MixTTA: Low-Rank Cross-Channel Mixing for Reliable Test-Time Adaptation

Mansoo Jung[1], Youngwook Kim[2⋆], and Jungwoo Lee[1,3⋆]

[1] Seoul National University, Seoul, Republic of Korea
{tlwh1179, junglee}@snu.ac.kr
[2] Kookmin University, Seoul, Republic of Korea
youngwook@kookmin.ac.kr
[3] HodooAI Labs, Seoul, Republic of Korea

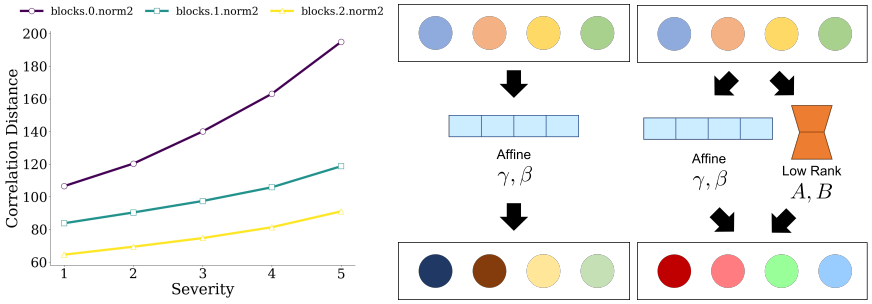
Abstract. Test-Time Adaptation (TTA) methods commonly update the affine parameters of normalization layers to adapt deployed models under distribution shifts. However, per-channel affine parameters perform axis-aligned scaling and shifting, making them geometrically incapable of correcting cross-channel structural changes induced by distribution shift. To address this limitation, we propose MixTTA, a lightweight plug-in module that equips normalization layers with a low-rank cross-channel transformation, enabling inter-channel mixing at each layer. To ensure that the low-rank branch captures only cross-channel interactions, we also propose Decoupling Projection that enforces strict separation from the diagonal affine path, along with Spectral Projection that prevents rank-1 collapse under non-stationary test streams. MixTTA can be seamlessly integrated into any existing normalization-based TTA method. Experiments in both standard and wild TTA settings show consistent improvements over strong baselines while mitigating adaptation failure under challenging conditions. The source code is publicly available at https://github.com/delta6189/MixTTA.

Keywords: Test-time Adaptation · Correlation-aware Adaptation · Low-rank Adaptation

1 Introduction

Deep neural networks have achieved remarkable progress over the past decade, demonstrating exceptional performance across a wide range of tasks in computer vision [4, 8, 17, 18, 20, 22, 36, 37, 39]. However, most models are trained under the unrealistic assumption that the training and test distributions are identical. In real-world scenarios, distribution shifts, arising from environmental changes, sensor noise, or data corruption, can lead to severe performance degradation [9]. Consequently, addressing domain shifts has become a central challenge in modern machine learning, motivating extensive research on adaptation frameworks [2, 6, 25, 34, 35, 38, 40, 45]. Among these, test-time adaptation (TTA) [5, 14, 16, 21, 31,

⋆ Corresponding authors.



(a) Correlation Distance **(b)** Tent **(c)** MixTTA (Ours)

Fig. 1: (a) Correlation distance comparison across three layers of a pre-trained ViT on ImageNet-C *Gaussian noise* under varying severity levels. (b) and (c) provide overviews of Tent [40] and our proposed MixTTA, respectively, where each colored circle represents a feature dimension. Tent applies only channel-wise affine modulation without mixing across dimensions (colors remain unchanged while intensities vary), whereas MixTTA introduces an additional low-rank cross-channel transform that mixes feature dimensions, leading to changes in both color and intensity.

40, 42, 48] has emerged as a promising direction since it allows a deployed model to adjust itself online without access to source data or labels.

One of the foundational methods of TTA is Tent [40], which sharpens predictions by minimizing prediction entropy while updating only the affine parameters of normalization layers. Its popularity stems from the fact that constraining updates to per-channel scale and bias maintains a compact update space, stabilizing adaptation and reducing the risk of overfitting. Subsequent works have typically retained the same set of learnable parameters, incorporating additional strategies such as regularization [29], sample selection [23,30], or region-level confidence estimation [11]. Despite these advances, the structural form of the affine transformation itself has remained largely unexamined.

In this paper, we revisit this design choice by examining how distribution shift affects intermediate feature representations. Specifically, we measure the correlation distance $D = \frac{1}{N}\sum_{n=1}^{N} \|\Sigma_n^s - \Sigma_n^t\|_F$, where Σ_n^s and Σ_n^t are the per-sample covariance matrices of the source and target features after each normalization layer, N is the number of test samples, and $\|\cdot\|_F$ denotes the Frobenius norm. As illustrated in Fig. 1a, the correlation distance increases progressively with corruption severity across all layers, revealing that distribution shift does not merely alter per-channel variances but fundamentally changes the inter-channel correlation structure. Since the learnable affine parameters of normalization layers perform axis-aligned per-channel scaling and shifting, they are *geometrically incapable of correcting such cross-channel structural changes*. Moreover, earlier layers exhibit substantially larger correlation distances than deeper layers, indicating that earlier layers are most affected by correlation disruption. While TCA [46] proposes aligning feature correlations at test time, it operates only on

the final representation, leaving the pronounced early-layer misalignment unaddressed.

Motivated by these observations, we propose **MixTTA**, a lightweight module that extends normalization layer adaptation to capture cross-channel feature interactions. As illustrated in Figs. 1b and 1c, MixTTA equips the existing per-channel affine transform with a residual cross-channel term parameterized by low-rank matrices, enabling cross-channel mixing at each normalization layer throughout the network. The low-rank design offers a parameter-efficient alternative to full cross-channel mixing, avoiding the instability of unconstrained adaptation while retaining sufficient capacity to capture the dominant correlation shifts. To maintain a clear functional separation between the diagonal affine path and the low-rank branch, we propose **Decoupling Projection** that constrains the low-rank component to purely off-diagonal corrections. Additionally, we further propose **Spectral Projection** that suppresses rank-1 collapse induced by the dominant principal component, stabilizing adaptation under biased or non-stationary test streams.

The proposed module operates as a plug-in that can be integrated into any existing normalization-based TTA method that updates axis-aligned per-channel parameters. Extensive experiments across standard and wild TTA scenarios demonstrate that MixTTA achieves state-of-the-art adaptation performance. Notably, under wild scenarios involving class imbalance, single-sample adaptation, and mixed domains, MixTTA substantially mitigates the prediction collapse that commonly afflicts existing methods.

The contributions of our work are summarized as follows:

- We show that distribution shift induces substantial changes in inter-channel feature correlations, and that this disruption is most pronounced in earlier layers. This reveals a key limitation of the per-channel affine paradigm in TTA, as per-channel modulation cannot correct cross-channel structural changes.
- We introduce MixTTA, a residual low-rank cross-channel transform that extends normalization layer adaptation to model inter-channel dependencies. We additionally propose Decoupling Projection to enforce strict diagonal/off-diagonal separation and Spectral Projection to prevent rank-1 collapse under non-stationary test streams.
- When integrated into diverse TTA baselines, MixTTA yields consistent accuracy gains across standard and wild scenarios with improved robustness and minimal overhead.

2 Related Work

2.1 Test-Time Adaptation via Normalization Layers

Test-time adaptation (TTA) has recently emerged as a promising paradigm for improving model robustness under distribution shifts without requiring access to source data or target labels. Tent [40] is a foundational method that minimizes

prediction entropy while updating only the affine parameters of normalization layers. Subsequent works have extended Tent along complementary directions. MEMO [48] augments each test instance and adapts by minimizing the marginal entropy across augmentations. EATA [29] introduces a Fisher regularizer to preserve critical parameters and selects reliable test samples to mitigate catastrophic forgetting. SAR [30] achieves robust online adaptation through batch-agnostic normalization and sharpness-aware optimization. DeYO [23] proposes the Pseudo-Label Probability Difference (PLPD) metric, weighting samples by shape consistency when entropy-based confidence is unreliable. COME [49] addresses the overconfidence failure mode of entropy minimization by modeling a Dirichlet prior over predictions. ReCAP [11] further improves robustness in wild scenarios by leveraging region-level confidence as an adaptation proxy. Despite their advances, these approaches all operate within Tent's per-channel affine update regime. Our method, MixTTA, extends this regime by introducing cross-channel mixing through a lightweight low-rank module while maintaining parameter efficiency.

2.2 Correlation Alignment under Distribution Shift

Correlation alignment aims to match feature distributions between source and target domains by aligning their covariance structures. CORAL [38] pioneered this direction by minimizing the distance between second-order statistics, and subsequent works extended it to higher-order moments (HoMM [3], CMD [47]) and entropy-aware formulations [28]. Recently, TCA [46] adapted correlation alignment to the TTA scenario by transforming test features to match pseudo-source correlations estimated from high-confidence samples. However, TCA operates only on the final feature representation, leaving early-layer correlation misalignment unaddressed. In contrast, our approach places cross-channel correction at each normalization layer, enabling hierarchical alignment throughout the network.

2.3 Low-Rank Adaptation

Low-rank adaptation (LoRA) [10] factorizes weight updates into low-rank matrices, enabling efficient fine-tuning with minimal additional parameters. This paradigm has inspired a range of parameter-efficient adaptation methods such as AdapterFusion [32], Prompt Tuning [24], and Visual Prompt Tuning [15]. Recent efforts have incorporated low-rank modules into test-time adaptation scenarios as well. Imam *et al.* [12] inserted low-rank adapters into transformer attention and optimized them via confidence maximization on unlabeled test data, while Kojima *et al.* [19] performed low-rank test-time training of visual encoders in vision–language models with an auxiliary self-supervised objective. Additionally, ViDA [26] targets Continual TTA and injects high-rank and low-rank adapters into linear or convolutional layers, with a Homeostatic Knowledge Allotment strategy to mitigate catastrophic forgetting. These methods apply

low-rank factorization to network weights, whereas MixTTA applies it to inter-channel dependencies within normalization layers, targeting correlation structure rather than weight adaptation.

3 Methodology

3.1 Preliminary

Given a source model with parameters θ and target data x, Tent [40] performs test-time adaptation by minimizing the prediction entropy:

$$\mathcal{L}_{\text{Ent}}(x) \;=\; \text{Ent}_\theta(x), \tag{1}$$

where $\text{Ent}_\theta(x)$ denotes the prediction entropy of the model. To optimize this objective, Tent updates only the affine parameters of normalization layers (*e.g.*, BatchNorm [13], GroupNorm [44], or LayerNorm [1]), while keeping all other weights frozen to avoid overfitting and catastrophic collapse. Formally, Tent modulates the normalized feature $x \in \mathbb{R}^{C\times T}$ as

$$y \;=\; \gamma \odot x \;+\; \beta, \tag{2}$$

where $\odot$ denotes the Hadamard product, $y \in \mathbb{R}^{C\times T}$ denotes the modulated activation, $\gamma, \beta \in \mathbb{R}^C$ are learnable per-channel scale and bias parameters, C is the number of channels, and T is the number of tokens (*e.g.*, $T = HW$ for CNNs or sequence length for Vision Transformers).

3.2 Low-Rank Cross-Channel Mixing

Per-channel modulation in Tent and its follow-ups [23, 29, 30] is parameter-efficient and effective, but it cannot explicitly capture cross-channel dependencies since it only rescales each channel independently. This limitation is directly reflected in Eq. (2), which can be rewritten in the following matrix form:

$$y \;=\; \Gamma x + \beta, \tag{3}$$

where $\Gamma \in \mathbb{R}^{C\times C}$ is the diagonal matrix with diagonal entries γ. This highlights that Tent restricts the channel transform to be diagonal and cannot model off-diagonal cross-channel mixing. When distribution shifts alter inter-channel correlations, such limited representational capacity can lead to under-adaptation.

A naïve remedy for this limitation is to replace Γ with a full channel-mixing transform $W \in \mathbb{R}^{C\times C}$ applied after normalization. Although this modification allows the model to capture inter-channel correlations, it introduces C^2 parameters per layer, which substantially increases adaptation cost and makes the model vulnerable to overfitting or collapse during adaptation. Moreover, learning a dense W may dilute the influence of the pretrained affine parameters (γ, β) that normalization layers already provide, removing the strong inductive bias of

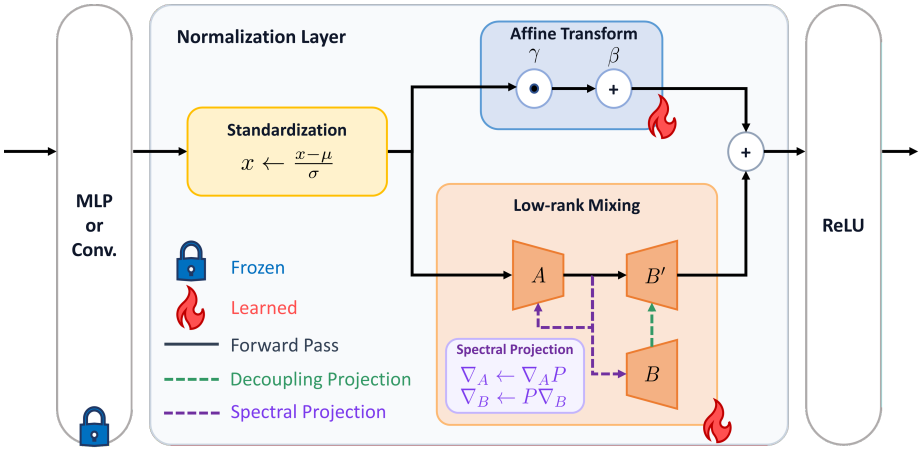


Fig. 2: Overall structure of MixTTA module, which operates within normalization layers of the backbone network. Standardized features are modulated through both the affine and residual low-rank branches. Decoupling Projection ensures strict separation between diagonal and off-diagonal updates, while Spectral Projection filters collapse-prone directions during adaptation.

a diagonal initialization. To address this, we introduce a *low-rank channel mixing* that extends Tent's diagonal scaling by a compact rank-r perturbation:

$$y \;=\; \gamma \odot x \;+\; (AB)^{\top} x \;+\; \beta, \tag{4}$$

where $A \in \mathbb{R}^{C\times r}$ and $B \in \mathbb{R}^{r\times C}$ are learnable low-rank matrices, and $r \in \mathbb{N}$ denotes the subspace dimension with $r \ll C$. This parameterization decomposes W as $W = \Gamma + \Delta$, where $\Delta = (AB)^{\top}$. The diagonal term Γ retains Tent's affine updates, while the low-rank perturbation Δ captures cross-channel dependencies by projecting features onto a compact subspace and reconstructing them to the full channel space. This formulation yields a unified modulation framework that jointly captures per-channel and cross-channel transformations with only a modest overhead in parameters. In practice, we initialize B as a zero matrix so that $W = \Gamma$ at the start of adaptation, recovering Tent and ensuring stable warm-start behavior. Fig. 2 illustrates the overall structure of the proposed MixTTA module.

3.3 Decoupling Projection

Although Δ is intended to capture cross-channel mixing, its non-zero diagonal elements can still induce per-channel modulation and thus duplicate the *functionality* of Γ. To enforce a clear functional decoupling of Γ and Δ, we impose the constraint

$$\mathrm{diag}(\Delta) = \mathbf{0}, \tag{5}$$

where $\mathrm{diag}(\cdot)$ denotes the diagonal vector of a matrix. Noting that $\Delta_{ii} = (AB)_{ii} = a_i^\top b_i$, where $a_i \in \mathbb{R}^r$ is the i-th row of A and $b_i \in \mathbb{R}^r$ is the i-th column of B, Eq. (5) is satisfied if $a_i^\top b_i = 0$ for all i. Accordingly, before the forward pass, we project each b_i onto the orthogonal complement of a_i:

$$b_i' \leftarrow b_i - \mathrm{sg}\left(\frac{a_i^\top b_i}{\|a_i\|_2^2 + \epsilon}\, a_i\right), \qquad \forall i \in \{1, \ldots, C\}, \tag{6}$$

where $\mathrm{sg}(\cdot)$ denotes the stop-gradient operator, $\|\cdot\|_2$ denotes the L2-norm, and $\epsilon > 0$ is a small constant for numerical stability. This projection enforces the diagonal components of Δ to be zero while preserving the off-diagonal mixing capacity of the low-rank branch.

3.4 Spectral Projection

While MixTTA enables the model to exploit cross-channel interactions during test-time adaptation, its entropy-driven updates remain susceptible to collapsing into degenerate modes. In practice, when the target data stream is biased or non-stationary, the adapted model tends to concentrate its optimization along dominant *rank-1* directions.

We now analyze why unconstrained MixTTA can collapse into a rank-1 mode. Let the upstream gradient $g := \partial \mathcal{L}_{\mathrm{Ent}} / \partial y \in \mathbb{R}^{C \times T}$. By standard backpropagation through $B^\top A^\top x$, the gradients of the low-rank factors take the coupled form

$$\nabla_A \mathcal{L}_{\mathrm{Ent}} = x\big(Bg\big)^\top, \qquad \nabla_B \mathcal{L}_{\mathrm{Ent}} = (A^\top x)\, g^\top. \tag{7}$$

Therefore, given step size δ, SGD updates of (A, B) follow an alternating dynamics:

$$A \leftarrow A - \delta\, x(Bg)^\top, \qquad B \leftarrow B - \delta\, (A^\top x)\, g^\top. \tag{8}$$

Crucially, the update of A depends on B (through Bg), and the update of B depends on A (through $A^\top x$), which creates a positive feedback loop that repeatedly reinforces whichever direction dominates the subspace features. When the subspace statistics become highly anisotropic (*i.e.*, $\lambda_1 \gg \lambda_2$), this alternation tends to align the columns of A and B to the same dominant axis, suppressing cross-channel diversity and effectively degenerating the low-rank update to rank-1. Empirically, we observe that the top-1 principal component tends to dominate the feature energy and diverge when left unregularized. To mitigate this excessive rank-1 concentration, we introduce the **Spectral Projection**.

Given the subspace feature $z := A^\top x$, we compute the dominant eigenvector $u_1 \in \mathbb{R}^r$ of the feature covariance as follows:

$$u_1 := \mathrm{eigvec}_{\max}(\mathrm{Cov}(z)), \tag{9}$$

where $\mathrm{Cov}(z)$ denotes the covariance matrix of z and $\mathrm{eigvec}_{\max}(\cdot)$ denotes the eigenvector corresponding to the largest eigenvalue, which can be efficiently obtained via the power iteration method [27]. We then construct a projection operator P as follows:

$$P := I - \eta \frac{u_1 u_1^\top}{\|u_1\|_2^2 + \epsilon} \in \mathbb{R}^{r \times r}, \tag{10}$$

where $I \in \mathbb{R}^{r \times r}$ denotes the identity matrix, $\eta \in [0, 1]$ controls the projection strength. During adaptation, the low-rank gradients are projected as

$$\nabla_A \leftarrow \nabla_A P, \qquad \nabla_B \leftarrow P \nabla_B, \tag{11}$$

thereby suppressing the dominant rank-1 direction and stabilizing adaptation. For batched inputs, we compute a per-sample projection matrix and use their average $\bar{P}$ as the batch-level projection operator. Although $\bar{P}$ is not an exact projector (*i.e.*, $\bar{P}^2 \neq \bar{P}$), we find that it sufficiently suppresses the dominant component and stabilizes adaptation.

4 Experiments

4.1 Benchmarks, Baselines, and Test Scenarios

We evaluated our method on ImageNet-C [9] and ImageNet-Sketch [41], two standard benchmarks for TTA. ImageNet-C is derived from the original ImageNet dataset [33] and designed to assess model robustness under a wide range of common corruptions. It includes 15 corruption types each applied at five severity levels. ImageNet-Sketch, on the other hand, contains sketch-based depictions of the 1000 ImageNet categories, providing substantial appearance-level deviations from natural images. This benchmark focuses on cross-domain generalization rather than simple corruption. For adaptation strategies, we evaluated MixTTA by integrating it into representative TTA baselines, including Tent [40], EATA [29], SAR [30], DeYO [23], and ReCAP [11]. We additionally included LinearTCA and LinearTCA$^+$ [46] as a baseline that addresses cross-channel correlation at test time. Following the convention in [46], where LinearTCA$^+$ denotes LinearTCA applied to the best-performing baseline, we instantiated LinearTCA$^+$ using ReCAP (*i.e.*, ReCAP+LinearTCA) to enable a fair and consistent comparison with a state-of-the-art base method. Finally, we evaluated under the mild and three wild test scenarios proposed by Niu et al. [30]. Wild scenarios include (i) *online imbalanced label shift*, which models fluctuations in the ground-truth label distribution, (ii) *single-sample* setting (*i.e.*, batch size 1), which evaluates a model's ability to adapt without relying on batch statistics, and (iii) *mixed shift*, which combines multiple types of distribution shifts simultaneously.

4.2 Implementation Details

Unless otherwise noted, we followed the experimental protocol of [23]. All experiments were repeated with three random seeds and we report their mean. We employed ViT [4] with Layer Normalization [1] as backbone models and pretrained weights were obtained from the `timm` library [43]. The learning rate was set to 0.001 and it was doubled under the batch size 1 scenario to compensate for the reduced reliability of batch-level statistics. We initialized the matrix A using the Xavier scheme [7] and set B to the zero matrix. MixTTA module was

Mild	Noise			Blur				Weather				Digital				Avg.
	Gauss.	Shot	Impulse	Defocus	Glass	Motion	Zoom	Snow	Frost	Fog	Brit.	Contr.	Elastic	Pixel	JPEG	
No Adapt.	16.8	12.0	16.5	29.2	23.6	34.0	27.3	15.8	26.6	47.5	55.4	44.3	31.0	45.1	49.1	31.6
• LinearTCA	17.9	13.0	17.4	29.8	24.2	35.7	28.8	17.0	27.8	49.8	56.0	45.8	33.0	46.0	49.8	32.8 ± 0.0
• LinearTCA$^+$	54.3	54.6	55.4	58.3	58.4	63.2	60.1	66.8	65.7	73.4	78.1	68.3	67.4	73.1	70.2	64.5 ± 0.3
• Tent	44.6	**42.8**	**45.2**	**52.1**	47.8	55.3	50.3	18.1	21.2	66.4	75.0	64.8	52.7	66.9	64.4	51.2 ± 0.0
+MixTTA	**47.9**	38.7	41.3	51.3	**48.3**	**56.2**	**51.0**	**57.2**	**41.6**	**68.0**	**75.3**	**64.9**	**53.6**	**68.2**	**65.1**	**55.2** ± 1.4
• EATA	51.7	51.6	52.4	55.5	55.8	60.2	57.7	63.1	61.0	71.2	75.4	67.1	64.2	70.5	67.9	61.7 ± 0.3
+MixTTA	**53.9**	**53.5**	**55.0**	**56.3**	**57.4**	**62.1**	**59.7**	**66.5**	**64.6**	**72.4**	**77.5**	**67.6**	**65.1**	**72.7**	**68.8**	**63.5** ± 0.1
• SAR	46.3	45.0	46.9	**52.9**	**50.0**	**55.9**	**51.5**	57.0	53.5	66.5	74.8	**64.4**	**55.2**	66.8	**64.5**	56.7 ± 0.2
+MixTTA	**47.8**	**46.9**	**49.1**	51.7	49.0	55.7	50.4	**58.9**	**57.4**	**67.0**	74.8	64.1	53.9	**67.4**	64.2	**57.2** ± 0.3
• DeYO	54.7	55.1	55.7	**58.4**	58.9	63.4	45.8	67.2	65.8	73.3	78.3	68.0	68.0	73.4	70.4	63.8 ± 0.6
+MixTTA	**56.0**	**56.6**	**57.0**	58.2	**59.3**	**64.6**	**62.2**	**68.6**	**67.1**	**74.0**	**78.5**	**68.6**	**68.4**	**74.2**	**71.0**	**65.6** ± 0.1
• ReCAP	53.9	54.2	55.0	**57.9**	58.1	62.6	58.9	66.4	65.1	72.9	78.0	67.9	67.0	72.7	69.9	64.0 ± 0.1
+MixTTA	**55.4**	**55.7**	**56.4**	57.6	**58.6**	**63.6**	**60.9**	**67.9**	**66.4**	**73.4**	**78.2**	**68.4**	**67.1**	**73.5**	**70.4**	**64.9** ± 0.1

Table 1: Comparisons with baselines on ImageNet-C at severity level 5 under the mild scenario regarding accuracy (%). **Bold** values denote the top-performing results.

Label Shifts	Noise			Blur				Weather				Digital				Avg.
	Gauss.	Shot	Impulse	Defocus	Glass	Motion	Zoom	Snow	Frost	Fog	Brit.	Contr.	Elastic	Pixel	JPEG	
No Adapt.	16.7	11.9	16.3	29.3	23.6	33.9	27.3	15.9	26.5	47.2	54.9	44.1	30.9	45.0	48.9	31.5
• LinearTCA	18.0	13.0	17.4	30.1	24.4	36.0	29.0	17.2	28.0	50.0	55.6	45.9	33.1	46.2	49.8	32.9 ± 0.0
• LinearTCA$^+$	55.1	55.1	56.0	58.3	58.9	63.8	61.9	67.9	66.3	73.3	78.0	67.1	69.0	73.5	70.4	65.0 ± 0.1
• Tent	**46.4**	**46.8**	**46.1**	**54.6**	52.2	58.2	53.0	10.2	10.5	69.6	76.2	66.1	58.1	69.4	66.7	52.3 ± 0.3
+MixTTA	37.8	34.0	29.4	54.2	**53.6**	**59.6**	**55.1**	**45.5**	**21.0**	**70.9**	**76.5**	**66.7**	**60.1**	**71.0**	**68.1**	**53.6** ± 2.5
• EATA	39.3	37.6	38.9	44.1	45.2	**44.4**	46.8	54.7	54.1	**60.0**	72.5	**25.9**	58.0	66.3	64.0	50.1 ± 0.7
+MixTTA	**47.9**	**46.2**	**48.9**	**50.3**	**52.6**	31.6	**54.2**	**62.9**	**60.6**	58.7	**76.0**	19.1	**62.6**	**71.0**	**66.1**	**53.9** ± 3.3
• SAR	50.0	49.1	50.6	**55.4**	54.1	59.1	54.6	56.9	48.6	69.7	76.3	66.2	**60.8**	69.7	66.9	59.2 ± 0.6
+MixTTA	**52.0**	**51.2**	**52.7**	54.6	**54.6**	**59.5**	**54.8**	**63.1**	**58.4**	**70.4**	76.3	**66.3**	60.1	**70.5**	**67.5**	**60.8** ± 0.2
• DeYO	54.5	55.1	55.8	**58.0**	58.9	63.8	61.4	67.9	66.2	73.1	78.0	66.7	69.1	73.6	70.5	64.8 ± 0.1
+MixTTA	**56.0**	**56.3**	**56.9**	57.9	**59.3**	**64.7**	**62.8**	**68.9**	**67.2**	**73.9**	**78.4**	**67.4**	**69.2**	**74.2**	**71.1**	**65.6** ± 0.1
• ReCAP	54.4	54.9	55.6	58.1	58.9	63.6	46.9	67.8	66.1	73.1	77.9	66.8	68.7	73.4	70.4	63.8 ± 1.7
+MixTTA	**55.8**	**56.1**	**56.8**	57.8	**59.3**	**64.9**	**63.5**	**69.0**	**67.2**	**73.8**	**78.4**	**67.9**	**69.3**	**74.3**	**71.0**	**65.7** ± 0.1

Table 2: Comparisons with baselines on ImageNet-C at severity level 5 under online imbalanced label shifts with imbalance ratio ∞ regarding accuracy (%).

applied to the second normalization layer in each of the first five transformer encoder blocks. The subspace dimension r and spectral projection strength η were fixed to 4 and 0.9 across all experiments, respectively. Due to the page limit, full ResNet results are provided in the supplementary material; MixTTA consistently improves most baselines on ImageNet-C and ImageNet-Sketch across all four adaptation scenarios.

4.3 Main Results

Comparison on Mild Scenario Table 1 presents the results on the ImageNet-C dataset at severity level 5 under the mild scenario. Integrating MixTTA improves the average accuracy across all five baselines, confirming its generality as a plug-in module. In particular, for ReCAP, MixTTA delivers a larger improvement (+0.9%p) in average accuracy than applying LinearTCA (+0.5%p). Furthermore, the gains are especially pronounced on challenging *Noise* and *Weather* corruptions. For example, MixTTA improves Tent from 18.1% to 57.2% on *Snow* and from 21.2% to 41.6% on *Frost*. These results suggest that MixTTA is particularly beneficial under corruptions where baseline adaptation is most fragile.

Batch Size 1	Noise			Blur				Weather				Digital				Avg.
	Gauss.	Shot	Impulse	Defocus	Glass	Motion	Zoom	Snow	Frost	Fog	Brit.	Contr.	Elastic	Pixel	JPEG	
No Adapt.	16.8	12.0	16.5	29.2	23.6	34.0	27.3	15.8	26.6	47.5	55.4	44.3	31.0	45.1	49.1	31.6
• LinearTCA	17.9	13.0	17.4	29.8	24.2	35.7	28.8	17.0	27.8	49.8	56.0	45.8	33.0	46.0	49.8	32.8 ± 0.0
• LinearTCA$^+$	56.1	56.9	57.5	59.1	60.4	65.6	61.1	69.3	67.5	74.0	78.6	68.0	70.3	74.6	71.5	66.0 ± 0.0
• Tent	**44.7**	**42.9**	**45.4**	**52.5**	48.1	55.5	50.6	15.6	18.6	66.6	75.1	64.9	52.9	67.1	64.6	51.0 ± 0.1
+MixTTA	41.7	36.7	41.0	51.8	**48.8**	**56.4**	**51.3**	**57.9**	**38.8**	**68.3**	**75.4**	**65.2**	**54.1**	**68.4**	**65.4**	**54.8** ± 2.5
• EATA	36.4	30.7	35.4	44.7	40.0	46.4	41.8	38.5	39.1	61.8	65.9	61.6	47.1	59.7	59.7	47.2 ± 0.5
+MixTTA	**49.7**	**47.6**	**51.0**	**55.2**	**54.4**	**60.5**	**56.3**	**63.8**	**62.3**	**71.5**	**77.1**	**66.7**	**61.8**	**72.0**	**68.4**	**61.2** ± 0.3
• SAR	45.6	43.2	46.3	**53.6**	**50.5**	**57.6**	**53.0**	58.7	54.9	68.9	75.4	**65.7**	**58.1**	**69.0**	**66.4**	57.8 ± 0.3
+MixTTA	**47.1**	**44.9**	**48.4**	52.0	49.5	57.0	51.6	**59.5**	**58.7**	**69.0**	**76.0**	65.4	56.0	67.4	66.1	**57.9** ± 0.2
• DeYO	55.2	55.6	56.3	**58.9**	59.4	64.3	44.2	68.1	66.5	73.8	78.4	68.3	69.0	73.9	70.9	64.2 ± 0.6
+MixTTA	**56.2**	**56.8**	**57.2**	58.6	**59.9**	**65.6**	**63.9**	**69.3**	**67.6**	**74.5**	**78.8**	**69.0**	**69.4**	**74.7**	**71.5**	**66.2** ± 0.1
• ReCAP	56.1	56.8	57.2	**59.2**	60.0	65.5	57.0	69.2	67.3	74.0	78.5	67.8	70.1	74.4	71.3	65.6 ± 0.8
+MixTTA	**56.6**	**57.5**	**57.9**	58.6	**60.6**	**66.3**	**66.5**	**70.1**	**68.3**	**74.5**	**78.8**	**67.9**	**70.8**	**75.0**	**71.8**	**66.8** ± 0.1

Table 3: Comparisons with baselines on ImageNet-C at severity level 5 under batch size 1 regarding accuracy (%).

Comparison on Wild Scenario Table 2 reports results under the online imbalanced label shift setting (imbalance ratio ∞) on ImageNet-C at severity level 5. Across a wide range of corruptions, MixTTA consistently improves existing TTA methods, with ReCAP increasing from 63.8% to 65.7% and DeYO from 64.8% to 65.6% on average. Beyond the average gains, MixTTA is particularly beneficial in difficult regimes where vanilla adaptation becomes fragile. For example, Tent on *Snow* drops accuracy from 15.9% (No Adapt.) to 10.2%, but combining it with MixTTA recovers and boosts the accuracy to 45.5%. These results indicate that our proposed MixTTA can effectively counteract drift and instability induced by heavily skewed target streams.

Table 3 summarizes the ImageNet-C results at severity level 5 in the batch size 1 setting, where adaptation must rely solely on instance-level statistics. MixTTA consistently improves strong baselines, boosting the average accuracy of DeYO to 66.2% and ReCAP to 66.8%. Notably, the gains are particularly evident for methods that degrade under single-sample updates (*e.g.*, EATA improves from 47.2% to 61.2% on average), and MixTTA yields broad improvements across most corruption categories, including several *Weather* and *Digital* corruptions. Overall, these results indicate that MixTTA remains effective even under highly constrained adaptation, improving robustness while maintaining stable adaptation dynamics.

The results under the mixed-shift setting are summarized in Table 4. MixTTA consistently strengthens adaptation performance across all baselines. Notably, while LinearTCA$^+$ yields only a marginal improvement (+0.2%p) in this setting, MixTTA provides substantially larger gains (+2.8%p for DeYO and +2.3%p for ReCAP), highlighting the benefit of its cross-channel correction under mixed shifts. Taken together, these observations suggest that MixTTA is particularly effective when multiple types of distribution shift co-occur and per-channel adaptation alone is most insufficient.

Results on ImageNet-Sketch Table 5 reports accuracy on ImageNet-Sketch under mild and wild settings. MixTTA consistently improves existing TTA baselines, with gains becoming markedly larger in wild regimes. The most pro-

Mixed Shifts	Accuracy (%)
No adapt.	31.6
• LinearTCA	32.8 ± 0.1
• LinearTCA$^+$	60.3 ± 0.2
• Tent	16.2 ± 2.5
+ MixTTA	**26.9** ± 2.7
• EATA	56.3 ± 0.4
+ MixTTA	**61.1** ± 0.1
• SAR	57.7 ± 0.0
+ MixTTA	**58.9** ± 0.1
• DeYO	59.7 ± 0.1
+ MixTTA	**62.5** ± 0.1
• ReCAP	60.1 ± 0.0
+ MixTTA	**62.4** ± 0.1

Table 4: Comparisons with baselines on ImageNet-C at severity level 5 under a mixture of 15 types of corruption regarding accuracy (%).

ImageNet -Sketch	Mild	Label Shifts	Batch Size 1
No Adapt.	18.2	18.3	18.2
• LinearTCA	19.3 ± 0.0	19.6 ± 0.0	19.3 ± 0.0
• LinearTCA$^+$	42.2 ± 0.2	44.6 ± 0.4	43.8 ± 0.1
• Tent	8.8 ± 0.4	5.3 ± 1.4	6.1 ± 0.8
+ MixTTA	**34.3** ± 0.2	**29.3** ± 11.6	**34.1** ± 0.3
• EATA	36.9 ± 0.1	33.5 ± 0.0	21.5 ± 0.1
+ MixTTA	**41.3** ± 0.3	**37.2** ± 0.2	**36.2** ± 0.5
• SAR	26.3 ± 2.9	13.6 ± 8.8	18.3 ± 6.3
+ MixTTA	**35.6** ± 0.1	**39.3** ± 0.1	**35.5** ± 0.0
• DeYO	43.6 ± 0.3	43.2 ± 1.0	43.4 ± 0.8
+ MixTTA	**43.7** ± 0.1	**44.5** ± 0.3	**44.0** ± 0.0
• ReCAP	42.1 ± 0.1	44.4 ± 0.4	42.9 ± 0.1
+ MixTTA	**42.9** ± 0.0	**45.1** ± 0.1	**45.9** ± 0.1

Table 5: Comparisons with baselines on ImageNet-Sketch under mild and wild scenarios regarding accuracy (%).

Method	r					Full rank
	1	2	4	8	16	
Tent+MixTTA	53.7	54.7	**55.2**	55.0	54.6	0.1
EATA+MixTTA	62.4	62.5	**63.5**	62.7	62.5	0.1
SAR+MixTTA	55.0	55.9	**57.2**	56.6	55.9	0.1
DeYO+MixTTA	64.0	65.1	**65.6**	65.4	64.9	0.1
ReCAP+MixTTA	63.6	64.2	**64.9**	64.1	63.6	0.1

Table 6: Comparison of average accuracy of MixTTA performance with varying r values and baselines on ImageNet-C severity level 5. Full rank denotes replacing Γ with a full channel mixing transform $W \in \mathbb{R}^{C\times C}$.

nounced effect is on Tent, where MixTTA recovers severely degraded performance by at least 24 percentage points across all three settings, indicating that it effectively stabilizes fragile entropy-minimization updates. Moreover, ReCAP with MixTTA outperforms LinearTCA$^+$ across all settings, demonstrating that MixTTA provides complementary benefits beyond explicit correlation alignment. These results suggest that the cross-channel correction provided by MixTTA is not limited to corruption-based shifts but extends to style-level domain gaps.

4.4 Analysis

Sensitivity of Subspace Dimension r. Table 6 studies the sensitivity of baselines with MixTTA to the low-rank parameterization rank r on ImageNet-C at severity level 5. We observe that performance improves as r increases from 1 to 4, reaching the best average accuracy at $r = 4$, indicating that a modest rank is sufficient to capture the dominant cross-channel shift required at test time. However, further increasing the rank to 8 or 16 degrades accuracy, suggesting that overly expressive channel mixing introduces unnecessary degrees of freedom

Method	# Params	GPU Time (50,000)	Acc. (%)
DeYO	27,648	477 seconds	63.8
DeYO (Full block)	36,864	481 seconds	62.1
DeYO+MixTTA	46,080	496 seconds	**65.6**
ReCAP	27,648	451 seconds	64.0
ReCAP (Full block)	36,864	455 seconds	62.4
ReCAP+MixTTA	46,080	470 seconds	**64.9**

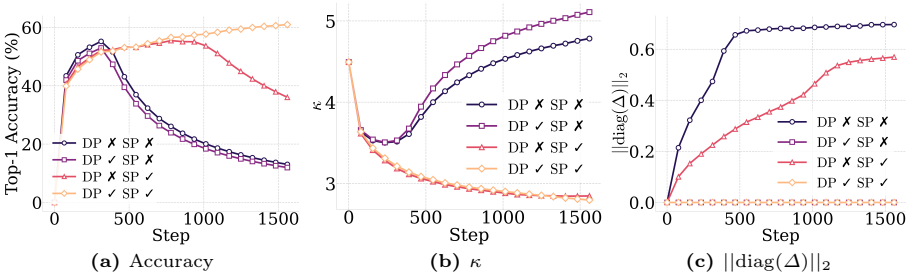
Table 7: Comparison of learnable parameter counts, runtime, and accuracy for DeYO, DeYO+MixTTA, ReCAP, and ReCAP+MixTTA, respectively. We also report baselines extended to update all transformer encoder blocks. The practical runtime is evaluated using a single A5000 GPU.

and leads to less stable adaptation. At the limit, the results demonstrate that replacing the diagonal affine scale matrix Γ in Eq. (3) with a full channel-mixing transform $W \in \mathbb{R}^{C\times C}$ ("Full rank") causes catastrophic collapse. We attribute this to the substantially larger parameterization, which weakens the inductive bias of affine-style modulation and leads to unstable online optimization and severe overfitting. These results collectively suggest that the low-rank parameterization effectively regularizes the update dynamics, maintaining robustness under severe corruption. Notably, $r = 4$ yields the best accuracy across all five baselines simultaneously, eliminating the need for per-method hyperparameter tuning.

Overhead and Parameter-controlled Analysis. Table 7 reports the learnable parameter count, GPU running time (50,000 samples), and average accuracy on the ImageNet-C dataset at severity level 5 of DeYO and ReCAP with and without MixTTA. MixTTA adds approximately 4% runtime overhead for both DeYO and ReCAP, confirming that the additional computation is negligible in practice.

We further disentangle the effect of MixTTA from simply increasing the number of learnable parameters by comparing against the full-block update variants of DeYO and ReCAP. Naively extending adaptation to all transformer encoder blocks increases the parameter count but degrades performance for both methods, suggesting that deeper online updates can be unstable or prone to overfitting. In contrast, integrating MixTTA consistently improves these baselines, indicating that the gains stem from the structured low-rank cross-channel transformation of MixTTA rather than from simply increasing model capacity.

Ablation Study of Projection Strategies. To examine the individual efficacy of Decoupling Projection (DP) and Spectral Projection (SP), we track accuracy, feature covariance condition number $\kappa = \lambda_{\max}/\lambda_{\min}$, and the norm of the diagonal component $\|\operatorname{diag}(\Delta)\|_2$ over adaptation steps on ImageNet-C *snow* corruption at severity level 5 under imbalanced label shift, across all DP/SP configurations.



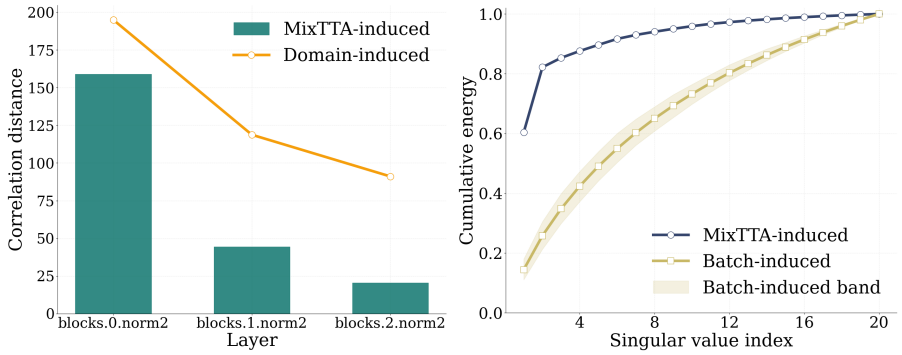
(a) Accuracy **(b)** κ **(c)** $||\text{diag}(\Delta)||_2$

Fig. 3: Comparison of (a) top-1 accuracy, (b) the condition number κ of the feature covariance at an intermediate layer, and (c) $||\,\text{diag}(\Delta)||_2$ for adapted models on ImageNet-C *snow* (severity 5) under imbalanced label shift, across different projection configurations.

Fig. 3a shows that enabling both DP and SP yields the most stable accuracy trajectory, preventing the performance collapse observed in other configurations. Particularly, when SP is disabled, accuracy begins to drop sharply, and this transition aligns with the rapid growth of the condition number in Fig. 3b, indicating that the failure coincides with a rank-1-dominant covariance regime where the leading eigenvalue overwhelms the spectrum. In contrast, enabling SP keeps κ bounded and thereby stabilizes the adaptation dynamics, explaining its critical role in preventing collapse.

Consistent with the design goal of DP, enabling DP effectively suppresses diagonal leakage in the low-rank branch, yielding a markedly smaller $\|\,\text{diag}(\Delta)\|_2$ than DP-off configurations (Fig. 3c). When DP is disabled, $\|\,\text{diag}(\Delta)\|_2$ grows progressively over adaptation steps, indicating increasing diagonal leakage. In the DP-off/SP-on setting, accuracy degradation coincides with a drastic rise in $\|\,\text{diag}(\Delta)\|_2$. This suggests that diagonal leakage may partially undermine the intended cross-channel mixing effect of the low-rank branch by reintroducing per-channel modulation.

Analysis of Correlation Shifts. To examine how MixTTA modifies cross-channel correlations during adaptation, we compute the correlation changes on ImageNet-C *Gaussian noise* at severity level 5. In Fig. 4a, we visualize two layer-wise quantities: the MixTTA-induced shift, computed as the correlation distance between Tent and Tent+MixTTA on the same corrupted input, and the domain-induced shift, computed as the correlation distance between source and corrupted features under the source model. Although these two correlation shifts have different origins, both quantities peak at the earliest block and decrease in deeper blocks, indicating that MixTTA applies larger corrections at layers where correlation disruption is most severe. Notably, this layer-wise allocation arises purely from entropy-driven optimization, without any explicit layer-dependent weighting, suggesting that the low-rank branch adaptively responds to cross-channel misalignment across layers.



(a) Correlation Distance **(b)** Cumulative Energy

Fig. 4: Layer-wise and spectral analysis of feature correlation changes on ImageNet-C *Gaussian noise* at severity level 5. (a) Layer-wise correlation distances. (b) Cumulative energy of the singular-value spectrum at `blocks.0.norm2`.

Fig. 4b further examines the spectral structure of the MixTTA-induced correlation shift. Specifically, we plot how rapidly the cumulative spectral energy of the MixTTA-induced correction concentrates as a function of singular-value index. As a control, we estimate batch-induced variation by randomly splitting a target mini-batch into two halves and measuring the correlation difference between the two half-batch means; we report its mean curve together with the q=0.1–0.9 quantile band across random splits. Compared to this batch-noise baseline, the MixTTA-induced correction concentrates energy in markedly fewer singular directions, forming a structured, low-rank adjustment rather than a diffuse perturbation. This is consistent with the intended design of the low-rank module.

5 Conclusion

We identified that distribution shift disrupts inter-channel feature correlations in a way that conventional per-channel affine updates cannot correct, and that this disruption is most severe in earlier network layers. Based on this finding, we proposed MixTTA, a plug-in module that extends normalization layers with a low-rank cross-channel transformation, enabling inter-channel mixing across multiple layers. Together with Decoupling Projection to enforce strict diagonal/off-diagonal separation and Spectral Projection to prevent rank-1 collapse, MixTTA consistently improves existing normalization-based TTA methods across standard and wild scenarios on ImageNet-C and ImageNet-Sketch, while incurring only minimal overhead. One future direction is to develop an adaptive rank selection strategy that adjusts the subspace dimension to the severity and structure of the encountered domain shift. We hope that the perspective of cross-channel correlation modeling opens new directions for more expressive and reliable test-time adaptation in the community.

Acknowledgements. This work was supported by the Institute of Information & Communications Technology Planning & Evaluation (IITP) under the AI Star Fellowship (Kookmin University) (RS-2025-02219317 (10%)) and the Leading Generative AI Human Resources Development (IITP-2026-RS-2026-25546026 (10%)) grant funded by the Korea government (MSIT), and the "Advanced GPU Utilization Support Program" funded by the Government of the Republic of Korea (Ministry of Science and ICT), the National Research Foundation of Korea (NRF, RS-2024-00451435 (10%), RS-2024-00413957 (10%)), the Institute of Information & Communications Technology Planning & Evaluation (IITP, RS-2025-02305453 (15%), RS-2025-02273157 (15%), RS-2025-25442149 (15%), RS-2021-II211343 (15%)) grant funded by the Ministry of Science and ICT (MSIT), the Institute of New Media and Communications (INMAC), and the BK21 FOUR program of the Education, the Artificial Intelligence Graduate School Program (Seoul National University), and the Research Program for Future ICT Pioneers, Seoul National University in 2026.

References

1. Ba, J.L., Kiros, J.R., Hinton, G.E.: Layer normalization. arXiv preprint arXiv:1607.06450 (2016)
2. Cha, J., Chun, S., Lee, K., Cho, H.C., Park, S., Lee, Y., Park, S.: Swad: Domain generalization by seeking flat minima. Advances in Neural Information Processing Systems **34**, 22405–22418 (2021)
3. Chen, C., Fu, Z., Chen, Z., Jin, S., Cheng, Z., Jin, X., Hua, X.: Homm: Higher-order moment matching for unsupervised domain adaptation. In: The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020. pp. 3422–3429. AAAI Press (2020). https://doi.org/10.1609/AAAI.V34I04.5745, https://doi.org/10.1609/aaai.v34i04.5745
4. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N.: An image is worth 16x16 words: Transformers for image recognition at scale. In: 9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021. OpenReview.net (2021), https://openreview.net/forum?id=YicbFdNTTy
5. Fan, X., Jiang, J., Chen, Z., Huang, F., Chen, X., Jiang, Q., Zhang, B., Tang, X., Wang, Z.: Moetta: Test-time adaptation under mixed distribution shifts with moe-layernorm. In: Koenig, S., Jenkins, C., Taylor, M.E. (eds.) Fortieth AAAI Conference on Artificial Intelligence, Thirty-Eighth Conference on Innovative Applications of Artificial Intelligence, Sixteenth Symposium on Educational Advances in Artificial Intelligence, AAAI 2026, Singapore, January 20-27, 2026. pp. 21011–21019. AAAI Press (2026). https://doi.org/10.1609/AAAI.V40I25.39243, https://doi.org/10.1609/aaai.v40i25.39243
6. Ganin, Y., Lempitsky, V.: Unsupervised domain adaptation by backpropagation. In: International conference on machine learning. pp. 1180–1189. PMLR (2015)

7. Glorot, X., Bengio, Y.: Understanding the difficulty of training deep feedforward neural networks. In: Proceedings of the thirteenth international conference on artificial intelligence and statistics. pp. 249–256. JMLR Workshop and Conference Proceedings (2010)
8. He, K., Gkioxari, G., Dollár, P., Girshick, R.: Mask r-cnn. In: Proceedings of the IEEE international conference on computer vision. pp. 2961–2969 (2017)
9. Hendrycks, D., Dietterich, T.: Benchmarking neural network robustness to common corruptions and perturbations. In: International Conference on Learning Representations (2019)
10. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W.: Lora: Low-rank adaptation of large language models. In: International Conference on Learning Representations (2022)
11. Hu, Z., Hu, Y., Li, X., Tang, S., Duan, L.: Beyond entropy: Region confidence proxy for wild test-time adaptation. In: International Conference on Machine Learning. pp. 24371–24390. PMLR (2025)
12. Imam, R., Gani, H., Huzaifa, M., Nandakumar, K.: Test-time low rank adaptation via confidence maximization for zero-shot generalization of vision-language models. In: 2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). pp. 5449–5459. IEEE (2025)
13. Ioffe, S., Szegedy, C.: Batch normalization: Accelerating deep network training by reducing internal covariate shift. In: International conference on machine learning. pp. 448–456. pmlr (2015)
14. Iwasawa, Y., Matsuo, Y.: Test-time classifier adjustment module for model-agnostic domain generalization. Advances in Neural Information Processing Systems **34**, 2427–2440 (2021)
15. Jia, M., Tang, L., Chen, B.C., Cardie, C., Belongie, S., Hariharan, B., Lim, S.N.: Visual prompt tuning. In: European Conference on Computer Vision (ECCV). pp. 709–727 (2022)
16. Jung, M., Jeong, S., Kim, Y., Lee, J.: Edas: Effective data augmentation strategies for test-time adaptation. ICT Express (2025)
17. Kim, Y.: Leveraging single positive class label supervision for weakly supervised semantic segmentation. IEEE Access (2025)
18. Kim, Y., Kim, S., Ro, Y., Lee, J.: Instance-dependent multilabel noise generation for multilabel remote sensing image classification. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing **17**, 17087–17098 (2024)
19. Kojima, Y., Xu, J., Zou, X., Wang, X.: Lora-ttt: Low-rank test-time training for vision-language models. arXiv preprint arXiv:2502.02069 (2025)
20. Kwon, E., Zhou, M., Xu, W., Rosing, T., Kang, S.: Rl-ptq: Rl-based mixed precision quantization for hybrid vision transformers. In: Proceedings of the 61st ACM/IEEE Design Automation Conference. pp. 1–6 (2024)
21. Lee, D., Yoon, J., Hwang, S.J.: Becotta: Input-dependent online blending of experts for continual test-time adaptation. In: Salakhutdinov, R., Kolter, Z., Heller, K.A., Weller, A., Oliver, N., Scarlett, J., Berkenkamp, F. (eds.) Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024. Proceedings of Machine Learning Research, vol. 235, pp. 27072–27093. PMLR / OpenReview.net (2024), https://proceedings.mlr.press/v235/lee24ab.html
22. Lee, H., Jang, C., Lee, D.B., Lee, J.: Dimension agnostic neural processes. In: International Conference on Learning Representations. vol. 2025, pp. 12094–12127 (2025)

23. Lee, J., Jung, D., Lee, S., Park, J., Shin, J., Hwang, U., Yoon, S.: Entropy is not enough for test-time adaptation: From the perspective of disentangled factors. In: The Twelfth International Conference on Learning Representations (2024), https://openreview.net/forum?id=9w3iw8wDuE
24. Lester, B., Al-Rfou, R., Constant, N.: The power of scale for parameter-efficient prompt tuning. In: Proceedings of the 2021 conference on empirical methods in natural language processing. pp. 3045–3059 (2021)
25. Liang, J., Hu, D., Feng, J.: Do we really need to access the source data? source hypothesis transfer for unsupervised domain adaptation. In: International conference on machine learning. pp. 6028–6039. PMLR (2020)
26. Liu, J., Yang, S., Jia, P., Zhang, R., Lu, M., Guo, Y., Xue, W., Zhang, S.: Vida: Homeostatic visual domain adapter for continual test time adaptation. In: The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024. OpenReview.net (2024), https://openreview.net/forum?id=sJ88Wg5Bp5
27. Miyato, T., Kataoka, T., Koyama, M., Yoshida, Y.: Spectral normalization for generative adversarial networks. In: International Conference on Learning Representations (2018)
28. Morerio, P., Cavazza, J., Murino, V.: Minimal-entropy correlation alignment for unsupervised deep domain adaptation. In: International Conference on Learning Representations (2018)
29. Niu, S., Wu, J., Zhang, Y., Chen, Y., Zheng, S., Zhao, P., Tan, M.: Efficient test-time model adaptation without forgetting. In: International conference on machine learning. pp. 16888–16905. PMLR (2022)
30. Niu, S., Wu, J., Zhang, Y., Wen, Z., Chen, Y., Zhao, P., Tan, M.: Towards stable test-time adaptation in dynamic wild world. In: International Conference on Learning Representations (2023), https://openreview.net/forum?id=g2YraF75Tj
31. Park, M., Won, H., Ro, W.W., Kim, S.: Rethinking entropy in test-time adaptation: The missing piece from energy duality. Advances in Neural Information Processing Systems **38**, 47123–47140 (2026)
32. Pfeiffer, J., Kamath, A., Rücklé, A., Cho, K., Gurevych, I.: Adapterfusion: Non-destructive task composition for transfer learning. In: Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics (EACL). pp. 487–503 (2021)
33. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., et al.: Imagenet large scale visual recognition challenge. International journal of computer vision **115**(3), 211–252 (2015)
34. Saito, K., Watanabe, K., Ushiku, Y., Harada, T.: Maximum classifier discrepancy for unsupervised domain adaptation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 3723–3732 (2018)
35. Shi, Y., Seely, J., Torr, P.H.S., Narayanaswamy, S., Hannun, A.Y., Usunier, N., Synnaeve, G.: Gradient matching for domain generalization. In: The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022. OpenReview.net (2022), https://openreview.net/forum?id=vDwBW49HmO
36. Shin, J., Kim, Y., Hong, S., Lee, J.: Learning dual hierarchical representation for 3d surface reconstruction. In: Proceedings of the Asian Conference on Computer Vision. pp. 4422–4438 (2024)
37. Shin, J., Kim, Y., Lee, J.: A plug-in curriculum scheduler for improved deformable medical image registration. IEEE Access (2025)

38. Sun, B., Saenko, K.: Deep coral: Correlation alignment for deep domain adaptation. In: European conference on computer vision. pp. 443–450. Springer (2016)
39. Um, S., Kim, B., Ye, J.C.: Boost-and-skip: A simple guidance-free diffusion for minority generation. In: International Conference on Machine Learning. pp. 60561–60589. PMLR (2025)
40. Wang, D., Shelhamer, E., Liu, S., Olshausen, B., Darrell, T.: Tent: Fully test-time adaptation by entropy minimization. In: International Conference on Learning Representations (2021), https://openreview.net/forum?id=uXl3bZLkr3c
41. Wang, H., Ge, S., Lipton, Z., Xing, E.P.: Learning robust global representations by penalizing local predictive power. Advances in neural information processing systems **32** (2019)
42. Wang, Q., Fink, O., Van Gool, L., Dai, D.: Continual test-time domain adaptation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 7201–7211 (2022)
43. Wightman, R.: Pytorch image models. https://github.com/rwightman/pytorch-image-models, accessed 2026-06-23 (2019). https://doi.org/10.5281/zenodo.4414861
44. Wu, Y., He, K.: Group normalization. In: Proceedings of the European conference on computer vision (ECCV). pp. 3–19 (2018)
45. Yang, S., Van de Weijer, J., Herranz, L., Jui, S., et al.: Exploiting the intrinsic neighborhood structure for source-free domain adaptation. Advances in neural information processing systems **34**, 29393–29405 (2021)
46. You, L., Lu, J., Huang, X.: Test-time correlation alignment. In: International Conference on Machine Learning. pp. 72700–72729. PMLR (2025)
47. Zellinger, W., Grubinger, T., Lughofer, E., Natschläger, T., Saminger-Platz, S.: Central moment discrepancy (cmd) for domain-invariant representation learning. In: International Conference on Learning Representations (ICLR) (2017)
48. Zhang, M., Levine, S., Finn, C.: Memo: Test time robustness via adaptation and augmentation. Advances in neural information processing systems **35**, 38629–38642 (2022)
49. Zhang, Q., Bian, Y., Kong, X., Zhao, P., Zhang, C.: COME: test-time adaption by conservatively minimizing entropy. In: The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025. OpenReview.net (2025), https://openreview.net/forum?id=506BjJ1ziZ