	**53.80**	21.55	**61.99**	**60.17**	41.83
ViTBase-LN	35.09	32.16	35.87	31.42	25.31	39.45	31.55	24.47	30.13	54.74	64.48	48.98	34.20	53.17	56.45	39.83
+ Tent	52.58	51.58	53.55	52.75	47.82	56.47	48.01	10.00	31.78	67.38	74.28	67.19	50.79	66.75	64.73	53.04
+ SAR	51.92	51.41	52.89	51.62	48.59	55.45	49.61	12.84	49.90	66.79	73.03	65.62	52.62	63.90	63.15	53.96
+ EATA	**55.87**	**56.14**	**56.96**	**57.36**	53.28	62.00	**58.63**	61.99	59.87	71.47	75.54	**68.66**	63.23	69.20	66.57	62.44
+ DeYO	49.77	53.27	54.71	48.79	50.17	55.87	51.82	57.64	61.09	64.78	75.67	62.34	58.48	67.41	67.12	58.59
+ DualTTA	54.45	55.32	55.47	55.74	**54.54**	**62.01**	57.48	**62.58**	**67.39**	**71.29**	**76.85**	67.55	**64.78**	**71.32**	**68.05**	**63.00**

Table 3. Average accuracy on ImageNet-C at severity level 5 under imbalanced label distribution and mixed distribution.

	Imbalanced label			Mixed shift		
Methods	ResNet-BN	ResNet-GN	ViTBase-LN	ResNet-BN	ResNet-GN	ViTBase-LN
No adapt	32.11	31.44	39.83	0.14	31.44	54.37
Tent	42.80	29.08	52.47	2.35	13.16	52.82
SAR	46.27	**45.67**	62.10	26.10	35.09	54.33
EATA	**47.14**	34.86	53.60	19.62	38.50	58.64
DeYO	41.03	41.54	58.68	11.63	**39.26**	69.73
DualTTA	43.83	38.88	**63.30**	**27.51**	38.92	**70.90**

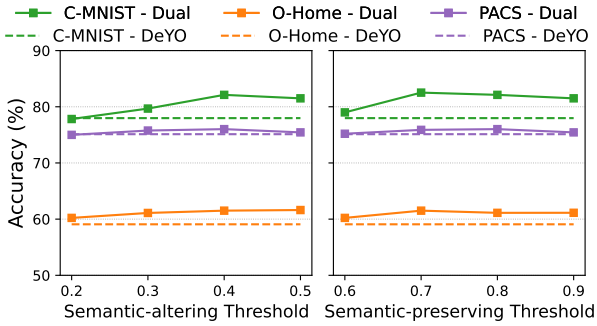

Figure 3. Sensitivity of DualTTA to semantic-altering and semantic-preserving thresholds on C-MNIST, O-Home, and PACS. DualTTA maintains stable performance across a wide range of values, showing low sensitivity to these parameters. DeYO [13] is shown for reference and is outperformed by DualTTA.

Fig. 4, $\mathcal{D}^+$ is dominated by samples with predictions that correctly match the ground truth, at 71% vs 29% incorrect, while $\mathcal{D}^-$is primarily composed of samples with incorrect predictions, at 82% vs 18% correct.

4.3.2. Impacts of utilizing likely-incorrect samples.

To evaluate the effectiveness of our proposed sample selection method, particularly the use of likely-incorrect samples, we compare our approach with a variant that only minimizes entropy on the likely-correct samples $\mathcal{D}^+$. The results in Table 6 show that the number of samples used

Table 4. Impact of dual transformation filtering of DualTTA on different datasets. Each column indicates whether semantic-altering or semantic-preserving transformations are used (✓) or not (✗). Results show that employing both transformations consistently yields the highest accuracy, highlighting the roles of semantic-altering and semantic-preserving filters in DualTTA.

DualTTA on	Semantic-Altering	Semantic-Preserving	Acc
ColoredMNIST	✓	✓	**82.12**
	✗	✓	74.63
	✓	✗	79.17
PACS	✓	✓	**76.02**
	✗	✓	75.80
	✓	✗	74.48
OfficeHome	✓	✓	**61.51**
	✗	✓	61.20
	✓	✗	59.70

Table 5. Impact of utilizing likely-incorrect samples on different methods. The additional column indicates whether the method incorporates likely-incorrect samples (✓) or not (✗). Results show that selectively leveraging these samples improves performance across all methods, confirming their complementary role of entropy maximization in test-time adaptation.

Original Method	Likely-Incorrect	Colored MNIST	PACS	Office-Home
DualTTA (proposed)	✓	**82.12**	**76.02**	**61.51**
DualTTA (ablated)	✗	80.65	75.52	61.10
DeYO (original)	✗	77.98	75.16	59.08
DeYO (improved)	✓	79.17	75.65	59.70
EATA (original)	✗	60.65	72.46	52.34
EATA (improved)	✓	62.31	73.47	55.72

Table 6. Data efficiency analysis. *"%corr-adapt"* represents the share of likely-correct samples whose predictions match the ground truth and likely-incorrect samples whose predictions differ from the ground truth. *"% adapt"* represents the share of test sets for adaptation.

Data efficiency analysis						
Method	ColoredMNIST		Office-Home		ImageNet-C	
	%adapt	%corr-adapt	%adapt	%corr-adapt	%adapt	%corr-adapt
EATA	15.6	10.6	29.7	14.5	11.7	7.0
SAR	19.6	12.1	31.8	18.9	13.1	8.2
DeYO	18.8	14.4	31.7	19.9	13.7	9.2
DualTTA	**33.1**	**27.3**	**42.6**	**33.6**	**25.8**	**19.8**

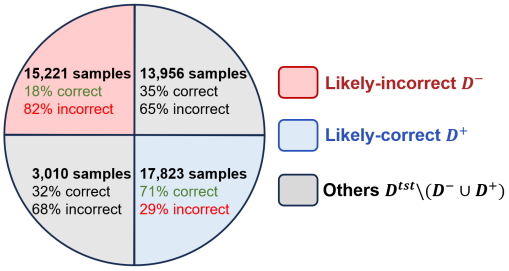

Figure 4. Quadrant-based accuracy analysis of 50000 impulse noise samples from ImageNet-C using our selection strategies. The blue region (**likely-correct** $\mathcal{D}^+$) contains a majority of correct samples, while the red region (**likely-incorrect** $\mathcal{D}^-$) is dominated by incorrect predictions, demonstrating that DualTTA effectively separates high- and low-accuracy predictions.

for the adaptation process in the proposed method TTA is greater than that of all baseline methods DeYO and EATA. The performance of DualTTA is also improved 1.57% when leveraging knowledge from likely-incorrect samples.

We further evaluate the impact of our proposed dual-optimization strategy by integrating it into existing baselines that utilize entropy-based sample selection, including EATA [19] and DeYO [13]. To achieve this, we define likely-correct and likely-incorrect sample sets based on the selection criteria established by these methods. For DeYO, likely-correct samples are defined as:

$$\mathcal{D}^+_{DeYO} = \{\mathbf{x} \mid \text{Ent}(\mathbf{x}) < \tau_{Ent},\ \text{PLPD}(\mathbf{x}, \mathbf{x}') > \tau_{\text{PLPD}}\},$$

where $\mathbf{x}'$ represents a perturbed version of $\mathbf{x}$; the likely-incorrect samples are defined as:

$$\mathcal{D}^-_{DeYO} = \{\mathbf{x} \mid \text{Ent}(\mathbf{x}) < \tau_{Ent},\ \text{PLPD}(\mathbf{x}, \mathbf{x}') < \tau_{\text{PLPD}}/2\},$$

For EATA, the likely-correct sample set is given by:

$$\mathcal{D}^+_{EATA} = \{\mathbf{x} \mid \text{Ent}(\mathbf{x}) < \tau_{Ent}, \cos(\theta(\mathbf{x}), \mathbf{n}^{t-1}) < \epsilon\},$$

$$\text{where} \quad \mathbf{n}^t = \begin{cases} \bar{y}^1, & \text{if } t = 1 \\ \alpha \bar{y}^t + (1-\alpha)\mathbf{n}^{t-1}, & \text{if } t > 1 \end{cases}$$

Here, $\bar{y}^t$ represents the average model prediction for a batch of B test samples at iteration t. The likely-incorrect samples are defined as:

$$\mathcal{D}^-_{EATA} = \{\mathbf{x} \mid \text{Ent}(\mathbf{x}) < \tau_{Ent}, \cos(\theta(\mathbf{x}), \mathbf{n}^{t-1}) > \frac{3}{2}\epsilon\}.$$

Table 5 shows that across all experimental scenarios, incorporating adaptation with likely-incorrect samples using our proposed method consistently enhances the performance of both DeYO and EATA compared to their original versions. The maximum performance improvement achieved by leveraging likely-incorrect samples reaches up to 1.19% for DeYO and 3.48% for EATA, highlighting the effectiveness of our approach.

4.3.3. Impact of semantic altering, preserving threshold.

We conduct experiments with different values of τ^{sa} and τ^{sp} across four datasets: ColoredMNIST, Waterbirds, PACS, and Office-Home, presented in Fig. 3. We find that the model's performance remains relatively stable despite variations in τ^{sa} and τ^{sp}. Notably, regardless of the threshold value, DualTTA consistently outperforms DeYO. To improve generalizability, we set $\tau^{sa} = 0.4$, $\tau^{sp} = 0.7$ for all experimental scenarios. See the Supplementary for analysis on other hyper parameters.

4.3.4. Efficient adaptation analysis.

Table 6 presents an analysis of adaptation efficiency across TTA methods, using two key metrics: %adapt, which indicates the proportion of test samples selected for adaptation, and %corr-adapt, which measures the accuracy of the selected samples. The visualization has been shown in Fig. 1.

Across all benchmarks, DualTTA substantially outperforms the baselines in both sample coverage and adaptation quality. On ImageNet-C, it adapts to 25.8% of test samples, compared with 13.7% of DeYO, while the share of correctly adapted samples rises to 19.8%, versus 9.2%. A similar gap is observed on Office-Home, where DualTTA improves coverage from 31.7% to 42.6% and correct-adapt from 19.9% to 33.6%.

5. Conclusions

We introduced DualTTA, an innovative framework designed to harness a larger and more diverse set of test samples for adaptation. Our approach pioneers a novel criterion that uses prediction stability to classify test samples as either likely-correct or likely-incorrect, based on their response to semantic-altering and semantic-preserving transformations. We proposed a dual optimization strategy: reinforcing alignment for likely-correct samples through entropy minimization while applying corrective adjustments for likely-incorrect samples via entropy maximization. Extensive experiments across multiple datasets demonstrated that DualTTA surpasses state-of-the-art TTA methods, achieving a significant performance boost, with a gap of up to **8.06%** over the second-best.

Acknowledgments. This project was initiated with support from the University of Adelaide's Global Engagement Fund. Subsequent support was provided in part by the Australian Institute for Machine Learning (University of Adelaide), the Centre for Augmented Reasoning (an initiative of the Department of Education, Australian Government), and Hanoi University of Science and Technology grant number T2024-TD-002.

References

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

[2] Younggeol Cho, Youngrae Kim, Junho Yoon, Seunghoon Hong, and Dongman Lee. Feature augmentation based test-time adaptation. In *Proceedings of the IEEE Workshop on Applications of Computer Vision*, 2025. 2

[3] Ekin D Cubuk, Barret Zoph, Jonathon Shlens, and Quoc V Le. Randaugment: Practical automated data augmentation with a reduced search space. In *Advances in Neural Information Processing Systems*, 2020. 2

[4] Abolfazl Farahani, Sahar Voghoei, Khaled Rasheed, and Hamid R. Arabnia. A brief review of domain adaptation. *arXiv preprint arXiv:2010.03978*, 2020. 1

[5] Pierre Foret, Ariel Kleiner, Hossein Mobahi, and Behnam Neyshabur. Sharpness-aware minimization for efficiently improving generalization. In *Proceedings of International Conference on Learning and Representation*, 2021. 2

[6] Shang-Hua Gao, Qi Han, Duo Li, Ming-Ming Cheng, and Pai Peng. Representative batch normalization with feature calibration. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2021. 5

[7] Leon A Gatys, Alexander S Ecker, and Matthias Bethge. Image style transfer using convolutional neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016. 2

[8] Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. In *Proceedings of International Conference on Learning and Representation*, 2019. 1, 5

[9] Xun Huang and Serge Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *Proceedings of the International Conference on Computer Vision*, 2017. 2, 5

[10] Philip T Jackson, Amir Atapour-Abarghouei, Stephen Bonner, Toby P Breckon, and Boguslaw Obara. Style augmentation: Data augmentation via style randomization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019. 2

[11] Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard L. Phillips, Irena Gao, et al. Wilds: A benchmark of in-the-wild distribution shifts. In *Proceedings of the International Conference on Machine Learning*, 2021. 1

[12] Jungsoo Lee, Debasmit Das, Jaegul Choo, and Sungha Choi. Towards open-set test-time adaptation utilizing the wisdom of crowds in entropy minimization. In *Proceedings of the International Conference on Computer Vision*, 2023. 2

[13] Jonghyun Lee, Dahuin Jung, Saehyung Lee, Junsung Park, Juhyeon Shin, Uiwon Hwang, and Sungroh Yoon. Entropy is not enough for test-time adaptation: From the perspective of disentangled factors. In *Proceedings of International Conference on Learning and Representation*, 2024. 1, 2, 3, 5, 6, 7, 8

[14] Boyi Li, Felix Wu, Ser-Nam Lim, Serge Belongie, and Kilian Q. Weinberger. On feature normalization and data augmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2021. 5

[15] Da Li, Yongxin Yang, Yi-Zhe Song, and Timothy M. Hospedales. Deeper, Broader and Artier Domain Generalization. In *Proceedings of the International Conference on Computer Vision*, 2017. 5

[16] Jian Liang, Ran He, and Tieniu Tan. A Comprehensive Survey on Test-Time Adaptation under Distribution Shifts. *International Journal of Computer Vision*, 133(1):31–64, 2024. 1

[17] Hyesu Lim, Byeonggeun Kim, Jaegul Choo, and Sungha Choi. Ttn: A domain-shift aware batch normalization in test-time adaptation. In *Proceedings of International Conference on Learning and Representation*, 2022. 2

[18] Yuejiang Liu, Parth Kothari, Bastien van Delft, Baptiste Bellot-Gurlet, Taylor Mordan, and Alexandre Alahi. Ttt++: When does self-supervised test-time training fail or thrive? In *Advances in Neural Information Processing Systems*, 2021. 1

[19] Shuaicheng Niu, Jiaxiang Wu, Yifan Zhang, Yaofo Chen, Shijian Zheng, Peilin Zhao, and Mingkui Tan. Efficient test-time model adaptation without forgetting. In *Proceedings of the International Conference on Machine Learning*, 2022. 1, 2, 3, 6, 8

[20] 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 *Proceedings of International Conference on Learning and Representation*, 2023. 1, 2, 3, 5, 6

[21] Sunghyun Park, Seunghan Yang, Jaegul Choo, and Sungrack Yun. Label shift adapter for test-time adaptation under covariate and label shifts. In *Proceedings of the International Conference on Computer Vision*, 2023. 2

[22] Claude Elwood Shannon. A mathematical theory of communication. *The Bell System Technical Journal*, 1948. 1, 3

[23] Shiliang Sun, Honglei Shi, and Yuanbin Wu. A survey of multi-source domain adaptation. *Information Fusion*, 24:84–92, 2015. 1

[24] Yu Sun, Xiaolong Wang, Zhuang Liu, John Miller, Alexei Efros, and Moritz Hardt. Test-time training with self-supervision for generalization under distribution shifts, 2020. 2

[25] Yu Sun, Xiaolong Wang, Zhuang Liu, John Miller, Alexei Efros, and Moritz Hardt. Test-time training with self-supervision for generalization under distribution shifts. In

Proceedings of the International Conference on Machine Learning, 2020. 1

[26] Nam Duong Tran, Nam Nguyen Phuong, Hieu H Pham, Phi Le Nguyen, and My T Thai. Conststyle: Robust domain generalization with unified style transformation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3174–3183, 2025. 1

[27] Hemanth Venkateswara, Jose Eusebio, Shayok Chakraborty, and Sethuraman Panchanathan. Deep Hashing Network for Unsupervised Domain Adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017. 5

[28] Dequan Wang, Evan Shelhamer, Shaoteng Liu, Bruno Olshausen, and Trevor Darrell. Tent: Fully test-time adaptation by entropy minimization. In *Proceedings of International Conference on Learning and Representation*, 2021. 1, 2, 5, 6

[29] Qin Wang, Olga Fink, Luc Van Gool, and Dengxin Dai. Continual test-time domain adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2022. 2

[30] Shuai Wang, Daoan Zhang, Zipei Yan, Jianguo Zhang, and Rui Li. Feature alignment and uniformity for test time adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2023. 2

[31] Ximei Wang, Ying Jin, Mingsheng Long, Jianmin Wang, and Michael I. Jordan. Transferable normalization: Towards improving transferability of deep neural networks. In *Advances in Neural Information Processing Systems*, 2019. 5

[32] Yanchao Yang and Stefano Soatto. Fda: Fourier domain adaptation for semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2020. 2

[33] Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoon Yoo. Cutmix: Regularization strategy to train strong classifiers with localizable features. In *Proceedings of the International Conference on Computer Vision*, 2019. 2

[34] 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. 2

[35] Zekun Zhang and Minh Hoai. Object detection with self-supervised scene adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2023. 1

[36] Zekun Zhang, Vu Quang Truong, and Minh Hoai. Efficiency-preserving scene-adaptive object detection. In *Proceedings of the British Machine Vision Conference*, 2024. 1, 2

[37] Kaiyang Zhou, Yongxin Yang, Yu Qiao, and Tao Xiang. Domain generalization with mixstyle. In *Proceedings of International Conference on Learning and Representation*, 2021. 1

[38] Kaiyang Zhou, Ziwei Liu, Yu Qiao, Tao Xiang, and Chen Change Loy. Domain generalization: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022. 1