TWO-STAGE ISOTROPIC WHITENING FOR SOURCE-FREE TEST-TIME ADAPTATION

Anonymous authors
Paper under double-blind review

ABSTRACT

When deploying machine learning models in real-world scenarios, distribution shift poses a critical challenge. Test data often deviates from training distributions and severely degrades model performance. This issue becomes especially acute in source-free test-time adaptation (SF-TTA) because models must adapt to unlabeled target data without source data or annotations. To address this challenge, we introduce Two-Stage Isotropic Whitening (TSIW) to facilitate target feature learning. TSIW integrates Whitening Batch Normalization (WBN) with Whitening Contrastive Learning (WCL). At the feature level, WBN enforces isotropic feature distributions via ZCA whitening. This suppresses domain-specific covariance structures and improves stability under distribution shifts. At the representation level, WCL extends standard contrastive learning through global feature whitening. This eliminates redundant feature correlations and optimizes representation geometry on a hypersphere to preserve semantic relationships. Through this two-stage process, WBN handles low-level feature standardization while WCL refines global representation geometry. Together, they yield highly generalizable features. Extensive experiments demonstrate that TSIW achieves state-of-the-art performance on major benchmarks, including VisDA-C, DomainNet-126, ImageNet-C, and CIFAR-100-C.

1 INTRODUCTION

Deep learning models typically assume that training and test data follow similar distributions Zou et al. (2023); Zhao et al. (2019). This assumption often fails in real-world deployment due to changes in environments, acquisition devices, and data quality He et al. (2016); Long et al. (2015). Such distribution shifts can cause severe performance degradation on unseen target domains. Domain adaptation aims to address this issue by transferring knowledge from labeled source data to target domains Ning et al. (2025); Li et al. (2026). However, conventional methods require access to source data during adaptation. This is often impractical due to privacy, storage, and computational constraints. Test-time adaptation (TTA) Chen et al. (2026); Han et al. (2026) provides an alternative by adapting a pre-trained model with unlabeled target data during inference. Source-free TTA (SF-TTA) further assumes that only the pre-trained source model and unlabeled target samples are available at test time. This setting improves practical applicability while making reliable adaptation under distribution shifts more challenging.

Existing TTA methods Honarmand et al. (2025) mainly improve target-domain performance by revising model predictions or adapting feature representations. For example, Tent Wang et al. (2021b) minimizes the entropy of target predictions by updating normalization parameters, while other approaches exploit prototype adjustment Iwasawa & Matsuo (2021), information maximization Liang et al. (2020), or batch normalization statistics Li et al. (2018) to improve the consistency of target predictions. Contrastive learning has also been introduced into TTA to construct more discriminative target representations. AdaContrast Chen et al. (2022) combines contrastive learning with pseudo-label refinement to progressively adapt the source model to unlabeled target samples. ADAPT Zhang et al. (2026) leverages a backpropagation-free test-time adaptation framework that models class-conditional feature distributions under a Gaussian assumption. Although these methods have demonstrated promising performance, they primarily focus on prediction uncertainty including class-level alignment or instance-level discrimination. They generally overlook a fundamental prop-

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

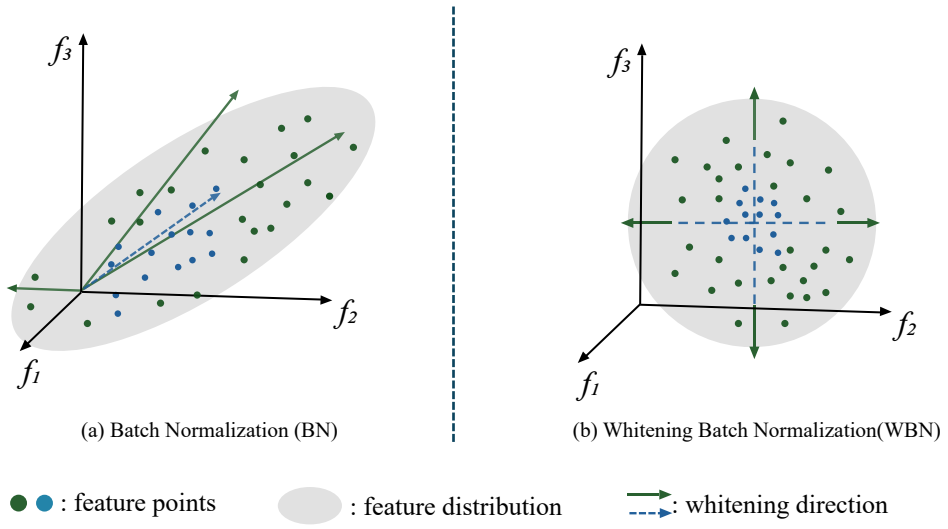


Figure 1: Illustrations of multiple normalization methods on centered data. (a) BN performs standardization by stretching/squeezing the data along the axes, such that each dimension has a unit variance. (b) Whitening BN performs ZCA whitening by stretching/squeezing the data along the eigenvectors, such that the covariance matrix is identical.

erty of the learned feature space that the target representations exhibit undesirable correlations and domain-specific covariance structures.

Feature distributions in deep neural networks are often highly anisotropic. Different feature dimensions may exhibit strong correlations, and these covariance structures can vary across domains. Such domain-dependent correlations may introduce redundant and domain-specific information into learned representations. This makes the model sensitive to distribution shifts. Therefore, we revisit SF-TTA from the perspective of feature whitening. Feature whitening removes correlations and promotes an isotropic feature distribution. Unlike standard Batch Normalization (BN), which independently normalizes each feature dimension. Whitening Batch Normalization (WBN) Huang et al. (2018) further decorrelates features through ZCA whitening. As illustrated in Figure 1, BN performs normalization along the original feature axes, while WBN performs whitening along the eigenvectors of the covariance matrix. This allows WBN to suppress domain-specific second-order statistics and provide a more stable feature space for target-domain adaptation.

However, feature-level whitening alone does not explicitly regulate the global geometry of the representation space. Although WBN removes redundant feature correlations, the resulting representations may still exhibit unfavorable spatial structures. This can limit semantic discrimination and consistency. We therefore introduce a second whitening mechanism at the representation level through Whitening Contrastive Learning (WCL) Ermolov et al. (2021). WCL reduces redundant dependencies and encourages representations to form a well-structured hyperspherical space while preserving semantic relationships between positive samples. Thus, WCL complements WBN by further refining the global representation geometry. Based on these observations, we propose Two-Stage Isotropic Whitening (TSIW) for source-free test-time adaptation. TSIW first applies WBN to intermediate target features and then employs WCL to refine the global representations. The two stages address different aspects of distribution shift. WBN focuses on feature-level distribution and suppresses domain-specific covariance structures. WCL focuses on representation-level geometry and promotes a compact and well-organized embedding space. This progressive whitening process enables robust target adaptation without access to source data or target labels. Our main contributions can be summarized as follows:

- We address source-free test-time adaptation (SF-TTA) by removing feature redundancy and domain-specific covariance structures, reducing reliance on domain-dependent correlations.

- We introduce TSIW, combining feature-level Whitening Batch Normalization (WBN) to standardize low-level features and representation-level Whitening Contrastive Learning (WCL) to optimize global hyperspherical geometry without representation collapse.
- We conduct extensive experiments on VisDA-C Peng et al. (2017), DomainNet-126 Peng et al. (2019), ImageNet-C Hendrycks & Dietterich (2019), and CIFAR-100-C Hendrycks & Dietterich (2019). TSIW consistently outperforms existing SF-TTA methods across different types of distribution shifts and achieves state-of-the-art performance on most benchmarks.

2 METHOD

2.1 MOTIVATION AND PRELIMINARIES

This work presents a extensive investigation of feature whitening for TTA Liang et al. (2025); Karmanov et al. (2024); Farina et al. (2024), a largely underexplored perspective for addressing distribution shifts. While conventional TTA methods primarily focus on statistical alignment or entropy minimization, we identify feature whitening as a powerful but neglected mechanism to address the fundamental challenge of distribution shift. However, these approaches largely overlook the geometric structure of the learned feature space, where feature correlations and domain-dependent covariance structures can introduce redundant and domain-specific information. We therefore revisit SF-TTA from the perspective of feature whitening and propose Two-Stage Isotropic Whitening, which progressively adjusts feature distributions at both the feature transformation and representation levels. In the first stage, WBN performs batch whitening at the feature transformation level to promote isotropic feature distributions. Unlike standard BN, which independently normalizes each feature dimension, WBN additionally decorrelates feature dimensions by whitening their covariance structure. Specifically, it estimates the batch mean and covariance, computes a whitening transformation and projects the intermediate activations into a decorrelated feature space. This operation reduces the dependence on domain-specific second-order statistics and alleviates redundant feature correlations, thereby providing a more stable feature space for target-domain adaptation. In contrast to BN, which normalizes features along the original coordinate axes, WBN operates on the covariance structure of the feature space and transforms correlated features into a more isotropic distribution. In the second stage, WCL further adjusts the global geometry of the latent representation space. WCL introduces a whitening constraint that encourages the covariance of the learned embeddings to approach the identity matrix, thereby preventing dimensional collapse and reducing redundant dependencies. Meanwhile, a positive-pair consistency objective encourages augmented views of the same sample to remain close in the embedding space. These two complementary objectives enable WCL to preserve semantic relationships while promoting a well-structured and approximately hyperspherical representation space. Thus, WBN focuses on removing domain-dependent correlations at the feature transformation level and WCL refines the global geometry of the learned representation.

TSIW is inspired by previous whitening methods such as DBN Huang et al. (2018) and WSS Ermolov et al. (2021), but is specifically designed for the practical constraints of SF-TTA. Unlike existing whitening approaches developed primarily for supervised representation learning or conventional training settings, TSIW operates without access to source data or target labels during adaptation. Moreover, it is designed for online adaptation under limited target batches, where covariance estimation can be unstable, especially when the batch size is small. To address these challenges, WBN updates target-domain statistics through moving averages, while WCL employs stable covariance estimation and whitening constraints to maintain reliable representation geometry. By progressively whitening intermediate features and global representations, TSIW reduces domain-specific feature dependencies while preserving semantic structure, leading to more robust and generalizable target representations under distribution shifts.

We are tackling the problem of closed-set test-time adaptation in image recognition, where the adaptation process does not utilize any source data. The source model is trained on pairs of images and labels $\{x_i^s, y_i^s\}_{i=1}^{n_s} \in \mathcal{D}_s$, where $x_i^s \in \mathcal{X}_s$ and $y_i^s \in \mathcal{Y}_s$. Our objective is to adapt the trained source model to unlabeled target data $\{x_i^t\}_{i=1}^{n_t} \in \mathcal{X}_t$, while the corresponding labels $\{y_i^t\}_{i=1}^{n_t} \in \mathcal{Y}_t$ are only accessible for evaluation. In the closed-set scenario, both the source and target domains have the same label space $\mathcal{Y}_s = \mathcal{Y}_t = \mathcal{Y}$. This assumption differentiates our work from more challenging open-set or partial-set adaptation scenarios, where the target domain may contain novel categories not

present in the source domain. The preservation of identical label spaces enables reliable knowledge transfer while maintaining well-defined evaluation protocols.

2.2 WHITENING BATCH NORMALIZATION (WBN)

2.2.1 BATCH NORMALIZATION

For a Batch Normalization (BN) layer that takes as its input a batch of D-dimensional vectors $X = (x_1, \cdots, x_B) \in \mathcal{R}^{D \times B}$, its output is a batch of vectors $Y = (y_1 \cdots, y_B) \in \mathcal{R}^{D \times B}$, computed as follows

$$y_{i,k} = \frac{x_{i,k} - \mu_k}{\sqrt{\sigma_k^2 + \epsilon}} \cdot \gamma_k + \beta_k \tag{1}$$

where γ and β are learnable affine parameters, ϵ is a small constant originally proposed for numerical stability. In training time, μ_k, σ_k^2 are mean and variance computed over the k-th row of the input batch X, and in inference time, running estimations from training time is used.

Dimensional collapse can harm feature utility and demands explicit mitigation. By definition, it is associated with strong correlations across feature dimensions. Under distribution shifts, such redundant correlations often encode domain-specific noise, making target representations sensitive to environmental variation. To restore representation geometry without relying on source domain, we propose a feature-level whitening that explicitly decorrelates dimensions, encouraging a spherical and highly discriminative feature space during adaptation.

2.2.2 DOMAIN-SPECIFIC WHITENING TRANSFORMATION

As previously mentioned, BN involves standardizing each sample $\mathbf{x}_i \in B$ based on its dimensions. However, this means that even after normalization, the features of batch samples may still be correlated. We aim to address the domain-shift problem with feature normalization, and therefore argue that simple standardization is insufficient for aligning the source and target marginal distributions. To better achieve this alignment, we leverage Batch Whitening (BW) as an alternative to BN. Essentially, BW involves transforming the data to have zero means and unit variances, while also decorrelating the features via principal component analysis.

$$\textit{BW}(x_{i,k}; \Omega) = \gamma_k \hat{x}_{i,k} + \beta_k, \tag{2}$$

$$\hat{\mathbf{x}}_i = W_B(\mathbf{x}_i - \boldsymbol{\mu}_B). \tag{3}$$

where $\boldsymbol{\mu}_B$ is the mean of the elements in batch B and the matrix W_B is determined such that $W_B^\top W_B = \Sigma_B^{-1}$, where Σ_B is the covariance matrix computed using B. $\Omega = (\boldsymbol{\mu}_B, \Sigma_B)$ represents the first and second-order statistics specific to the batch. Eq. 3 is utilized to whiten $\mathbf{x}_i$, resulting in a set of vectors $\hat{B} = \{\hat{\mathbf{x}}_1, ..., \hat{\mathbf{x}}_m\}$ that are distributed spherically.

Our network takes as input the batches of target data, randomly extracted from $\mathcal{X}_t$. Specifically, given any arbitrary layer l in the network, let $B^t = \{\mathbf{x}_1^t, ..., \mathbf{x}_k^t\}$ denote the batch of intermediate output activations from layer l. Using Eq. 2 and Eq. 3, we can now define our WBN. Let x^s denote the inputs to the WBN layer from the target domain. Our WBN is defined as follows

$$\textit{WBN}(x^t; \Omega^t) = BW(x^t, \Omega^t). \tag{4}$$

B^t and use them for whitening the corresponding activations, projecting the batches into a common spherical distribution.

To compute W_B^t, we adopt the method proposed in Roy et al. (2019), which employs Cholesky decomposition Higham (1990). Unlike the approach taken in Huang et al. (2018), we eliminate the coloring step following the whitening process and instead apply simple scale and shift operations to avoid introducing additional network parameters. Additionally, we utilize feature grouping to enhance the accuracy of batch-statistics estimation when the number of samples m is small and the feature dimension d is large, distinguishing our method from Huang et al. (2018). During training, the WBN layers continually update the target domain statistics through a moving average of the batch statistics.

2.3 WHITENING CONTRASTIVE LEARNING (WCL)

To further optimize the spatial structure of target representations, we introduce Whitening Contrastive Learning (WCL). Our objective is to extract an embedding $\mathbf{z} = f(x;\theta)$ from an input target image x using an encoder network $f(\cdot;\theta)$ with parameters θ. To obtain meaningful target representations under distribution shifts, we enforce two key properties: (1) preventing representation collapse, where all embeddings converge to a degenerate point, and (2) ensuring that positive image pairs (x_i, x_j) sharing similar semantic content cluster closely in the feature space. We formulate this constrained optimization problem as:

$$min_\theta \, \mathbb{E}[dist(\mathbf{z}_i, \mathbf{z}_j)], \tag{5}$$

$$s.t. \; cov(\mathbf{z}_i, \mathbf{z}_i) = cov(\mathbf{z}_j, \mathbf{z}_j) = \mathbf{I}, \tag{6}$$

where $\mathbf{I}$ denotes the identity matrix, and $(\mathbf{z}_i, \mathbf{z}_j)$ represents the embedding pair of a positive sample pair (x_i, x_j). The covariance constraint in Eq. 6 prevents dimensional collapse and decorrelates the individual dimensions of $\mathbf{z}$, encouraging each component to capture distinct semantic variation. We measure distance using the Mean Squared Error (MSE) over L_2-normalized vectors, which is equivalent to cosine distance:

$$\begin{aligned} dist(\mathbf{z}_i, \mathbf{z}_j) &= \left\| \frac{\mathbf{z}_i}{\|\mathbf{z}_i\|_2} - \frac{\mathbf{z}_j}{\|\mathbf{z}_j\|_2} \right\|_2^2 \\ &= 2 - 2\frac{\langle \mathbf{z}_i, \mathbf{z}_j \rangle}{\|\mathbf{z}_i\|_2 \cdot \|\mathbf{z}_j\|_2} \end{aligned} \tag{7}$$

To obtain positive samples with shared semantics, we follow a procedure similar to AdaContrast Ermolov et al. (2021) which utilize a standard image augmentation.Specifically, we randomly and independently select parameters $\mathbf{p}$ for each positive sample extracted from a single image x. We use the notation $pos(i, j)$ to indicate that x_i and x_j have been extracted from the same image.

In order to facilitate representation learning, we employ an unsupervised backbone encoder network $E(\cdot)$ within AdaContrast. We use a conventional ResNet He et al. (2016) as the encoder, with $\mathbf{h} = E(x)$ representing the output of the average-pooling layer. This approach is straightforward and easily reproducible, compared to other methods that use encoder architectures specifically tailored to a particular pretext task. We follow Ermolov et al. (2021) and employ a nonlinear projection head $g(\cdot)$ to project $\mathbf{h}$ into a lower-dimensional space: $\mathbf{v} = g(\mathbf{h})$. $g(\cdot)$ is implemented using a multi-layer perceptron (MLP) with a hidden layer and a BN layer, and the complete network $f(\cdot)$ is the composition of $g(\cdot)$ with $E(\cdot)$.

Let N represent the number of original images, and consider a batch of samples $B = \{x_1, ..., x_k\}$, where $K = N_d$. We obtain the corresponding batch of features, denoted by $V = \{\mathbf{v}_1, ...\mathbf{v}_k\}$, using the aforementioned approach. In our proposed W-MSE loss, we calculate the MSE across all $N_{d(d-1)/2}$ positive pairs, satisfying the constraint through reparameterization of the $\mathbf{v}$ variables with the whitened variables $\mathbf{z}$

$$L_{WCL}(V) = \frac{2}{N_{d(d-1)/2}} \sum dist(\mathbf{z}_i, \mathbf{z}_j), \tag{8}$$

where $(\mathbf{v}_i, \mathbf{v}_j) \in V$ and $\mathbf{z} = Whitening(\mathbf{v})$ representing the whitened variables. Specifically, we have:

$$Whitening(\mathbf{v}) = W_V(\mathbf{v} - \boldsymbol{\mu}_V). \tag{9}$$

Eq. 9 defines $\boldsymbol{\mu}_V$ as the mean of the elements in V: $\boldsymbol{\mu}_V = \frac{1}{K}\sum_k \mathbf{v}_k$. The matrix W_V is determined such that $W_V^\top W_V = \Sigma_V^{-1}$, where Σ_V denotes the covariance matrix of V.

$$\Sigma_V = \frac{1}{K-1} \sum_k (\mathbf{v}_k - \boldsymbol{\mu}_V)(\mathbf{v}_k - \boldsymbol{\mu}_V)^T. \tag{10}$$

The proposed loss function in Eq. 8 operates on the premise that it penalizes distant positive pairs, thereby encouraging $g(E(\cdot))$ to reduce their inter-positive distances. Meanwhile, since Z must conform to a spherical distribution, the remaining samples must be repositioned to satisfy constraint Eq. 6.

Table 1: Classification accuracy (%) on VisDA-C train → val. All methods use ResNet-101 backbone except the on-target rows, which use ResNet-18 as the student network. Bold is the highest. The proposed AdaContrast surpasses the previous state-of-the-art by 1.0% Avg. When applying an extra knowledge distillation stage following Wang et al. (2021a), we achieve the highest 87.8% with a small ResNet-18 backbone.

Method	source-free	plane	bcycl	bus	car	horse	knife	mcycl	person	plant	sktbrd	train	truck	Avg.
DANN	no	81.9	77.7	82.8	44.3	81.2	29.5	65.1	28.6	51.9	54.6	82.8	7.8	57.4
CDAN	no	85.2	66.9	83.0	50.8	84.2	74.9	88.1	74.5	83.4	76.0	81.9	38.0	73.9
CDAN+BSP	no	92.4	61.0	81.0	57.5	89.0	80.6	90.1	77.0	84.2	77.9	82.1	38.4	75.9
CAN	no	97.0	87.2	82.5	74.3	**97.8**	**96.2**	90.8	80.7	**96.6**	**96.3**	87.5	**59.9**	**87.2**
SWD	no	90.8	82.5	81.7	70.5	91.7	69.5	86.3	77.5	87.4	63.6	85.6	29.2	76.4
MCC	no	88.7	80.3	80.5	71.5	90.1	93.2	85.0	71.6	89.4	73.8	85.0	36.9	78.8
Source only	-	57.2	11.1	42.4	66.9	55.0	4.4	81.1	27.3	57.9	29.4	86.7	5.8	43.8
MA	yes	94.8	73.4	68.8	74.8	93.1	95.4	88.6	84.7	89.1	84.7	83.5	48.1	81.6
BAIT	yes	93.7	83.2	84.5	65.0	92.9	95.4	88.1	80.8	90.0	89.0	84.0	45.3	82.7
SHOT	yes	95.3	87.5	78.7	55.6	94.1	94.2	81.4	80.0	91.8	90.7	86.5	59.8	83.0
+ On-target	yes	96.0	**89.5**	84.3	67.2	95.9	94.2	91.0	81.5	93.8	89.9	89.1	58.2	85.9
AdaContrast (baselines)	yes	97.0	84.7	84.0	77.3	96.7	93.8	91.9	84.8	94.3	93.1	**94.1**	49.7	86.8
TSIW	yes	**97.6**	87.8	**87.3**	**82.5**	**96.9**	92.8	**94.1**	**87.2**	**95.3**	**91.2**	92.6	48.5	**87.8**

Table 2: Classification accuracy (%) on 7 domain shifts of DomainNet-126. All methods use ResNet-50 backbone. Bold is the highest. The proposed TSIW achieves the highest average performance.

Method	Source-free	R→C	R→P	P→C	C→S	S→P	R→S	P→R	Avg.
MCC	no	44.8	65.7	41.9	34.9	47.3	35.3	72.4	48.9
Source only	-	55.5	62.7	53.0	46.9	50.1	46.3	75.0	55.6
TENT	yes	58.5	65.7	57.9	48.5	52.4	54.0	67.0	57.7
SHOT	yes	67.7	68.4	66.9	60.1	66.1	59.9	80.8	67.1
AdaContrast (baseline)	yes	70.2	69.8	68.6	58.0	65.9	61.5	80.5	67.8
TSIW	yes	**72.3**	**70.4**	**69.7**	**61.5**	**66.2**	**63.7**	**81.2**	**69.2**

Table 3: Ablation study on DomainNet-126. AdaContrast denotes the original baseline. AdaContrast (+WBN) denotes whitening Batch Normalization applied on the baseline AdaContrast. AdaContrast (+WCL) denotes whitening contrastive learning applied on the baseline AdaContrast.

Method	R→C	R→P	P→C	C→S	S→P	R→S	P→R	Avg.
AdaContrast	70.2	69.8	68.6	58.0	65.9	61.5	80.5	67.8
AdaContrast (+WBN)	71.0	70.1	68.9	59.3	66.0	62.7	80.8	68.4
AdaContrast (+WCL)	71.4	70.2	69.3	60.9	66.1	62.8	80.7	68.7
TSIW	**72.3**	**70.4**	**69.7**	**61.5**	**66.2**	**63.7**	**81.2**	**69.2**

2.4 Overall Formulation

With the whitening features, we can define our final objective for effective test-time adaptation. First, a recent work of AdaContrast Chen et al. (2022), which proposes a novel way to leverage self-supervised contrastive learning to facilitate target feature learning, along with an online pseudo labeling scheme with a refinement that significantly denoises pseudo labels. AdaConstrast consists of the exclusion of same-class negative pairs loss $\mathcal{L}^{\text{ctr}}$, weak-strong consistency loss $\mathcal{L}^{\text{ce}}$ and diversity regularization loss $\mathcal{L}^{\text{div}}$. Combining all these loss functions together, we can get the objective of AdaContrast as

$$\mathcal{L}_{\text{Adacontrast}} = \mathcal{L}^{\text{ctr}} + \mathcal{L}^{\text{ce}} + \mathcal{L}^{\text{div}} \tag{11}$$

Then the WBN and WCL module is applied for AdaContrast, our final objective becomes

$$\mathcal{L}_{\text{TSIW}} = \mathcal{L}^{\text{Whitening}}_{\text{Adacontrast}} \tag{12}$$

3 Experiment

We conduct experiments of closed-set adaptation on major benchmarks. In the following, we first compare the proposed method TSIW with the previous state-of-the-art algorithms. Then, we discuss several desirable test-time properties of TSIW, followed by ablation and analysis of the important design modules that brought the gains.

Table 4: Classification accuracy (%) on 12 domain shifts of ImageNet-C and CIFAR-100C.

Method (ImageNet-C)	Gauss	Shot	Impul	Defoc	Glass	Motion	Zoom	Snow	Frost	Fog	Brit	Cont
Tent	53.9	53.9	55.3	55.9	51.9	59.8	52.6	58.7	61.4	71.3	78.2	68.0
Tent + Whitening	**54.5**	**54.4**	**56.8**	**56.1**	**52.6**	**60.7**	**54.0**	**59.6**	**63.4**	**72.5**	**79.4**	**69.5**
CoTTA	43.5	44.1	45.4	46.8	51.5	47.6	36.8	48.3	39.7	43.5	47.7	42.1
CoTTA + Whitening	**44.9**	**42.5**	**46.7**	**37.9**	**39.3**	**48.3**	**43.7**	**45.7**	**43.5**	**63.6**	**75.2**	**63.7**
MEMO	40.6	37.5	40.6	33.4	26.7	41.2	35.4	54.7	67.1	65.9	66.7	59.3
MEMO + Whitening	**41.5**	**38.6**	**41.3**	**34.7**	**28.3**	**36.5**	**29.4**	**34.5**	**55.6**	**68.3**	**56.4**	**60.4**
Method (CIFAR-100C)	Gauss	Shot	Impul	Defoc	Glass	Motion	Zoom	Snow	Frost	Fog	Brit	Cont
RoTTA	31.8	36.7	40.9	42.1	30.3	27.9	25.4	32.3	34.0	38.8	38.3	31.3
RoTTA + Whitening	**32.5**	**37.3**	**41.5**	**43.7**	**33.5**	**34.8**	**28.9**	**26.4**	**33.2**	**35.3**	**39.7**	**39.5**

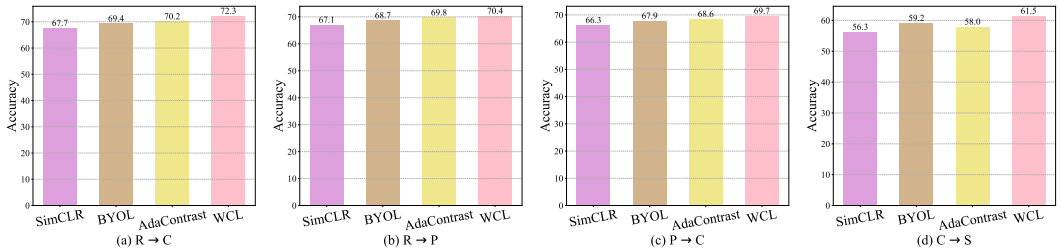


Figure 2: Classification accuracy on DomainNet-126 with different contrastive loss. (a) R → C, (b) R → P, (c) P → C, (d) C → S.

3.1 Experimental setup

Datasets and Metrics. We use VisDA-C Peng et al. (2017), DomainNet-126 Peng et al. (2019), ImageNet-C Hendrycks & Dietterich (2019) and CIFAR 100-C Hendrycks & Dietterich (2019) for evaluating our method and comparison.

VisDA-C Peng et al. (2017) is a large-scale and challenging benchmark, primarily designed for the 12-class synthesis-to-real object recognition task. The dataset consists of a source domain with 152,000 synthetic images generated by rendering 3D models and a target domain with 55,000 real object images sampled from Microsoft COCO.

DomainNet-126 is a subset of DomainNet Peng et al. (2019), which is currently the largest unsupervised domain adaptation dataset. It consists of six distinct domains and approximately 0.6 million images distributed among 345 categories. For our evaluation, we follow the approach of Saito et al. (2019) and select 126 classes from four domains, namely Real, Clipart, painting, and sketch.

ImageNet-C and CIFAR-100C Hendrycks & Dietterich (2019) were initially designed to evaluate the robustness of classification networks. Each dataset consists of 15 different corruption types, each with five severity levels. These corruptions were applied to images from the test set of the clean ImageNet-C or CIFAR-100 datasets.

It is important to note that the original DomainNet dataset has noisy labels. Therefore, we employ a subset of the dataset containing 126 classes from four domains, namely Real, Sketch, Clipart, and Painting, as recommended by the authors' follow-up work Saito et al. (2019). This subset is referred to as DomainNet-126. We evaluate the performance of the methods on seven domain shifts, as constructed from the domains, following the approach of Saito et al. (2019). We report the top-1 accuracy under each domain shift as well as the average across all seven shifts. For VisDA-C, we compare per-class top-1 accuracies and their average.

3.2 Compared Baseline

We perform a comparative analysis of our method against both classical unsupervised domain adaptation (UDA) baselines and source-free/test-time adaptation (TTA) baselines. For UDA methods, we compare our approach to DANN Ganin & Lempitsky (2015), CDAN Long et al. (2017), CDAN+BSP Chen et al. (2019), CAN Kang et al. (2019), SWD Lee et al. (2019), and MCC Jin et al. (2020). It

is important to note that all UDA methods have access to source data during adaptation. For TTA methods, we compare our approach to MA Li et al. (2020), BAIT Yang et al. (2020), TENT Wang et al. (2021b), SHOT Liang et al. (2020), and On-target Wang et al. (2021a), CoTTA Wang et al. (2022b), MEMO Zhang et al. (2022) and RoTTA Yuan et al. (2023). These methods represent a variety of approaches, including image generation, class prototypes, entropy minimization, pseudo labeling, and the combination of contrastive features and pseudo labeling.

3.3 Model Architecture

Our proposed method assumes a general architecture with a feature encoder followed by a classifier. For the purpose of comparison, we employ ResNet models He et al. (2016) as our backbones in various experiments. We use AdaContrast Chen et al. (2022) as our baseline and introduce a 256-dimensional bottleneck consisting of a fully-connected layer. To modify ResNet, we replace the first Batch Normalization (BN) layer and the BN layers in the first residual block with Weight Standardization (WBN) layers. We adopt the implementation and hyperparameter configuration from the AdaContrast method. As we utilize a lower-dimensional bottleneck, we remove the original projection heads from MoCo Chen et al. (2020b) without any observed performance decline.

3.4 Implementation

We adopt Pytorch Paszke et al. (2019) for all our implementations. For source training, we initialize the ResNet backbone with ImageNet-1K Deng et al. (2009) pre-trained weights from the Pytorch model zoo. We follow the approach of Liang et al. (2020) and randomly split the source data into a 9:1 ratio, with 90% used for training the source model and 10% used for validation. Source training is conducted for 10 and 60 epochs for VisDA-C and DomainNet-126, respectively. For target training, we use only 15 epochs for all datasets, unless otherwise indicated. For all experiments, we utilize the SGD optimizer with momentum 0.9 and weight decay 1e-4, and cosine annealing on the learning rate. The learning rate is decayed from its initial value to zero based on the training progress, as given by: $\eta = \eta_0 \cdot 0.5(\cos(a \cdot \pi/2) + 1)$. The newly added bottleneck and classifier layers have a learning rate 10 times that of the backbone. We set the initial learning rate for the backbone to 2e-4 and use a batch size of 128.

3.5 Results

VisDA-C. Table. 1 compares AdaContrast with state-of-the-art unsupervised domain adaptation and test-time adaptation methods on VisDA-C's "train" to "val" shift. For UDA, our method is on-par with a strong UDA baseline CAN Kang et al. (2019) and significantly outperforms the others by a large margin, even though we do not utilize source data at all during test-time adaptation. In the more challenging TTA setting, we achieve the highest per-class average accuracy by a notable margin (+1.0%) upon AdaContrast. This performance gap is especially significant given that VisDA-C represents one of the most challenging large-scale domain adaptation benchmarks, with substantial domain shift across all classes.

DomainNet-126. Table. 2 shows the comparison between TSIW and state-of-the-art UDA (first section) and TTA (second sections) methods. Without needing source data during the adaptation, our method outperforms the UDA method MCC Jin et al. (2020) by +20.3% on the averaged performance. This significant margin highlights the effectiveness of our approach in handling domain shifts without relying on source data, unlike traditional UDA techniques. When being compared to TTA methods TSIW outperforms TENT by +11.5% on the averaged performance. It achieves the best performance on four out of seven domain shifts as well as the highest averaged performance.

ImageNet-C and CIFAR 100-C. Table 4 demonstrates that integrating TSIW with existing TTA methods consistently enhances performance across a wide range of corruptions and datasets, particularly on ImageNet-C and CIFAR-100-C. For ImageNet-C, the Tent baseline shows moderate performance with an average accuracy of 61.2%. However, when whitening is added, the accuracy increases to 62.3%, with the largest gains observed in Motion, Fog, Brightness, and Contrast corruptions. This indicates that whitening enhances the model's robustness to distortions, improving its adaptability. Similarly, CoTTA, with its baseline performance of 49.1%, benefits from whitening, achieving an average accuracy of 50.3%. The improvement is particularly evident in Fog, Brightness,

and Motion corruptions. The MEMO baseline, which initially performs at 43.2%, also sees a notable increase when combined with whitening, reaching an average accuracy of 44.2%, with improvements in Fog, Brightness, and Motion.

When evaluated on CIFAR-100-C, RoTTA shows relatively low performance with an average accuracy of 35.0%. However, adding whitening boosts the accuracy to 36.1%, with significant gains in Motion, Brightness, and JPEG corruptions. These results confirm that whitening is effective in mitigating the adverse effects of domain shifts and corruptions in a variety of challenging scenarios. Overall, the integration of whitening consistently improves the robustness and adaptability of existing TTA methods, providing significant performance boosts across diverse corruption types and domain shifts in both ImageNet-C and CIFAR-100-C datasets. These improvements highlight the generalizability of the method and its potential for real-world applications where source data is unavailable during adaptation.

3.6 Analysis

Ablation Study. We performed an ablation study on the DomainNet-126 dataset to assess the effectiveness of each component in our TSIW. Table. 3 presents the classification results of the two modules of our method. We established a baseline result with WBN and WCL modules. We removed WBN and WCL individually to understand how the whitening module impacts the learning framework. Our observations indicate that the accuracy of the model with WBN (WCL) module is 0.6% (0.9%) higher than the baseline. Additionally, compared to the baseline, there is a clear performance gain of 1.4%. This implies that whitening from the same class closer in the embedding space is beneficial for improving the model's generalization power.

Domain Generalization Tasks. To evaluate the effectiveness of the proposed method on domain generalization, we conduct experiments over Digits-DG tasks. We use a recently proposed method SADML Wang et al. (2022a) as our baseline and apply the proposed method WBN and WCL to SADML. Specifically, on the hardest target domains SVHN and SYN, which involve cluttered digits and low image qualities, our method outperforms SADML 2.13% and 1.77%, respectively. The success of our method indicates that the whitening module applied to the domain generalization task can significantly promote its performance on leave-one-domain-out images.

Compared with Different Contrastive Loss. We show the DomainNet-126 results with different contrastive loss, and we compare WCL with results of other state-of-the-art approaches of contrastive learning such as SimCLR Chen et al. (2020a), BYOL Grill et al. (2020), AdaContrast, Despite some configuration details are different, in all cases, the encoder is a ResNet. As illustrated in Figure 2, WCL achieves highly competitive performance despite two key advantages: (1) we do not perform intensive hyperparameter tuning, and (2) our framework is significantly simpler than many existing approaches (e.g., it avoids the asymmetric architectures or momentum encoders used in BYOL). These results suggest that WCL effectively balances performance and simplicity, making it a practical choice for contrastive representation learning.

4 Conclusion

We introduced TSIW, a novel test-time adaptation approach for closed-set DA in image classification. In this paper, we tackle the challenge of adapting to a new domain without labeled data or access to the source domain. To address this problem, we propose Test-time Adaptation (TTA) as a solution that enables our model to adapt to the new data distribution during testing, using only unlabeled test data batches. Our approach introduces a novel method to facilitate target feature learning, which involves whitening Batch Normalization and whitening contrastive learning schemes. To enhance the optimization efficiency and generalization capabilities of our models, we employ ZCA whitening for feature decorrelation. Additionally, we enforce the constraint that batch samples should conform to a spherical distribution by whitening the latent space features in whitening contrastive learning. Our method achieves state-of-the-art performance on prominent benchmarks such as VisDA-C and DomainNet-126, and offers several advantages over existing works.

AI USE STATEMENT

Large language models, including ChatGPT, were only used during manuscript revision for language polishing. The authors reviewed and revised the AI-assisted content and take full responsibility for the research ideas, experimental design, data analysis, and scientific conclusions.

REFERENCES

Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016.

Collin Burns and Jacob Steinhardt. Limitations of post-hoc feature alignment for robustness. In *CVPR*, pp. 2525–2533, 2021.

Woong-Gi Chang, Tackgeun You, Seonguk Seo, Suha Kwak, and Bohyung Han. Domain-specific batch normalization for unsupervised domain adaptation. In *CVPR*, pp. 7354–7362, 2019.

Dian Chen, Dequan Wang, Trevor Darrell, and Sayna Ebrahimi. Contrastive test-time adaptation. In *CVPR*, pp. 295–305, 2022.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *ICML*, pp. 1597–1607, 2020a.

Xiao Chen, Zhongjing Du, Jiazhen Huang, Xu Jiang, Li Lu, Jingyan Jiang, and Zhi Wang. Neural collapse in test-time adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10567–10576, 2026.

Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297*, 2020b.

Xinyang Chen, Sinan Wang, Mingsheng Long, and Jianmin Wang. Transferability vs. discriminability: Batch spectral penalization for adversarial domain adaptation. In *ICML*, pp. 1081–1090. PMLR, 2019.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, pp. 248–255, 2009.

Aleksandr Ermolov, Aliaksandr Siarohin, Enver Sangineto, and Nicu Sebe. Whitening for self-supervised representation learning. In *ICML*, pp. 3015–3024, 2021.

Matteo Farina, Gianni Franchi, Giovanni Iacca, Massimiliano Mancini, and Elisa Ricci. Frustratingly easy test-time adaptation of vision-language models. *Advances in Neural Information Processing Systems*, 37:129062–129093, 2024.

Yaroslav Ganin and Victor Lempitsky. Unsupervised domain adaptation by backpropagation. In *ICML*, pp. 1180–1189. PMLR, 2015.

Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre H Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Daniel Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent: A new approach to self-supervised learning. *NeurIPS*, 2020.

Zongbo Han, Jialong Yang, Guangyu Wang, Junfan Li, Qianli Xu, Mike Zheng Shou, and Changqing Zhang. Dota: Distributional test-time adaptation of vision-language models. *Advances in Neural Information Processing Systems*, 38:142772–142798, 2026.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pp. 770–778, 2016.

Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. *ICLR*, 2019.

Nicholas J Higham. Analysis of the cholesky decomposition of a semi-definite matrix. 1990.

Mohammadmahdi Honarmand, Onur Cezmi Mutlu, Parnian Azizian, Saimourya Surabhi, and Dennis Wall. Selective test-time domain adaptation using fisher information for robust facial expression recognition in-the-wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5846–5856, 2025.

Lei Huang, Dawei Yang, Bo Lang, and Jia Deng. Decorrelated batch normalization. In *CVPR*, pp. 791–800, 2018.

Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *ICML*, pp. 448–456, 2015.

Yusuke Iwasawa and Yutaka Matsuo. Test-time classifier adjustment module for model-agnostic domain generalization. *NeurIPS*, 34:2427–2440, 2021.

Ying Jin, Ximei Wang, Mingsheng Long, and Jianmin Wang. Minimum class confusion for versatile domain adaptation. In *ECCV*, pp. 464–480, 2020.

Guoliang Kang, Lu Jiang, Yi Yang, and Alexander G Hauptmann. Contrastive adaptation network for unsupervised domain adaptation. In *CVPR*, pp. 4893–4902, 2019.

Adilbek Karmanov, Dayan Guan, Shijian Lu, Abdulmotaleb El Saddik, and Eric Xing. Efficient test-time adaptation of vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14162–14171, 2024.

Chen-Yu Lee, Tanmay Batra, Mohammad Haris Baig, and Daniel Ulbricht. Sliced wasserstein discrepancy for unsupervised domain adaptation. In *CVPR*, pp. 10285–10295, 2019.

Jindong Li, Yongguang Li, Yali Fu, Jiahong Liu, Yixin Liu, Menglin Yang, and Irwin King. Clip-powered domain generalization and domain adaptation: A comprehensive survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2026.

Rui Li, Qianfen Jiao, Wenming Cao, Hau-San Wong, and Si Wu. Model adaptation: Unsupervised domain adaptation without source data. In *CVPR*, pp. 9641–9650, 2020.

Yanghao Li, Naiyan Wang, Jianping Shi, Xiaodi Hou, and Jiaying Liu. Adaptive batch normalization for practical domain adaptation. *Pattern Recognition*, 80:109–117, 2018.

Jian Liang, Dapeng Hu, and Jiashi Feng. Do we really need to access the source data? source hypothesis transfer for unsupervised domain adaptation. In *ICML*, pp. 6028–6039, 2020.

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, 2025.

Hyesu Lim, Byeonggeun Kim, Jaegul Choo, and Sungha Choi. Ttn: A domain-shift aware batch normalization in test-time adaptation. *arXiv preprint arXiv:2302.05155*, 2023.

Sheng Liu, Xiao Li, Yuexiang Zhai, Chong You, Zhihui Zhu, Carlos Fernandez-Granda, and Qing Qu. Convolutional normalization: Improving deep convolutional network robustness and training. *NeurIPS*, 34:28919–28928, 2021.

Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *CVPR*, pp. 3431–3440, 2015.

Mingsheng Long, Zhangjie Cao, Jianmin Wang, and Michael I Jordan. Conditional adversarial domain adaptation. *ICML*, 2017.

Siwei Lyu and Eero P Simoncelli. Nonlinear image representation using divisive normalization. In *CVPR*, pp. 1–8, 2008.

A Tuan Nguyen, Thanh Nguyen-Tang, Ser-Nam Lim, and Philip HS Torr. Tipi: Test time adaptation with transformation invariance. In *CVPR*, pp. 24162–24171, 2023.

Liangbo Ning, Zuowei Zhang, Weiping Ding, Dian Shao, and Yining Zhu. Multilevel distribution alignment for multisource universal domain adaptation. *IEEE Transactions on Neural Networks and Learning Systems*, 36(9):17365–17379, 2025.

Shuaicheng Niu, Jiaxiang Wu, Yifan Zhang, Yaofo Chen, Shijian Zheng, Peilin Zhao, and Mingkui Tan. Efficient test-time model adaptation without forgetting. In *ICML*, pp. 16888–16905, 2022.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *NeurIPS*, 32:8026–8037, 2019.

Xingchao Peng, Ben Usman, Neela Kaushik, Judy Hoffman, Dequan Wang, and Kate Saenko. Visda: The visual domain adaptation challenge. *CVPR*, 2017.

Xingchao Peng, Qinxun Bai, Xide Xia, Zijun Huang, Kate Saenko, and Bo Wang. Moment matching for multi-source domain adaptation. In *ICCV*, pp. 1406–1415, 2019.

Subhankar Roy, Aliaksandr Siarohin, Enver Sangineto, Samuel Rota Bulo, Nicu Sebe, and Elisa Ricci. Unsupervised domain adaptation using feature-whitening and consensus loss. In *CVPR*, pp. 9471–9480, 2019.

Kuniaki Saito, Donghyun Kim, Stan Sclaroff, Trevor Darrell, and Kate Saenko. Semi-supervised domain adaptation via minimax entropy. In *ICCV*, pp. 8050–8058, 2019.

Wenqi Shao, Tianjian Meng, Jingyu Li, Ruimao Zhang, Yudian Li, Xiaogang Wang, and Ping Luo. Ssn: Learning sparse switchable normalization via sparsestmax. In *CVPR*, pp. 443–451, 2019.

Dequan Wang, Shaoteng Liu, Sayna Ebrahimi, Evan Shelhamer, and Trevor Darrell. On-target adaptation. *arXiv preprint arXiv:2109.01087*, 2021a.

Dequan Wang, Evan Shelhamer, Shaoteng Liu, Bruno Olshausen, and Trevor Darrell. Tent: Fully test-time adaptation by entropy minimization. In *ICLR*, 2021b.

Mengzhu Wang, Jianlong Yuan, Qi Qian, Zhibin Wang, and Hao Li. Semantic data augmentation based distance metric learning for domain generalization. In *ACMMM*, pp. 3214–3223, 2022a.

Qin Wang, Olga Fink, Luc Van Gool, and Dengxin Dai. Continual test-time domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7201–7211, 2022b.

Ximei Wang, Ying Jin, Mingsheng Long, Jianmin Wang, and Michael I Jordan. Transferable normalization: Towards improving transferability of deep neural networks. *NeurIPS*, 32, 2019.

Yuxin Wu and Kaiming He. Group normalization. In *ECCV*, pp. 3–19, 2018.

Shiqi Yang, Yaxing Wang, Joost van de Weijer, Luis Herranz, and Shangling Jui. Unsupervised domain adaptation without source data by casting a bait. *arXiv preprint arXiv:2010.12427*, 2020.

Longhui Yuan, Binhui Xie, and Shuang Li. Robust test-time adaptation in dynamic scenarios. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15922–15932, 2023.

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

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.

Youjia Zhang, Youngeun Kim, Young-Geun Choi, Hongyeob Kim, Huiling Liu, and Sungeun Hong. Backpropagation-free test-time adaptation via probabilistic gaussian alignment. *Advances in Neural Information Processing Systems*, 38:76280–76310, 2026.

Zhong-Qiu Zhao, Peng Zheng, Shou-tao Xu, and Xindong Wu. Object detection with deep learning: A review. *IEEE transactions on neural networks and learning systems*, 30(11):3212–3232, 2019.

Zhengxia Zou, Keyan Chen, Zhenwei Shi, Yuhong Guo, and Jieping Ye. Object detection in 20 years: A survey. *Proceedings of the IEEE*, 111(3):257–276, 2023.

APPENDIX

A RELATED WORK

A.1 TEST-TIME ADAPTATION

In this setting, the unlabeled target data is not available during the training phase of the source model. Instead, the model only has access to the target data when deployed in the target environment, as the stream of test data batches. As a consequence, the unlabeled target data is not available in large quantities, but only in small batches. So far, there are two main lines of work dealing with this online adaptation problem. The first approach is based on the batch normalization layer in the network, with the idea of using the test-time batch statistics for these layers instead of the source statistics. This can partially help with the representation misalignment between the source and target distributions but is only applicable when the network has BatchNorm layers and the batch size is large. The other line of work is to use an unsupervised surrogate loss to "refine" the model on the target data. For example, TENT minimizes the prediction entropy of the model on the target data. However, as pointed out by its authors, TENT needs a large batch size to avoid collapsing to a trivial prediction, due to the characteristics of this surrogate objective. In TENT, this is achieved with a simple entropy minimization loss, which informs the optimization of scale and bias parameters of batch normalization layers. As for batch normalization layers' statistics, they are re-estimated on the test data, similarly to what is done in adaptive batchnorm (AdaBN) methods Hendrycks & Dietterich (2019); Burns & Steinhardt (2021), which have shown strong performance on the perturbations of ImageNet-C Hendrycks & Dietterich (2019). In a similar spirit, Liang *et al*. Liang et al. (2020) updates the parameters of the feature extractor of a given model by maximizing a mutual information objective. Meanwhile, TTT and TTT+++ train a self-supervised task alongside the main task on the source domain and continue to train the self-supervised task during the online adaptation phase. Therefore, TTT and TTT+++ are not fully off-the-shelf test-time adaptation methods, since they require modifications to the training phase of the source domain. Recently, TIPI Nguyen et al. (2023) propose to use an invariance regularize as the surrogate loss during test-time adaptation, motivated by our theoretical results regarding the model's performance under input transformations. There also have been several works Niu et al. (2022) that study datapoints selection strategies for optimizing the unsupervised surrogate loss which only optimize the loss for reliable and informative datapoints.

A.2 NORMALIZATION TECHNIQUES

It is widely in CNNs to make them learn faster, more stable, and increase their generalization ability. Representative methods include BN Ioffe & Szegedy (2015), Layer Normalization Ba et al. (2016), Adaptive BN Li et al. (2018), Group Normalization Wu & He (2018), Switchable Normalization Shao et al. (2019) and Representation Normalization Lyu & Simoncelli (2008). For better adaptation, researchers have devised novel designs to mitigate the shortcomings in BN. AutoDIAL uses the statistics of two domains channel by channel in order to align the source and target feature distributions. Domain Specific BN Chang et al. (2019) normalizes the source and target representations completely individually, including affine parameters. Transferable Normalization Wang et al. (2019) utilizes the statistics of two domains to calculate corresponding channel attention, which is all detached from the computation graph. ConvNorm Liu et al. (2021) proposes an adaptation layer $\mathcal{A}$ to whiten and color source domain data, then $\mathcal{A}$ is fine-tuned on the target domain. DWT Roy et al. (2019) uses two co-variance matrices to whiten feature maps from source and target domains, respectively. TTN Lim et al. (2023) employs test batch statistics mitigates the performance degradation caused by the domain shift and present a strategy for mixing conventional batch normalization and transductive batch normalization based on the interpolating weight derived from the optimization procedure utilizing the sensitivity to domain shift and show that our method significantly outperforms other normalization techniques in various realistic evaluation settings. DomainAdapter Zhang et al. (2023) dynamically fuses statistics between source and test batch statistics for an accurate statistics estimation and generalizatized entropy minimization loss effectively enhances the adaptation ability. Different from these existing normalization counterparts, we focus on decoupling batch normalization in CNNs for TTA. In this paper, we leverage Decoupling Batch Normalization (DBN), which extends Batch Normalization to include whitening over mini-batch data.