INTRINSIC SELF-DISTILLATION FOR TEST-TIME ADAPTATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Test-time adaptation has emerged as a promising approach for addressing distribution shifts in target domains by improving predictions during inference on online data streams. For decades, the utilization of unlabeled data has been widely explored in semi-supervised and self-supervised learning, and provides a blueprint for test-time adaptation. Building on pioneering insights, recent solutions achieve notable improvements with data augmentation or auxiliary modules. However, theses approaches demand additional computational complexity and reduce practical applicability. In this paper, we introduce intrinsic self-distillation for adapting to the target domain using intermediate representations, without not only augmentation but also auxiliary modules. Our method leverages intermediate representations as intrinsic learning signals, motivated by two empirical observations: first, the outputs of deeper layers exhibit progressively higher accuracy. Second, predictions from intermediate representations are more robust to noise. We demonstrate the efficiency and effectiveness of our method across various architectures and benchmarks.

1 INTRODUCTION

Addressing distribution shifts in practical scenarios has been a long-standing challenge in machine learning across traditional methods and modern deep neural networks (DNNs) to enable robust real-world applicability (Blitzer et al., 2007; Duan et al., 2009; Gong et al., 2012). Central to these approaches is transferring knowledge from the source to the target domain to mitigate domain shift (Ben-David et al., 2006; 2010). Recently, test-time adaptation (TTA) has emerged as a promising approach that shifts conventional offline domain adaptation to an online paradigm, enabling immediate adaptation without prior knowledge of the target domain (Wang et al., 2021; Shu et al., 2022; Wang et al., 2024). Given that TTA is designed for settings with limited or no access to source data, it is inherently challenging to directly address domain discrepancies (Mitsuzumi et al., 2024). Consequently, TTA shares conceptual similarities with self-training and requires stable self-prediction refinement on test data with computational efficiency.

In this context, TTA is formulated as an online unsupervised learning paradigm under computational constraints, building on extensive research in unsupervised learning (Wang et al., 2022a; Niu et al., 2025; Chen et al., 2025). In particular, entropy minimization (Grandvalet & Bengio, 2004) is the primary approach in TTA (Wang et al., 2021), and semi-supervised learning techniques such as pseudo-labeling (Lee et al., 2013) and temporal ensembling (Laine & Aila, 2017; Tarvainen & Valpola, 2017) are also widely used (Wang et al., 2022a; Goyal et al., 2022; Hu et al., 2025a). Meanwhile, self-supervised learning inspired approaches leveraging contrastive learning, data augmentation, and asymmetric architectures have been proposed to mitigate model collapse caused by trivial solutions (Chen et al., 2022; Han et al., 2025; Chen et al., 2025). Notably, the asymmetric approach provided a direction for advancement toward efficient TTA by suggesting a method to improve the final prediction utilizing only a single forward pass (Chen et al., 2025). However, under the constraint of a single forward pass, adaptation based solely on final model predictions is limited, thereby requiring additional intrinsic signals for effective refinement.

The utilization of intermediate representations in DNNs has been proposed to alleviate gradient vanishing and improve the discriminative capability of intermediate layers (Szegedy et al., 2015; Lee et al., 2015). Furthermore, self-knowledge distillation methods employing auxiliary classifiers at

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.

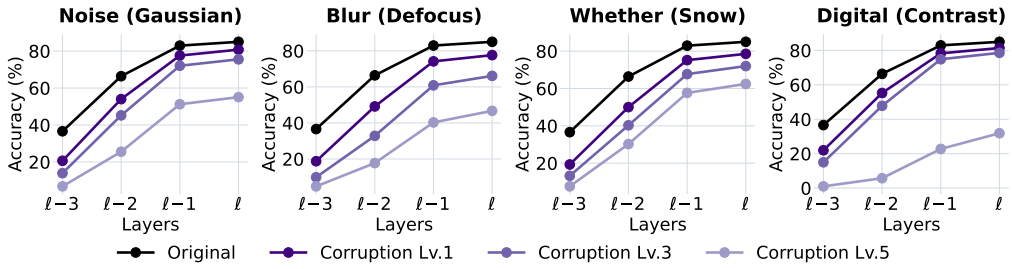


Figure 1: Layer-wise accuracy with respect to corruption severity. Across diverse corruption types in ImageNet-C, the ViT-Base source model exhibits a consistent degree of performance degradation across layers, indicating that responses to corruption shifts are consistently reflected across layers.

intermediate layers for improved self-prediction within a single network have achieved promising results (Zhang et al., 2019). Especially, Wei et al. (2021) provide theoretical justification for the usage of intermediate representations by proving that the generalization error can be bounded by the margins of perturbation stability across the entire layer. The theoretical foundations of generalization bounds have provided the basis for understanding weak-to-strong generalization (Lang et al., 2024) and mitigating catastrophic forgetting in continual learning (Zheng et al., 2024). Furthermore, prior studies Mirza et al. (2023); Niu et al. (2024) on TTA have mitigated domain shift by aligning the statistics of intermediate features between the source and target domains. In this paper, we focus on the fully TTA setting, where source domain data are inaccessible, and explore the potential of intermediate features for TTA.

Figure 1 presents the classification accuracy obtained by applying a shared classifier to feature representations from relatively shallow layers. Empirical observations show that performance increases gradually with layer depth across different source domains and various corruption shifts, and that the rate of performance degradation changes uniformly with increasing corruption severity. This suggests consistent responses to domain shift across all layers rather than confinement to the final output, indicating the potential for improvement from both final and intermediate representations. Our analysis shows lower sensitivity of intermediate layers to noisy supervision signals compared to final predictions, along with higher initial accuracy in deeper layers (see Section 3.2).

Motivated by the aforementioned observations, we introduce **i**ntrinsic **s**elf-**d**istillation (**ISD**), a unified self-distillation framework for TTA. Our approach is designed with a focus on adaptability, stability, and efficiency. Adaptability is achieved through gradual feature distillation, transferring supervision from deeper layers with higher accuracy to shallower layers that are less sensitive to noise. Stability is ensured via debiased logit distillation, which corrects directional bias in the features and thereby mitigates model collapse to a single vector. Efficiency is preserved by maintaining a single forward pass without additional modules. Compared to state-of-the-art methods, our ISD achieves improved performance across diverse TTA scenarios, demonstrating its effectiveness.

2 Related Works

The main goal of TTA is to improve predictions under domain shift by adapting to input data during inference. Over the past five years, it has emerged as an important research topic, demonstrating practical benefits in addressing domain shifts encountered in the wild. In this section, we categorize existing approaches into three groups to clarify the differences between our method and prior work.

Prediction-based approaches. Entropy minimization is a central approach in TTA, based on the observation that lower entropy under domain shift correlates with lower negative log-likelihood (Wang et al., 2021). Nonetheless, entropy minimization can suffer from instability caused by model collapse, where predictions converge to a trivial solution irrespective of the input. To address this issue, several methods utilize thresholding strategies that restrict model updates to low-entropy samples (Niu et al., 2022), adopting sharpness-aware minimization (Niu et al., 2023), and leverage prediction consistency under object-destructive transformations for identifying helpful samples (Lee et al., 2024). In addition, statistics-based methods provide computational efficiency by statistical

estimation without backpropagation. The approaches refine decision boundaries via the concave-convex procedure (Boudiaf et al., 2022) or align statistical distributions through covariance matrix adaptation and embedding alignment (Xiao et al., 2026).

Data-based approaches. Data augmentation techniques generate positive samples for regularization and have achieved promising results in semi-supervised and self-supervised learning, with broad use in unsupervised learning (Chen et al., 2020; Sohn et al., 2020). The first work on test-time training (Sun et al., 2020) adopts rotation prediction for adaptation, provided the groundwork for TTA solutions. In particular, leveraging averaged predictions over diverse augmentations is one of the key approaches in TTA (Zhang et al., 2022; Shu et al., 2022; Farina et al., 2024). Moreover, explicit augmentation strategies, including low-frequency masking and high-frequency enhancement (Niu et al., 2025), and context masking to strengthen input–prediction causal relationships (Han et al., 2025), improve adaptation stability.

Model-based approaches. Approaches incorporating additional modules or retaining the initial model mitigate catastrophic forgetting and demonstrate advantages in continual settings. For instance, Wang et al. (Wang et al., 2022a) propose the continual TTA scenario and employ a teacher–student framework to enable stable adaptation. Furthermore, to alleviate the inefficiency associated with additional models, recent approaches have increasingly focused on parameter-efficient adaptation, evolving towards more compact architectures (Song et al., 2023; Liu et al., 2024; Shin & Kim, 2024). Besides, asymmetric frameworks using a single multilayer perceptron predictor achieve both stability and efficiency, with model-based approaches improving efficiency by maintaining a single model and a single forward pass (Chen et al., 2025).

What matters in TTA? Adaptation mechanisms during inference demand computational efficiency, immediate adaptability, and stability. Prediction-based approaches (Wang et al., 2021; Niu et al., 2023) enable efficient adaptation but suffer from instability due to model collapse, while statistics-based methods (Boudiaf et al., 2022; Niu et al., 2024) require preprocess for source-domain statistics or exhibit limited adaptability across multiple domains. In contrast, data-based approaches (Marsden et al., 2024; Niu et al., 2025; Han et al., 2025) improve representations and offer strong adaptability, but induce computational overhead from additional forward passes. Finally, model-based approaches (Wang et al., 2022a; Song et al., 2023) enable stable adaptation and partially alleviate the efficiency issues caused by additional modules; however, the adaptability of these approaches is limited by conservative updates for stability. The strengths and weaknesses of these approaches are illustrated in Figure 2. In this work, our goal is to extend prediction-based approaches by leveraging reliable intrinsic signals within a single model, without data augmentation or additional modules, to achieve efficiency, adaptability, and stability.

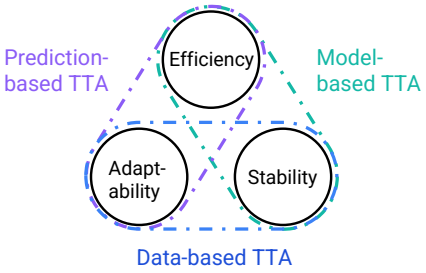


Figure 2: Pros and cons in TTA methods.

3 PRELIMINARY

3.1 PROBLEM DEFINITION

In TTA, we consider a source model pretrained on a labeled source dataset $\{\mathcal{X}_s, \mathcal{Y}_s\}$, where $x_s \in \mathcal{X}_s$ and $y_s \in \mathcal{Y}_s$ denote a source image and the corresponding label, respectively. The source model consists of a feature extractor $f_s(\cdot)$ and a classifier $g_s(\cdot)$. For the target domain, the feature extractor $f_t(\cdot)$ and classifier $g_t(\cdot)$ are initialized from the source model. During adaptation on the target test image x_t, the target labels y_t are unavailable and used only for evaluation in experimental settings. The objective is to minimize the discrepancy between the prediction $\hat{y}_t = \arg\max(p_t)$ and the corresponding label y_t, where $p_t = \sigma(z_t)$ is the predicted probability distribution, $z_t = g_t(f_t(x_t))$ denotes the logits, and $\sigma(\cdot)$ represents the softmax operator. In this paper, we consider intermediate representations whose feature dimensionality matches that of the final feature representation.

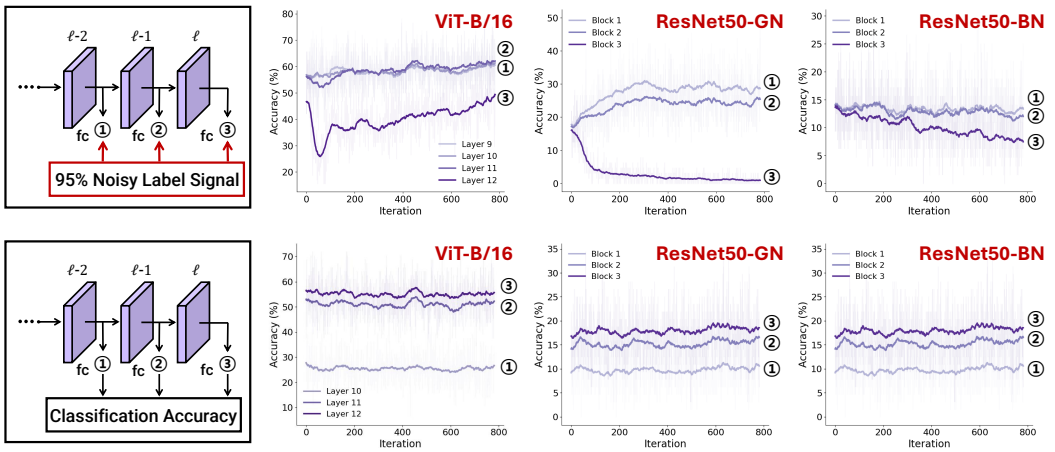


Figure 3: Noise sensitivity and initial performance of intermediate representations under corruption shift. (**Above**) We compare the final prediction performance obtained by optimizing the outputs of the final layer ℓ and intermediate layers ℓ-1, ℓ-2 using cross-entropy loss with 95% incorrect labels and 5% correct labels. (**Below**) We compare classification performance by applying a classifier trained on the final feature to lower layers without additional training. These results indicate that intermediate representations are less sensitive to noise while exhibiting progressively decreasing accuracy across diverse architectures.

3.2 Analysis of Intermediate Representations in TTA

Figure 3 shows the accumulated accuracy of ViT-Base (Dosovitskiy et al., 2021), ResNet50-GN (Wu & He, 2018), and ResNet50-BN (He et al., 2016) under Gaussian noise from ImageNet-C (Hendrycks & Dietterich, 2019), illustrating performance changes over the online test stream. To analyze the performance and adaptation characteristics of intermediate representations in the initial model without additional training modules, we compare the outputs of earlier layers that share the same dimension as the final feature. The feature extractor with ℓ layers is defined as $F_t^{\ell}(x_t) = f_t^{\ell} \circ f_t^{\ell-1} \circ \cdots \circ f_t^{1}(x_t)$, where intermediate representations refer to partial compositions of earlier layers (i.e., $F_t^{\ell-1}(x_t) = f_t^{\ell-1} \circ f_t^{\ell-2} \circ \cdots \circ f_t^{1}(x_t)$). The logits at layer $\ell - 1$ are defined as $z_t^{\ell-1}(x_t) = g_t(F_t^{\ell-1}(x_t))$, and we compare predictions across layers.

Intermediate representations are more robust. Unsupervised adaptation often suffers from confirmation bias, as incorrect predictions are reinforced (Arazo et al., 2020). Self-training on incorrect predictions can lead to model collapse, as observed in methods such as entropy minimization and pseudo-labeling (Niu et al., 2022; Han et al., 2025). The robustness of intermediate representations to incorrect predictions is evaluated under a setting with 95% noisy labels and 5% true labels (Figure 3, above). Remarkably, even with only 5% true labels, adaptation based on intermediate representations leads to improved or maintained performance across ViT-Base, ResNet50-GN, and ResNet50-BN, while performance at the final output degrades. This finding is consistent with prior empirical observations on noisy label training (Bai et al., 2021) and suggests the benefits of using intermediate representations for adaptation.

Deeper representations are more accurate. We now turn to the question of what evidence or learning signals can support adaptation via intermediate representations. Representations from deeper layers capture more fine-grained patterns and achieve better classification performance due to improved class discriminability (Zeiler & Fergus, 2014). The relationship between layer depth and performance is preserved under corruption shifts. We present the observed results in Figure 3 (below). This observation motivates the use of deeper-layer representations as a source of supervision for earlier layers. Under the observed conditions, transferring signals from deeper to shallower layers improves the lower bound of the final layer accuracy, thereby enabling progressive performance improvement.

4 INTRINSIC SELF-DISTILLATION

Motivated by the above analysis, we propose ISD, which achieves progressive performance improvement by using outputs from deeper layers as learning signals for earlier layers. Knowledge distillation transfers knowledge from a large teacher model to a small student model; it is widely used for knowledge transfer and model compression (Hinton et al., 2015; Gupta et al., 2016; Rebuffi et al., 2017; Gu et al., 2024).

Self-distillation applies knowledge distillation within a single architecture (Zhang et al., 2019; Xu & Liu, 2019). ISD leverages intrinsic representations within a single model and shares similarities with self-distillation, integrating both feature-level and logit-level distillation. As discussed in Section 2, efficiency, adaptability, and stability are primary objectives in this work. Gradual feature distillation (GFD) is designed to improve efficiency and adaptability, while debiased logit distillation (DLD) aims to enhance stability. Figure 4 illustrates the overall framework of ISD, with detailed descriptions provided in Sections 4.1 and 4.2.

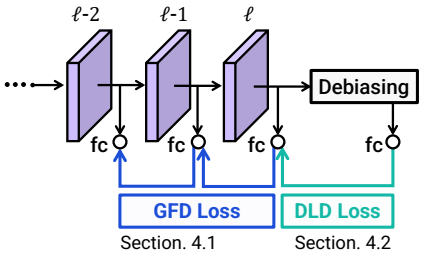


Figure 4: Intrinsic self-distillation.

4.1 GRADUAL FEATURE DISTILLATION

For the final feature, each class logit can be decomposed into magnitude and angular components with respect to the corresponding classifier weight vector. Specifically, we define the classifier weight matrix as $W = [w_1, w_2, \ldots, w_C]$, where w_j denotes the classifier weight vector for class j and C is the number of classes. The logit vector is given by $z_t^\ell = [z_1^\ell, z_2^\ell, \ldots, z_C^\ell]^\top$, where each class logit is formulated as follows:

$$z_j^\ell = w_j^\top F_t^\ell(x_t) \; = \|w_j\|_2, \|F_t^\ell(x_t)\|_2 \cos\left(w_j, F_t^\ell(x_t)\right), \quad j = 1, \ldots, C,$$

where $\cos(\cdot, \cdot)$ denotes cosine similarity and $\|\cdot\|_2$ denotes the L_2-norm. For simplicity, the bias term is omitted. The formulation can be extended to intermediate layers, yielding $z_t^{\ell-k}$ for $k \geq 1$, for layers whose output features have dimension matching the final feature. Since intermediate representations are not aligned with the classifier, scale mismatch may induce bias during adaptation. For example, for a fixed cosine similarity, an increase in the output norm leads to sharper predictions, enabling a reduction in the loss without meaningful structural changes. To mitigate this issue, we replace the scale of intermediate outputs with that of the final output, thereby preventing bias during training. Therefore, the adjusted logit $\bar{z}$ is formalized as follows:

$$\bar{z}_j^{\ell-k} = \|w_j\|_2 \, \|F_t^\ell(x_t)\|_2^{\text{stop-grad}} \, \cos\left(w_j, F_t^{\ell-k}(x_t)\right), \tag{1}$$

where $\|F_t^\ell(x_t)\|_2^{\text{stop-grad}}$ denotes the L_2-norm of the final feature with stop-gradient operation. In implementation, we update only the affine layers of the feature extractor while keeping the classifier weights W fixed. As a result, logits of adjusted intermediate representations are computed using a shared classifier.

Building on this, we leverage outputs from deeper layers as learning signals for earlier layers, thereby exploiting the high accuracy of final representations and the robustness of intermediate representations. Formally, the proposed GFD loss is defined as:

$$\mathcal{L}_{\text{GFD}} = \sum_{i=1}^{k} H\left(\sigma(\bar{z}_t^{\ell-i+1})^{\text{stop-grad}}, \sigma(\bar{z}_t^{\ell-i})\right), \tag{2}$$

where $H(p, q) = -p \log q$ and k denotes the number of preceding layers used for distillation. GFD gradually refines the outputs of lower layers, thereby indirectly improving the final prediction. Compared to entropy minimization, which often induces abrupt changes in final predictions, GFD mitigates overconfidence and improves calibration capability by providing structured learning signals.

4.2 DEBIASED LOGIT DISTILLATION

In TTA, improving predictions based on self-predictions may negatively impact performance under persistent errors, rendering models vulnerable to collapse into a single dominant prediction. While

Table 1: . Comparisons with state-of-the-art methods on ImageNet-C (severity level 5) with ResNet50-GN and ViT-Base regarding accuracy (%). We summarize the average accuracy (%) and ECE (%) across 15 domains, with the best performance highlighted in bold.

	Noise			Blur				Weather				Digital				Average	
Method	Gauss.	Shot	Impul.	Defoc.	Glass	Motion	Zoom	Snow	Frost	Fog	Brit.	Contr.	Elastic	Pixel	JPEG	Acc.	ECE
ResNet50-GN	17.98	19.84	17.88	19.76	11.35	21.41	24.92	40.43	47.30	33.60	69.28	36.27	18.62	28.39	52.29	30.62	8.37
Tent [ICLR'21]	5.03	6.60	5.38	15.22	9.57	22.71	22.66	26.53	34.41	3.13	70.37	42.02	11.61	48.21	54.35	25.19	48.76
EATA [ICML'22]	36.90	39.56	38.77	28.03	27.33	36.66	38.87	50.61	49.05	55.57	71.99	50.02	41.56	55.59	58.20	45.24	25.02
SAR [ICLR'23]	27.96	30.59	29.76	18.70	17.22	30.30	30.70	41.57	43.35	10.97	70.71	43.94	14.91	48.70	55.16	34.30	12.29
DeYO [ICLR'24]	39.04	42.23	40.61	22.35	23.83	38.40	37.68	50.87	49.55	1.04	73.08	49.92	41.16	56.08	57.81	41.58	13.53
ROID [WACV'24]	39.36	42.01	40.19	20.85	25.79	38.09	39.99	54.16	52.56	58.91	73.06	51.60	45.56	57.16	58.82	46.54	46.44
SPA [ICML'25]	41.23	44.20	41.91	28.84	28.01	41.60	44.69	55.45	53.53	60.63	73.56	51.31	50.34	61.88	58.36	49.04	16.83
ZeroSiam [ICLR'26]	40.52	42.83	41.87	31.98	32.30	40.60	44.67	53.57	53.30	59.27	72.90	52.53	47.74	58.98	59.23	48.82	21.06
ISD	**43.27**	**45.47**	**44.01**	**35.79**	**34.66**	**44.98**	**48.49**	**58.38**	**56.33**	**62.92**	**74.01**	**55.66**	**53.11**	**61.95**	**61.58**	**52.04**	**4.55**
ViT-Base	55.34	56.23	56.00	46.48	34.78	52.87	44.20	62.69	62.66	65.56	77.70	32.04	69.11	66.72	66.67	56.60	10.87
Tent [ICLR'21]	61.09	62.66	62.44	59.15	56.73	63.46	58.82	61.98	64.34	3.79	79.26	67.38	61.10	72.94	70.48	60.37	17.63
EATA [ICML'22]	62.08	63.82	63.16	60.59	61.02	66.07	63.35	70.62	68.60	71.83	79.99	66.52	68.17	75.03	72.88	67.58	12.75
SAR [ICLR'23]	59.66	61.26	60.96	56.75	55.44	61.46	56.91	65.86	63.38	66.38	78.82	58.02	62.21	72.15	69.77	63.27	**5.63**
DeYO [ICLR'24]	58.84	61.55	60.97	57.19	58.14	64.72	61.74	68.95	66.37	71.01	78.79	0.40	69.80	74.14	71.88	61.63	16.97
ROID [WACV'24]	62.80	64.39	64.09	59.79	61.14	65.32	63.15	71.79	70.38	73.94	80.59	63.17	71.74	75.82	73.80	68.13	68.03
SPA [ICML'25]	63.32	64.82	64.43	61.24	**63.79**	**69.25**	**67.37**	**74.20**	**72.68**	75.78	**80.90**	65.29	**74.00**	**77.43**	**74.74**	69.95	42.13
ZeroSiam [ICLR'26]	62.34	63.76	63.70	59.90	60.43	66.00	63.39	70.27	68.30	72.33	80.14	65.74	69.75	75.43	72.37	67.59	14.41
ISD	**63.46**	**65.31**	**64.86**	**62.98**	63.26	68.53	66.77	72.95	71.01	**76.18**	80.56	**70.77**	72.72	76.60	73.94	**69.99**	13.78

we partially mitigate this issue through GFD by avoiding explicit confidence maximization, the risk of model collapse may still persist. Debiasing methods (Niu et al., 2021; Wang et al., 2022b; Lee & Kim, 2024) based on causal inference approximate counterfactual outcomes either by using momentum predictions obtained via exponential moving average of model outputs to calibrate final predictions (Wang et al., 2022b), or by subtracting the model output for non-informative inputs from the final logits to remove bias (Niu et al., 2021; Lee & Kim, 2024). Inspired by these approaches, we address bias induced by non-causal factors in model collapse and propose the DLD loss as follows:

$$\mathcal{L}_{\text{DLD}} = H\Big(\sigma(g(F_t^\ell(x_t) - \lambda\Delta))^{\text{stop-grad}}, \sigma(\bar{z}_t^\ell)\Big), \tag{3}$$

$$\Delta \leftarrow m\Delta + (1-m)F_t^\ell(x_t), \tag{4}$$

where Δ represents the feature-level bias estimated via an exponential moving average update with momentum parameter m, and λ is a hyperparameter. The overall objective is composed of two distillation losses:

$$\mathcal{L}_{\text{ISD}} = \mathcal{L}_{\text{GFD}} + \alpha\mathcal{L}_{\text{DLD}}, \tag{5}$$

where α is a hyperparameter. ISD is a unified self-distillation framework that combines feature-level self-distillation for efficiency and adaptability with logit-level self-distillation for stability.

5 Experiments

5.1 Experiments Settings

Datasets. We evaluate our method under domain shifts using ImageNet-C (Hendrycks & Dietterich, 2019), an algorithmically generated corruption dataset. ImageNet-C is a widely used standard benchmark for TTA and consists of 15 corruption types, for which we use severity level 5 in all experiments. In addition, we provide evaluations on datasets representing natural domain shifts in the Appendix A.4, including ImageNet (Deng et al., 2009) variants such as ImageNet-R (Hendrycks et al., 2021a), ImageNet-A (Hendrycks et al., 2021b), ImageNet-V2 (Recht et al., 2019), and ImageNet-Sketch (Wang et al., 2019), as well as Office-Home (Venkateswara et al., 2017) and DomainNet (Peng et al., 2019).

Evaluation protocols. We report performance under three TTA protocols: standard TTA, continual TTA, and realistic TTA. In the standard TTA protocol, adaptation and evaluation are performed online on test samples from a single target domain, and the accumulated average performance is reported. The continual TTA protocol operates over sequential target domains without resetting model parameters, and the realistic TTA protocol extends standard TTA to a practical setting with class-imbalanced test samples and a batch size = 1 suggested by Niu et al. (Niu et al., 2023).

Table 2: . Comparisons with state-of-the-art methods under Continual TTA protocols on ImageNet-C (severity level 5) with ViT-Base regarding accuracy (%). We summarize the average accuracy (%) and ECE (%) across 15 domains, with the best performance highlighted in bold.

Time	$t \longrightarrow$																	
	Noise			Blur				Weather				Digital				Average	
Method	Gauss.	Shot	Impul.	Defoc.	Glass	Motion	Zoom	Snow	Frost	Fog	Brit.	Contr.	Elastic	Pixel	JPEG	Acc.	ECE
Source	55.34	56.23	56.00	46.48	34.78	52.87	44.20	62.69	62.66	65.56	77.70	32.04	69.11	66.72	66.67	56.60	10.87
Tent [ICLR'21]	61.09	64.46	64.57	54.53	55.60	60.99	56.05	61.44	58.24	68.29	78.19	64.59	3.45	0.40	0.39	50.15	37.23
EATA [ICML'22]	62.08	65.44	65.22	59.13	60.79	65.18	64.21	69.44	69.38	72.65	80.12	49.63	68.16	74.52	73.32	66.62	21.85
SAR [ICLR'23]	59.66	61.64	61.36	53.53	55.15	58.78	55.34	61.20	62.40	64.27	76.92	58.39	57.76	68.63	68.30	61.55	14.52
DeYO [ICLR'24]	58.84	60.97	59.78	47.55	50.26	52.32	15.87	56.62	59.52	65.33	76.34	50.30	62.50	69.63	67.39	56.88	28.14
ROID [WACV'24]	62.80	64.45	64.10	59.32	61.70	65.84	63.64	**71.69**	**70.98**	74.23	**80.69**	63.35	71.79	75.77	**73.84**	68.28	68.18
SPA [ICML'25]	63.32	65.71	65.10	55.84	**63.17**	**67.16**	66.59	71.52	70.78	75.10	79.38	68.72	67.70	**76.51**	73.29	68.66	56.85
ZeroSiam [ICLR'26]	62.34	64.48	63.17	47.17	51.54	54.99	53.40	62.99	62.19	65.56	77.13	50.96	55.66	69.92	68.57	60.67	22.71
ISD	**63.46**	**65.97**	**65.46**	**60.11**	62.42	67.15	**67.19**	70.84	70.70	**75.71**	79.93	**70.07**	**71.86**	75.45	73.18	**69.30**	**4.95**

Table 3: . Comparisons with state-of-the-art methods on ImageNet-C (severity level 5) under label shifts (imbalance ratio =∞) scenario regarding accuracy (%). We summarize the average accuracy (%) and ECE (%) across 15 domains, with the best performance highlighted in bold.

	Noise			Blur				Weather				Digital				Average	
Method	Gauss.	Shot	Impul.	Defoc.	Glass	Motion	Zoom	Snow	Frost	Fog	Brit.	Contr.	Elastic	Pixel	JPEG	Acc.	ECE
ResNet50-GN	17.88	19.93	17.91	19.72	11.25	21.31	24.88	40.41	47.39	33.60	69.25	36.26	18.70	28.37	52.22	30.61	8.37
Tent [ICLR'21]	2.58	3.25	2.73	13.93	7.93	19.52	16.84	16.52	22.05	1.82	70.47	42.19	6.61	49.36	53.66	21.96	60.59
EATA [ICML'22]	23.17	28.51	27.14	16.61	19.07	24.37	27.75	31.40	35.52	41.48	63.63	36.44	26.12	39.13	46.60	32.46	44.91
SAR [ICLR'23]	33.54	36.01	36.08	19.20	18.61	33.80	22.69	23.79	45.84	48.75	71.87	46.60	8.10	51.61	56.34	36.86	15.24
DeYO [ICLR'24]	41.62	43.65	42.75	23.04	24.69	40.53	4.18	52.32	50.83	1.36	73.16	52.33	45.83	58.78	58.73	40.92	19.99
COME [ICLR'25]	17.73	19.69	17.74	19.65	11.20	21.22	24.83	40.22	47.80	33.69	68.89	35.62	18.55	27.66	51.83	30.42	8.81
SPA [ICML'25]	41.73	44.38	42.28	30.60	31.90	45.35	48.44	54.85	52.62	62.65	72.99	53.24	54.09	63.07	58.52	50.45	**5.68**
ZeroSiam [ICLR'26]	43.17	44.80	43.30	34.88	35.99	44.26	0.85	58.12	54.89	62.95	72.41	54.66	52.69	61.77	60.12	48.32	34.87
ISD	**43.41**	**45.52**	**44.06**	**36.39**	**37.19**	**47.12**	**50.12**	**59.12**	**56.21**	**64.41**	**73.75**	**56.61**	**56.22**	**63.74**	**61.93**	**53.05**	6.87
ViT-Base	28.89	29.65	31.59	31.25	25.23	39.41	31.51	24.51	29.97	54.39	64.50	48.42	34.20	52.54	55.05	38.74	**8.46**
Tent [ICLR'21]	53.73	54.03	54.60	55.31	52.16	59.22	46.70	6.43	17.75	70.08	75.56	68.65	57.63	69.09	66.18	53.81	25.18
EATA [ICML'22]	44.37	44.88	46.98	40.20	44.41	49.14	45.13	51.73	50.47	58.57	71.84	24.28	60.74	63.85	61.93	50.57	27.48
SAR [ICLR'23]	52.05	52.99	53.00	52.01	52.29	57.51	53.00	21.73	43.61	68.67	74.49	66.90	57.88	67.15	64.42	55.85	14.58
DeYO [ICLR'24]	52.41	52.81	53.41	54.26	54.70	62.15	49.96	64.90	63.41	71.06	76.91	63.88	66.06	71.61	67.93	61.70	17.26
COME [ICLR'25]	57.36	58.35	58.12	59.72	**61.40**	66.06	23.03	70.38	68.91	75.32	**79.39**	70.20	70.92	74.99	**72.21**	64.42	23.92
SPA [ICML'25]	54.03	55.00	54.52	58.50	60.49	**66.95**	**65.25**	69.83	68.44	**75.83**	79.43	67.94	71.04	**75.17**	69.96	66.16	22.80
ZeroSiam [ICLR'26]	57.51	58.07	**58.74**	58.88	59.86	65.27	63.51	69.84	**69.21**	74.80	79.55	66.86	69.93	75.15	71.92	66.61	16.65
ISD	**57.64**	**58.46**	58.72	**60.56**	61.15	66.70	**65.25**	**70.59**	68.86	75.39	79.26	**70.67**	**71.09**	74.84	71.30	**67.37**	10.23
ConvNeXt-Tiny	21.38	27.68	22.13	24.51	10.97	32.47	31.16	44.50	52.45	39.54	72.03	44.82	23.10	17.72	55.78	34.68	10.30
Tent [ICLR'21]	18.06	23.76	26.61	24.34	4.24	33.87	29.19	41.56	46.27	39.72	73.17	51.59	18.41	10.66	56.90	33.22	29.27
EATA [ICML'22]	34.60	36.88	34.95	26.56	18.98	36.88	33.40	47.63	48.33	50.52	72.96	50.74	28.55	31.25	56.89	40.61	19.09
SAR [ICLR'23]	34.84	36.73	34.91	27.20	4.14	34.97	30.69	44.34	47.56	8.47	73.07	51.10	19.42	24.16	56.66	35.22	23.87
DeYO [ICLR'24]	36.93	39.86	37.49	28.39	1.92	36.56	9.35	52.89	45.49	1.66	74.38	56.82	4.20	18.33	58.44	33.51	41.93
COME [ICLR'25]	42.54	44.43	42.34	**38.73**	1.18	46.07	11.85	57.14	56.88	61.20	75.61	60.63	5.08	**52.03**	**61.87**	43.84	24.04
SPA [ICML'25]	29.48	37.08	30.59	18.80	21.05	42.82	40.66	57.41	35.36	54.49	73.89	48.20	33.99	50.57	61.27	42.38	13.21
ZeroSiam [ICLR'26]	42.27	44.52	42.07	37.36	**33.49**	45.34	44.00	56.48	55.40	62.87	74.75	59.41	44.25	51.64	60.02	50.26	18.59
ISD	**44.35**	**46.31**	**44.43**	38.62	33.10	**46.69**	**46.05**	**59.50**	**59.49**	**65.28**	**75.97**	**60.64**	**46.49**	50.47	61.27	**51.91**	**4.47**

Comparison methods. We compare our results with state-of-the-art TTA approaches, including Tent (Wang et al., 2021), EATA (Niu et al., 2022), SAR (Niu et al., 2023), DeYO (Lee et al., 2024), ROID (Marsden et al., 2024), COME (Zhang et al., 2025), SPA (Niu et al., 2025), ZeroSiam (Chen et al., 2025). Detailed descriptions of each method are provided in the Appendix A.1. We report the average online accuracy for each target domain and summarize the overall performance using average accuracy and expected calibration error (ECE) (Naeini et al., 2015) across all domains. We compute the ECE by partitioning predictions into 15 confidence bins. The discrepancy between accuracy and confidence is measured within each bin and averaged over all bins. Lower ECE indicates better calibration capability.

Implementation details. All experiments are conducted on a single NVIDIA RTX A6000 GPU. For source models, we use ImageNet pretrained models from the timm (Wightman, 2019) library, including ResNet50-GN, ViT-Base, and ConvNeXt-Tiny. Our implementation is built upon the official open-source codebase of (Niu et al., 2025; Chen et al., 2025), with a batch size of 64. Following prior work (Niu et al., 2025; Chen et al., 2025), we employ the SGD optimizer with a momentum of 0.9. For a batch size of 64, the learning rates are set to 0.01, 0.0025, and 0.00025 for ResNet50-GN, ViT-Base, and ConvNeXt-Tiny, respectively. For batch size 1 experiments, the learning rates are scaled down by a factor of 32. During adaptation, we update the affine parameters only. For practical applicability, we adopt a fixed hyperparameters across all experiments, with α=1, λ=1, and m=0.95. More details are provided in the Appendix A.2 and the code is included in the supplementary material.

Table 4: . Comparisons with state-of-the-art methods on ImageNet-C (severity level 5) under batch size = 1 scenario regarding accuracy (%). We summarize the average accuracy (%) and ECE (%) across 15 domains, with the best performance highlighted in bold.

Method	Noise			Blur				Weather				Digital				Average	
	Gauss.	Shot	Impul.	Defoc.	Glass	Motion	Zoom	Snow	Frost	Fog	Brit.	Contr.	Elastic	Pixel	JPEG	Acc.	ECE
ResNet50-GN	17.99	19.84	17.88	19.74	11.35	21.42	24.92	40.43	47.30	33.60	69.27	36.26	18.62	28.39	52.28	30.62	8.37
Tent [ICLR'21]	2.50	2.89	2.51	13.47	3.61	18.43	17.62	15.37	22.99	1.41	70.45	42.25	6.18	49.18	53.85	21.51	61.97
EATA [ICML'22]	24.87	28.22	25.76	18.08	17.26	28.51	29.24	44.52	44.36	41.57	70.91	44.67	26.91	46.78	55.73	36.49	16.66
SAR [ICLR'23]	23.77	26.31	23.92	18.52	15.38	28.30	30.56	44.62	44.81	23.80	72.43	44.67	12.08	46.98	56.11	34.14	10.60
DeYO [ICLR'24]	41.79	44.57	42.77	22.86	24.61	41.63	10.05	53.62	51.90	1.90	73.35	53.64	47.99	60.48	59.74	42.06	18.30
COME [ICLR'25]	17.87	19.68	17.74	19.71	11.30	21.34	24.87	40.34	47.50	33.61	68.93	35.76	18.54	27.88	51.84	30.46	8.73
SPA [ICML'25]	41.35	45.17	42.80	30.15	32.75	46.77	48.32	56.14	53.48	62.76	**73.77**	52.66	53.28	63.89	57.11	50.69	**6.31**
ZeroSiam [ICLR'26]	41.52	43.09	42.44	34.19	35.54	45.28	49.64	58.44	55.84	64.19	73.73	55.13	54.79	63.12	61.32	51.88	21.11
ISD	**45.64**	**47.40**	**46.30**	**39.15**	**39.03**	**50.69**	**53.07**	**60.57**	**57.32**	**65.92**	73.75	**58.10**	**59.43**	**65.63**	**62.64**	**55.00**	8.87
ViT-Base	29.10	29.62	31.58	31.20	25.10	39.33	31.46	24.60	30.18	54.34	64.47	48.45	34.18	52.52	55.11	38.75	8.49
Tent [ICLR'21]	51.91	52.27	53.23	53.12	47.73	56.72	47.07	8.69	22.45	67.41	70.01	67.39	51.08	66.59	64.25	52.26	22.99
EATA [ICML'22]	46.12	46.85	48.83	45.24	39.76	49.86	43.44	36.92	42.14	64.83	68.32	64.57	50.33	60.67	60.79	51.24	14.18
SAR [ICLR'23]	50.16	50.54	51.78	51.53	48.93	56.70	50.73	31.92	54.38	67.32	73.46	65.70	55.06	66.43	64.93	55.97	11.01
DeYO [ICLR'24]	53.34	54.68	54.03	55.44	57.08	62.80	59.22	65.21	63.58	72.02	77.17	67.19	66.48	72.06	69.22	63.30	14.41
COME [ICLR'25]	57.51	58.50	58.62	60.47	60.25	65.70	61.58	69.68	68.01	75.18	79.43	71.06	69.58	74.78	72.05	66.83	20.44
SPA [ICML'25]	53.25	55.19	53.32	60.08	61.80	**68.51**	66.79	70.86	69.41	**76.66**	**79.75**	53.93	71.63	75.56	68.10	65.66	**3.47**
ZeroSiam [ICLR'26]	56.51	57.31	57.27	58.50	59.65	64.09	62.13	67.73	67.61	73.84	78.64	68.80	38.50	73.43	69.94	65.60	17.07
ISD	**58.98**	**60.06**	**60.23**	**62.37**	**62.68**	68.13	**67.02**	**71.70**	**70.28**	76.27	79.70	**72.18**	**72.21**	**75.69**	**72.26**	**68.65**	10.97
ConvNeXt-Tiny	21.46	27.75	22.17	24.56	10.89	32.52	30.64	44.43	52.50	39.81	72.03	44.66	23.21	17.57	55.67	34.68	10.31
Tent [ICLR'21]	30.13	33.22	32.11	24.90	7.34	33.59	29.89	43.96	47.23	42.86	72.92	49.68	21.38	18.26	56.35	36.25	17.55
EATA [ICML'22]	26.60	30.45	26.89	25.11	11.40	33.13	30.91	44.77	50.10	39.97	72.19	47.74	23.57	18.56	55.82	35.81	10.31
SAR [ICLR'23]	30.07	33.80	29.98	25.90	11.09	33.44	30.71	44.82	47.62	38.21	72.99	46.15	22.61	18.12	56.14	36.11	11.42
DeYO [ICLR'24]	36.64	39.48	36.81	28.63	3.40	38.02	16.13	51.68	47.59	4.95	73.96	55.79	7.55	25.08	57.73	34.90	33.99
COME [ICLR'25]	41.07	43.58	40.98	37.03	2.17	44.82	18.41	56.32	55.76	60.77	75.40	59.42	8.05	44.09	60.99	43.26	21.37
SPA [ICML'25]	36.29	38.99	28.66	24.41	23.21	44.00	41.48	57.53	54.64	55.08	74.23	50.31	37.66	**52.70**	60.02	45.95	8.72
ZeroSiam [ICLR'26]	41.48	44.02	41.48	36.95	32.17	44.64	42.94	55.59	54.39	61.82	74.38	58.99	42.34	51.17	59.31	49.44	18.91
ISD	**44.18**	**46.57**	**44.62**	**38.98**	**32.92**	**47.02**	**45.67**	**59.77**	**59.99**	**65.65**	**76.01**	**61.01**	**46.61**	51.07	**61.55**	**52.11**	**4.56**

5.2 Experimental Results

Standard TTA protocol.

Table 1 shows the average accuracy and ECE across 15 domains on ImageNet-C. Our ISD achieves the best accuracy and calibration on ResNet50-GN and yields competitive results on ViT-Base. Notably, ISD consistently outperforms ZeroSiam, which requires only a single forward pass, while achieving competitive accuracy compared with the multi-branch-based SPA across both architectures.

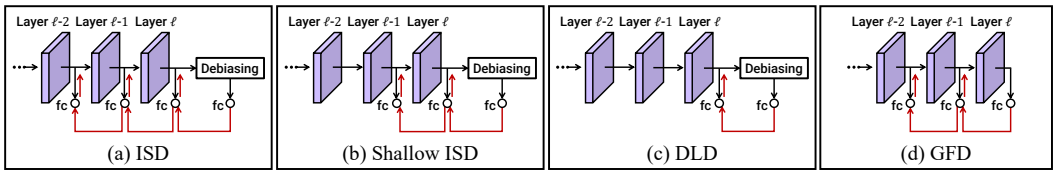
Continual TTA protocol. Table 2 reports the results of continual TTA over a sequence of 15 corruption domains on ImageNet-C. ISD achieves the best average accuracy and the lowest ECE, demonstrating strong performance across diverse domain shifts. In contrast to prior methods such as Tent, which suffer from model collapse in later domains, ISD maintains stable performance throughout the sequence. Compared to ZeroSiam, ISD shows improved adaptability to continuously changing domains, leading to consistently improved accuracy across the sequence. Overall, these results highlight the effectiveness of ISD in providing robust, adaptable, and well-calibrated predictions under continual adaptation settings.

Realistic TTA protocol. Tables 3 and 4 present results on practical TTA scenarios, including the imbalanced label shifts and batch size = 1 settings. Our method consistently improves accuracy on ResNet50-GN and ViT-Base and achieves improvements in both accuracy and ECE on ConvNeXt-Tiny. Prior methods achieve low ECE in specific cases, but their calibration is inconsistent across architectures and scenarios. By contrast, ISD consistently improves calibration and provides more reliable uncertainty estimation. These results demonstrate the robustness of our method in realistic settings, where label imbalance and limited batch statistics present challenges for stable adaptation.

5.3 Further Analysis

Effect of each component. Figure 5 illustrates the ablation study results to evaluate the effect of each component. With only GFD applied, performance is comparable to ISD when model collapse does not occur; however, collapse leads to degradation. In contrast, combining DLD and GFD yields consistent performance gains without model collapse. Furthermore, ISD with a progressive two-layer demonstrates the best performance, outperforming the shallow ISD that applies GFD to a single layer.

Hyperparameter sensitivity. We conduct a sensitivity analysis of α, λ, and m on the ImageNet-C using ResNet50-GN and ViT-Base to evaluate robustness with respect to the hyperparameters. We



Protocol	Standard-TTA		Continual-TTA	Realistic-TTA (label shifts)		
Model	RN50-GN	ViT-Base	ViT-Base	RN50-GN	ViT-Base	ConvNeXt-Tiny
(a) ISD	**52.04**	**69.99**	**69.30**	**53.05**	**67.37**	**51.91**
(b) Shallow ISD	51.61	69.68	69.14	52.90	67.23	49.19
(c) DLD	47.71	65.10	63.28	49.42	59.08	43.35
(d) GFD	47.64	60.43	36.89	46.34	67.23	35.54

Figure 5: Effect of each component. We present ablation results on ISD across diverse TTA scenarios regarding average accuracy (%). Results in red indicate model collapse.

Table 5: Hyperparameter sensitivity analysis. We study the sensitivity of α, λ, and m on ImageNet-C regarding average accuracy (%). RN50-GN stands for ResNet50-GN. Gray corresponds to only GFD, and purple denotes the hyperparameters used in the paper.

(a) Hyperparameter α.

α	RN50-GN	ViT-Base
0	47.64	60.43
0.5	51.34	69.97
1.0	52.04	69.99
1.5	52.49	69.99
2.0	**52.75**	**70.02**

(b) Hyperparameter λ.

λ	RN50-GN	ViT-Base
0	47.64	60.43
0.1	48.43	69.64
0.3	51.11	69.93
0.5	51.55	69.96
1.0	**52.04**	**69.99**

(c) Hyperparameter m.

m	RN50-GN	ViT-Base
0.9	**52.14**	**69.99**
0.95	52.04	**69.99**
0.99	51.56	69.94
0.995	51.23	69.95
1.0	47.64	60.43

report the average accuracy in Table 5. The coefficient α for the loss function of GFD and DLD shows stable performance over the range [0.5, 1.0, 1.5, 2.0]. The debiasing parameter λ and the momentum m for feature-level bias also exhibit stable behavior across the ranges of [0.1, 0.3, 0.5, 1.0] and [0.9, 0.95, 0.99, 0.995], respectively. Hyperparameters are not tuned separately for each dataset or model. Instead, we set α and λ to 1 and empirically set m to 0.95 to ensure robustness across diverse scenarios.

Efficiency comparison. Table 6 reports the efficiency of the proposed method compared to existing approaches. We measure the average inference time per batch on the Gaussian noise corruption of ImageNet-C using a single NVIDIA RTX A6000 GPU. Tent, Zerosiam, and ISD achieve comparably low latency due to their single forward-pass design. Despite this efficiency, ISD attains competitive performance compared to SPA, which requires three forward passes. These results demonstrate the effectiveness of the proposed ISD in achieving both efficiency and adaptability.

Table 6: Efficiency comparison.

Method	Latency (s/batch)	Acc. (%)
Tent	**0.329**	60.37
EATA	0.339	67.58
SAR	0.654	63.27
DeYO	0.470	61.63
SPA	0.854	69.95
ZeroSiam	0.335	67.59
ISD	0.332	**69.99**

6 Conclusion and Limitation

In this paper, we introduce ISD, an unified self-distillation framework for TTA. We demonstrate structured properties regarding robustness and accuracy for intermediate representations and unveil their value as an intrinsic learning signal. Our ISD provides a balanced solution regarding efficiency, adaptability, and stability via synergistic feature-level and logit-level self-distillation. Experimental evaluations show ISD not only outperforms existing state-of-the-art methods across various TTA scenarios and architectures but also achieves improved calibration performance.

Limitation. Our method can be applied to most modern DNNs; however, ISD requires access to intermediate representations, which limits the applicability in black-box settings where intermediate features are inaccessible and in architectures without compatible intermediate features.

AI USE STATEMENT

In this work, We used ChatGPT for minor english editing and language polishing of the manuscript. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

The source code is provided in the Supplementary Material. All implementation details are provided in Section 5.1 and Appendix A.2 to ensure reproducibility.

REFERENCES

Eric Arazo, Diego Ortego, Paul Albert, Noel E O'Connor, and Kevin McGuinness. Pseudo-labeling and confirmation bias in deep semi-supervised learning. In *2020 International joint conference on neural networks (IJCNN)*, pp. 1–8. IEEE, 2020.

Yingbin Bai, Erkun Yang, Bo Han, Yanhua Yang, Jiatong Li, Yinian Mao, Gang Niu, and Tongliang Liu. Understanding and improving early stopping for learning with noisy labels. *Advances in Neural Information Processing Systems*, 34:24392–24403, 2021.

Shai Ben-David, John Blitzer, Koby Crammer, and Fernando Pereira. Analysis of representations for domain adaptation. In B. Schölkopf, J. Platt, and T. Hoffman (eds.), *Advances in Neural Information Processing Systems*, volume 19. MIT Press, 2006.

Shai Ben-David, John Blitzer, Koby Crammer, Alex Kulesza, Fernando Pereira, and Jennifer Wortman Vaughan. A theory of learning from different domains. *Machine learning*, 79(1):151–175, 2010.

John Blitzer, Mark Dredze, and Fernando Pereira. Biographies, bollywood, boom-boxes and blenders: Domain adaptation for sentiment classification. In *Proceedings of the 45th annual meeting of the association of computational linguistics*, pp. 440–447, 2007.

Malik Boudiaf, Romain Mueller, Ismail Ben Ayed, and Luca Bertinetto. Parameter-free online test-time adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8344–8353, 2022.

Dian Chen, Dequan Wang, Trevor Darrell, and Sayna Ebrahimi. Contrastive test-time adaptation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 295–305, 2022.

Guohao Chen, Shuaicheng Niu, Deyu Chen, Jiahao Yang, Zitian Zhang, Mingkui Tan, Pengcheng Wu, and Zhiqi Shen. Zerosiam: An efficient siamese for test-time entropy optimization without collapse. *arXiv preprint arXiv:2509.23183*, 2025.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pp. 1597–1607. PmLR, 2020.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2009.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021.

Lixin Duan, Ivor W Tsang, Dong Xu, and Stephen J Maybank. Domain transfer svm for video concept detection. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1375–1381. IEEE, 2009.

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.

Boqing Gong, Yuan Shi, Fei Sha, and Kristen Grauman. Geodesic flow kernel for unsupervised domain adaptation. In *2012 IEEE conference on computer vision and pattern recognition*, pp. 2066–2073. IEEE, 2012.

Sachin Goyal, Mingjie Sun, Aditi Raghunathan, and J Zico Kolter. Test time adaptation via conjugate pseudo-labels. *Advances in Neural Information Processing Systems*, 35:6204–6218, 2022.

Yves Grandvalet and Yoshua Bengio. Semi-supervised learning by entropy minimization. *Advances in neural information processing systems*, 17, 2004.

Yuxian Gu, Li Dong, Furu Wei, and Minlie Huang. MiniLLM: Knowledge distillation of large language models. In *The Twelfth International Conference on Learning Representations*, 2024.

Saurabh Gupta, Judy Hoffman, and Jitendra Malik. Cross modal distillation for supervision transfer. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2827–2836, 2016.

Jisu Han, Jaemin Na, and Wonjun Hwang. Ranked entropy minimization for continual test-time adaptation. In *Forty-second International Conference on Machine Learning*, 2025.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. In *International Conference on Learning Representations*, 2019.

Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, Dawn Song, Jacob Steinhardt, and Justin Gilmer. The many faces of robustness: A critical analysis of out-of-distribution generalization. *ICCV*, 2021a.

Dan Hendrycks, Kevin Zhao, Steven Basart, Jacob Steinhardt, and Dawn Song. Natural adversarial examples. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021b.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

Yihao Hu, Congyu Qiao, Xin Geng, and Ning Xu. Selective label enhancement learning for test-time adaptation. In *The Thirteenth International Conference on Learning Representations*, 2025a.

Zixuan Hu, Yichun Hu, Xiaotong Li, Shixiang Tang, and Ling-Yu Duan. Beyond entropy: Region confidence proxy for wild test-time adaptation. *arXiv preprint arXiv:2505.20704*, 2025b.

Samuli Laine and Timo Aila. Temporal ensembling for semi-supervised learning. In *International Conference on Learning Representations*, 2017.

Hunter Lang, David Sontag, and Aravindan Vijayaraghavan. Theoretical analysis of weak-to-strong generalization. *Advances in neural information processing systems*, 37:46837–46880, 2024.

Chen-Yu Lee, Saining Xie, Patrick Gallagher, Zhengyou Zhang, and Zhuowen Tu. Deeply-supervised nets. In *Artificial intelligence and statistics*, pp. 562–570. Pmlr, 2015.

Dong-Hyun Lee et al. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In *Workshop on challenges in representation learning, ICML*, volume 3, pp. 896. Atlanta, 2013.

Hyuck Lee and Heeyoung Kim. Cdmad: Class-distribution-mismatch-aware debiasing for class-imbalanced semi-supervised learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 23891–23900, 2024.

Jonghyun Lee, Dahuin Jung, Saehyung Lee, Junsung Park, Juhyeon Shin, Uiwon Hwang, and Sungroh Yoon. Entropy is not enough for test-time adaptation: From the perspective of disentangled factors. *arXiv preprint arXiv:2403.07366*, 2024.

Jiaming Liu, Senqiao Yang, Peidong Jia, Renrui Zhang, Ming Lu, Yandong Guo, Wei Xue, and Shanghang Zhang. ViDA: Homeostatic visual domain adapter for continual test time adaptation. In *The Twelfth International Conference on Learning Representations*, 2024.

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

Muhammad Jehanzeb Mirza, Pol Jané Soneira, Wei Lin, Mateusz Kozinski, Horst Possegger, and Horst Bischof. Actmad: Activation matching to align distributions for test-time-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 24152–24161, 2023.

Yu Mitsuzumi, Akisato Kimura, and Hisashi Kashima. Understanding and improving source-free domain adaptation from a theoretical perspective. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 28515–28524, 2024.

Mahdi Pakdaman Naeini, Gregory Cooper, and Milos Hauskrecht. Obtaining well calibrated probabilities using bayesian binning. In *Proceedings of the AAAI conference on artificial intelligence*, 2015.

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

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

Shuaicheng Niu, Chunyan Miao, Guohao Chen, Pengcheng Wu, and Peilin Zhao. Test-time model adaptation with only forward passes. In *Forty-first International Conference on Machine Learning*, 2024. URL https://openreview.net/forum?id=qz1Vx1v9iK.

Shuaicheng Niu, Guohao Chen, Peilin Zhao, Tianyi Wang, Pengcheng Wu, and Zhiqi Shen. Self-bootstrapping for versatile test-time adaptation. In *Forty-second International Conference on Machine Learning*, 2025.

Yulei Niu, Kaihua Tang, Hanwang Zhang, Zhiwu Lu, Xian-Sheng Hua, and Ji-Rong Wen. Counterfactual vqa: A cause-effect look at language bias. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 12700–12710, 2021.

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

Sylvestre-Alvise Rebuffi, Alexander Kolesnikov, Georg Sperl, and Christoph H Lampert. icarl: Incremental classifier and representation learning. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pp. 2001–2010, 2017.

Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do imagenet classifiers generalize to imagenet? In *International conference on machine learning*. PMLR, 2019.

Jin Shin and Hyun Kim. L-TTA: Lightweight test-time adaptation using a versatile stem layer. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

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

Kihyuk Sohn, David Berthelot, Nicholas Carlini, Zizhao Zhang, Han Zhang, Colin A Raffel, Ekin Dogus Cubuk, Alexey Kurakin, and Chun-Liang Li. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. *Advances in neural information processing systems*, 33:596–608, 2020.

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

Yu Sun, Xiaolong Wang, Zhuang Liu, John Miller, Alexei Efros, and Moritz Hardt. Test-time training with self-supervision for generalization under distribution shifts. In *International conference on machine learning*, pp. 9229–9248. PMLR, 2020.

Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1–9, 2015.

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

Hemanth Venkateswara, Jose Eusebio, Shayok Chakraborty, and Sethuraman Panchanathan. Deep hashing network for unsupervised domain adaptation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 5018–5027, 2017.

Dequan Wang, Evan Shelhamer, Shaoteng Liu, Bruno Olshausen, and Trevor Darrell. Tent: Fully test-time adaptation by entropy minimization. In *International Conference on Learning Representations*, 2021.

Haohan Wang, Songwei Ge, Zachary Lipton, and Eric P Xing. Learning robust global representations by penalizing local predictive power. In *Advances in Neural Information Processing Systems*, 2019.

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, 2022a.

Xudong Wang, Zhirong Wu, Long Lian, and Stella X Yu. Debiased learning from naturally imbalanced pseudo-labels. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 14647–14657, 2022b.

Zhenyi Wang, Enneng Yang, Li Shen, and Heng Huang. A comprehensive survey of forgetting in deep learning beyond continual learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(3):1464–1483, 2024.

Colin Wei, Kendrick Shen, Yining Chen, and Tengyu Ma. Theoretical analysis of self-training with deep networks on unlabeled data. In *International Conference on Learning Representations*, 2021.

Ross Wightman. Pytorch image models. `https://github.com/rwightman/pytorch-image-models`, 2019.

Yuxin Wu and Kaiming He. Group normalization. In *Proceedings of the European conference on computer vision (ECCV)*, pp. 3–19, 2018.

MA Xiao, Young D. Kwon, Pan Zhou, and Dong Ma. Architecture-agnostic test-time adaptation via backprop-free embedding alignment. In *The Fourteenth International Conference on Learning Representations*, 2026.

Ting-Bing Xu and Cheng-Lin Liu. Data-distortion guided self-distillation for deep neural networks. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pp. 5565–5572, 2019.

Matthew D Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In *European conference on computer vision*, pp. 818–833. Springer, 2014.

Linfeng Zhang, Jiebo Song, Anni Gao, Jingwei Chen, Chenglong Bao, and Kaisheng Ma. Be your own teacher: Improve the performance of convolutional neural networks via self distillation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 3713–3722, 2019.

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.

Qingyang Zhang, Yatao Bian, Xinke Kong, Peilin Zhao, and Changqing Zhang. COME: Test-time adaption by conservatively minimizing entropy. In *The Thirteenth International Conference on Learning Representations*, 2025.

Bowen Zheng, Da-Wei Zhou, Han-Jia Ye, and De-Chuan Zhan. Multi-layer rehearsal feature augmentation for class-incremental learning. In *Forty-first International Conference on Machine Learning*, 2024.

A Appendix

A.1 Introduction to Comparison Methods

Tent (Wang et al., 2021): Tent introduces the TTA setting and proposes entropy minimization. By updating only the affine layers in an online manner for each batch, this approach enables efficient adaptation using a single model and a single forward pass; however, entropy minimization involves the risk of model collapse.

EATA (Niu et al., 2022): EATA proposes an active sample selection criterion to mitigate model collapse caused by entropy minimization by identifying reliable and non-redundant samples. Specifically, entropy minimization loss is applied only to low-entropy samples, and a Fisher information-based weight regularization term is introduced to alleviate the forgetting problem.

SAR (Niu et al., 2023): SAR extends TTA to more realistic settings by introducing the dynamic wild world scenario. Furthermore, to improve the stability of entropy minimization under diverse settings, SAR adopts a reliability-based sample selection strategy and employs the sharpness-aware minimization optimizer, which requires additional gradient computations for each update.

DeYO (Lee et al., 2024): DeYO extends entropy-based sample selection by introducing an additional criterion based on prediction consistency under input perturbations. Prediction distribution shifts induced by object-destructive transformations are utilized as a thresholding strategy to identify samples beneficial for adaptation.

ROID (Marsden et al., 2024): ROID proposes a universal TTA framework designed to address diverse and non-stationary environments. The framework improves adaptation stability through diversity-aware optimization, prior correction, and source-model weight ensembling.

COME (Zhang et al., 2025): COME introduces a conservative entropy minimization objective by modeling prediction uncertainty with a Dirichlet prior. This formulation regularizes the model to avoid overconfident predictions on unreliable samples and improves stability on TTA. The experimental results reported for COME correspond to the results of DeYO with COME.

PALSE (Hu et al., 2025a): PALSE proposes a selective label enhancement framework for TTA that models uncertainty using candidate pseudo-label sets instead of fixed one-hot labels. By progressively refining pseudo-labels during adaptation, PALSE alleviates error accumulation and improves robustness under domain shift.

SPA (Niu et al., 2025): SPA suggests generating deteriorated views in the frequency domain to construct stable signals during adaptation. The method applies low-frequency amplitude masking to remove dominant semantic components while preserving spatial structure, injects high-frequency noise to strengthen perturbations without destroying geometric consistency. Hence, SPA involves three forward passes, including the original image and two augmented views.

ZeroSiam (Chen et al., 2025): ZeroSiam mitigates the model collapse of entropy minimization-based approach via asymmetric Siamese architecture. ZeroSiam employs a single layer predictor to align online predictions with stop-gradient target predictions generated from shared feature representations, enabling stable adaptation without augmentations or auxiliary teacher models.

A.2 Additional Experimental Details

Experiments are conducted following the base protocol of SPA[1], with ZeroSiam[2] used for realistic TTA and PALSE[3] for domain generalization datasets. In the realistic TTA setting, data normalization is not aligned with that used by timm pretrained models, following prior work (Niu et al., 2023; Lee et al., 2024; Chen et al., 2025). This results in a performance gap on the initial source domain compared to the standard TTA protocol. For domain generalization datasets, we select the learning rate for each domain via grid search over the range from 1e-3 to 1e-6, and use a batch size of 128, following PALSE (Hu et al., 2025a). For experiments on ImageNet-C and ImageNet variants, we use learning rates of 0.00025, 0.001, and 0.000025 for Tent, EATA, SAR, DeYO, ROID, and COME on

[1] https://github.com/mr-eggplant/SPA

[2] https://github.com/Cascol-Chen/ZeroSiam

[3] https://github.com/palm-ml/PASLE

ResNet50-GN, ViT-Base, and ConvNeXt-Tiny, respectively. For SPA, we use learning rates of 0.0025, 0.01, and 0.00025 for the feature extractor, while 0.025, 0.05, and 0.025 are used for the predictor on ResNet50-GN, ViT-Base, and ConvNeXt-Tiny, respectively. For ZeroSiam, we use learning rates of 0.0025, 0.005, and 0.00025 for the feature extractor, while 0.025, 0.005, and 0.0025 are used for the predictor on ResNet50-GN, ViT-Base, and ConvNeXt-Tiny, respectively. All hyperparameter settings follow the original papers and (Niu et al., 2025; Chen et al., 2025).

A.3 Pseudo Code

Algorithm 1 illustrates the overall procedure of the proposed ISD framework. For each incoming test sample, intermediate and final features are extracted from the feature extractor, and adjusted logits are computed using a shared scaling factor derived from the final feature norm. The model is optimized using the proposed GFD and DLD losses. Finally, the feature-level bias Δ is updated via an exponential moving average. Since the final feature logit is identical to the model output, predictions are obtained without requiring an additional forward pass.

Algorithm 1 ISD: Intrinsic Self-Distillation

Input: test image $x_t \in \mathcal{X}_t$, feature extractor $f_t(\cdot)$, classifier $g_t(\cdot)$, zero initialized Δ
Compute features $F_t^l(x_t)$, $F_t^{\ell-1}(x_t)$, $F_t^{\ell-2}(x_t)$ via feature extractor $f(\cdot)$
Compute logit scaling factor $\|F_t^\ell(x_t)\|_2^{\text{stop-grad}}$
Compute debiased logit $z_\Delta = g(F_t^\ell(x_t) - \lambda\Delta)$
Compute losses $\mathcal{L}_{\text{GFD}}$ by Eq. 2 and $\mathcal{L}_{\text{DLD}}$ by Eq. 3
Compute total loss $\mathcal{L}_{\text{ISD}}$ via Eq. 5
Update affine parameters of $f(\cdot)$ using gradient $\nabla\mathcal{L}_{\text{ISD}}$
Update $\Delta \leftarrow m\Delta + (1-m)F_t^\ell(x_t)$
Output: prediction y_t from logit $z_t = g_t(F_t^l(x_t))$

A.4 Experimental Results on Natural Domain Shift Scenarios

We present results on natural domain shifts in Tables 7 and 8. On ImageNet variants, we achieve the best accuracy on ResNet50-GN, while demonstrating improved calibration capability on ViT-Base. We evaluate our method in a domain generalization setting by applying TTA methods to models trained with empirical risk minimization (ERM) on the other domains. Our method achieves the best performance on Office-Home and the best performance on DomainNet when applied with the pseudo labeling method PALSE. These results indicate that our method is broadly applicable across different models and datasets.

Table 7: Comparisons with state-of-the-art methods on ImageNet variants regarding accuracy (%). We summarize the average accuracy (%) and ECE (%) across 15 domains, with the best performance highlighted in bold.

Model	ResNet50-GN				Average		ViT-Base				Average	
Method	R	A	V2	Sketch	Acc.	ECE	R	A	V2	Sketch	Acc.	ECE
Source	40.80	14.01	68.90	29.17	38.22	13.80	59.49	50.84	**75.49**	44.89	57.68	6.30
Tent [ICLR'21]	41.18	14.35	68.90	26.56	37.75	27.41	63.92	52.77	75.14	49.01	60.21	11.56
SAR [ICLR'23]	42.49	14.41	68.93	30.70	39.13	**13.62**	63.94	52.69	75.01	48.65	60.07	7.44
DeYO [ICLR'24]	48.19	16.79	68.98	34.72	42.17	14.75	68.83	**55.23**	74.37	50.58	62.25	12.20
ROID [WACV'24]	47.29	15.69	68.62	37.70	42.33	42.02	**69.44**	54.60	74.61	53.21	62.97	62.65
SPA [ICML'25]	47.14	**18.28**	**69.07**	37.17	42.92	14.74	68.32	55.12	75.22	**53.45**	**63.03**	27.20
ZeroSiam [ICLR'26]	46.06	16.65	68.70	36.31	41.93	24.68	67.04	54.19	74.94	51.08	61.81	14.74
ISD	**50.05**	16.64	69.02	**40.70**	**44.10**	13.94	68.41	53.33	74.33	52.41	62.12	**5.43**

A.5 Experimental Results on Mixed Domain Shift

We evaluate ISD under a mixed domain shift protocol, where samples from the 15 corruption types in ImageNet-C are randomly mixed to simulate abrupt domain changes. Unlike the Continual-TTA protocol, in which each corruption domain is presented sequentially, this setting introduces frequent domain shifts within the test stream, providing a more challenging scenario for EMA-based bias

Table 8: Comparisons with state-of-the-art methods on domain generalization datasets with ResNet50-BN regarding accuracy (%). We summarize the average accuracy (%) across all domains, with the best performance highlighted in bold.

Dataset	Office-Home					DomainNet						
Method	Art	Clipart	Product	Real	Avg.	Clipart	Infograph	Painting	Quickdraw	Real	Sketch	Avg.
ERM	61.76	54.07	76.50	77.85	67.55	64.98	15.26	50.45	11.85	63.71	54.00	43.38
Tent [ICLR'21]	63.49	56.72	76.82	77.25	68.57	66.30	18.05	52.90	15.05	66.03	56.29	45.77
DeYO [ICLR'24]	63.45	57.64	77.87	77.87	69.21	**66.60**	18.82	**53.78**	15.66	66.39	**56.83**	46.35
COME [ICLR'25]	62.18	53.40	76.01	77.58	67.29	64.97	15.17	50.44	14.76	63.67	53.95	43.83
PALSE [ICLR'25]	64.85	56.40	78.46	77.58	69.32	66.53	21.54	53.55	18.41	**67.48**	56.82	47.39
ZeroSiam [ICLR'26]	63.08	56.79	76.66	**77.99**	68.63	65.88	19.22	52.35	17.29	65.58	55.61	45.99
ISD	**65.55**	**58.81**	77.77	77.81	**69.99**	65.86	18.89	52.73	15.41	65.38	56.19	45.74
ISD w/ PALSE	65.06	56.88	**78.53**	77.76	69.56	66.16	**22.44**	53.43	**20.65**	66.19	56.54	**47.57**

estimation. As shown in Table 9, the accuracy of ISD decreases from 69.30% under Continual-TTA to 64.28% under the mixed domain shift protocol. Nevertheless, ISD maintains competitive accuracy, achieving the second-highest performance among the compared methods. Notably, ISD achieves the lowest ECE of 4.31%, substantially outperforming the other methods in terms of calibration. These results demonstrate that ISD remains effective even when the test distribution changes abruptly within the data stream.

Table 9: Comparison results on the ImageNet-C under mixed domain shift scenario with ViT-Base.

Method	Tent	EATA	SAR	DeYO	COME	SPA	ZeroSiam	ISD
Acc. (%)	58.82	57.91	53.61	52.29	62.95	**64.85**	59.30	64.28
ECE (%)	24.55	26.55	20.82	31.55	24.99	21.87	23.54	**4.31**

A.6 Discussion with Thresholding-based Methods

We adopt an adaptation strategy that achieves stable adaptation while utilizing all test samples. Existing prediction-based methods typically select samples considered beneficial for adaptation by applying various thresholding criteria (Niu et al., 2022; 2023; Lee et al., 2024; Hu et al., 2025b). Specifically, ReCAP (Hu et al., 2025b) reduces prediction disparity by optimizing region confidence instead of instance-wise entropy, thereby promoting consistent adaptation across local semantic regions through probabilistic region modeling and an efficient proxy objective. Nevertheless, similar to DeYO (Lee et al., 2024), ReCAP employs a threshold-based sample selection strategy, which requires architecture-specific threshold tuning. Moreover, ReCAP relies on source data to compute the covariance matrix. While such sample selection strategies can improve adaptation stability by updating the model only with samples deemed beneficial for learning, their effectiveness may depend heavily on the data distribution and the source model. In particular, a selection criterion calibrated for one setting may become overly restrictive when applied to a different architecture or distribution, potentially leaving too few samples for effective adaptation.

In contrast, our ISD requires no access to source data and achieves stable adaptation while utilizing all test samples. To examine whether the aforementioned limitation arises in practice, we additionally conduct experiments using the lightweight MobileNetV3-Small-0.5 backbone without modifying any hyperparameters, with the results reported in Table 10. Under Gaussian Noise at severity level 5 on ImageNet-C, DeYO's thresholding strategy filters out 49,982 of the 50,000 test samples, leaving only 18 samples for adaptation. This result demonstrates that a fixed thresholding strategy can become excessively restrictive for certain architectures and may therefore require careful architecture-specific tuning. In contrast, the proposed ISD achieves stable adaptation even on a lightweight network without any additional modifications, by leveraging its intrinsic learning signal and utilizing all available test samples.

Table 10: Comparison results on ImageNet-C with MobileNetV3-Small-0.5.

Method	Source	SAR	DeYO	ReCAP ($\tau_{RE} = 1.0$)	ReCAP ($\tau_{RE} = 3.0$)	ISD
Acc. (%)	22.28	20.78	12.16	22.28	6.95	**24.21**

A.7 Discussion with Source Statistics-based Methods

Existing TTA methods have leveraged intermediate representations to mitigate domain shift (Mirza et al., 2023; Niu et al., 2024). In particular, ActMAD (Mirza et al., 2023) extends adaptation beyond the final layer by aligning the activation statistics of multiple intermediate layers. It stores the means and variances of source-domain representations and minimizes their discrepancy with the corresponding statistics of target batches, enabling fine-grained distribution alignment. However, such approaches rely on precomputed source-domain statistics, which may not always be available in practical deployment scenarios. Table 11 shows the comparison results on ImageNet-C with existing methods that utilize source-domain information. EATA uses source data to estimate the Fisher Information Matrix, while ActMAD relies on stored activation statistics for distribution alignment. In contrast, ISD achieves the best accuracy without requiring access to source-domain data.

Table 11: Comparison results on ImageNet-C with ViT-Base.

Method	EATA	ActMAD	ISD
Acc. (%)	67.58	66.34	**69.99**

A.8 Statistical Comparison

We evaluate the robustness of our results to data ordering by conducting five trials with random seeds [2021, 2022, 2023, 2024, 2025] on both ImageNet-C and DomainNet, and report the mean accuracy and standard deviation in Tables 12 and 13, respectively. As shown in Table 12, all methods exhibit relatively small standard deviations on ImageNet-C, except in cases where model collapse occurs. In particular, ZeroSiam shows large variations under Continual-TTA and Realistic-TTA with RN50-GN, whereas ISD consistently maintains low standard deviations across different architectures and protocols. These results demonstrate that ISD provides stable adaptation regardless of data ordering. Table 13 further evaluates the robustness on DomainNet. Both ISD and PALSE exhibit small standard deviations across all target domains. Moreover, combining ISD with PALSE consistently improves the mean accuracy over PALSE across all domains while maintaining low standard deviations. These results demonstrate that the performance gains obtained by integrating ISD with PALSE are consistent across different data orderings.

Table 12: Comparison results on ImageNet-C over 5 random seeds.

Protocol	Standard-TTA		Continual-TTA	Realistic-TTA (label shifts)	
Model	RN50-GN	ViT-Base	ViT-Base	RN50-GN	ViT-Base
SPA	49.08 ± 0.08	69.94 ± 0.07	68.57 ± 0.11	50.43 ± 0.36	66.32 ± 0.15
ZeroSiam	48.70 ± 0.07	67.56 ± 0.19	53.98 ± 8.53	47.53 ± 6.49	66.69 ± 0.20
ISD	**52.01** ± **0.02**	**69.98** ± **0.03**	**69.35** ± **0.03**	**53.18** ± **0.15**	**67.70** ± **0.25**

Table 13: omparison results on DomainNet over 5 random seeds. † denotes the use of a balancing coefficient, i.e., $\mathcal{L}_{\text{PALSE}} + 0.1\mathcal{L}_{\text{ISD}}$.

Acc (%)	Clipart	Infograph	Painting	Quickdraw	Real	Sketch	Avg.
PALSE	66.57 ± 0.10	21.53 ± 0.10	53.68 ± 0.07	18.44 ± 0.14	67.22 ± 0.21	56.86 ± 0.10	49.05 ± 0.06
ISD	65.97 ± 0.17	18.78 ± 0.13	52.69 ± 0.08	15.41 ± 0.08	65.44 ± 0.08	56.03 ± 0.04	47.39 ± 0.04
ISD w/ PALSE	65.99 ± 0.12	**22.38** ± **0.03**	53.57 ± 0.10	**20.62** ± **0.17**	66.49 ± 0.34	56.44 ± 0.07	49.28 ± 0.12
ISD w/ PALSE †	**66.78** ± **0.13**	21.89 ± 0.11	**54.09** ± **0.07**	18.99 ± 0.18	**67.71** ± **0.06**	**57.08** ± **0.09**	**49.47** ± **0.07**

A.9 Analysis of Intermediate Feature Accuracy

We further analyze how ISD affects the prediction quality of intermediate features. Specifically, we evaluate the classification accuracy obtained from the intermediate layers on ImageNet-C under Gaussian Noise, as reported in Table 14. ISD consistently improves the accuracy across all intermediate layers, with particularly substantial gains in earlier layers. These results demonstrate that ISD

improves the discriminative capability of intermediate representations throughout the network, rather than merely enhancing the prediction of the final layer.

Table 14: Layer wise accuracy on ImageNet-C Gaussian Noise with ViT-Base.

Acc. (%)	Layer 9	Layer 10	Layer 11	Layer 12
Source	6.81	25.55	51.23	55.09
ISD	17.34	43.03	61.36	63.46

A.10 Analysis of Calibration Performance

We investigate the effect of each component of ISD on calibration using ECE. Table 15 compares the calibration performance of the ISD with its component-wise variants across different architectures and TTA protocols. Among the individual components, GFD generally exhibits low ECE across different settings, supporting its effectiveness in alleviating overconfidence. In contrast, DLD shows relatively large variations in ECE depending on the architecture and protocol, as its primary role is to prevent model collapse rather than to improve calibration. Interestingly, Shallow ISD also achieves low ECE in several settings, indicating that exploiting intermediate representations does not inherently lead to overconfident predictions. Combining GFD and DLD does not necessarily yield a lower ECE than GFD alone, since the two components serve different objectives. Nevertheless, the full ISD avoids the severe calibration degradation observed in several existing TTA methods and maintains relatively stable ECE across different architectures and protocols. These results empirically demonstrate the distinct roles of the two components: GFD contributes to calibration, while DLD complements GFD by improving adaptation stability.

Protocol	Standard-TTA		Realistic-TTA (label shifts)		
Model	RN50-GN	ViT-Base	RN50-GN	ViT-Base	ConvNeXt-Tiny
Source	5.30	9.47	5.39	4.82	5.65
Tent	5.03	21.77	88.88	19.15	49.78
SPA	11.08	37.16	4.87	12.37	9.57
ZeroSiam	23.76	17.05	26.50	21.91	21.12
DLD	2.26	24.33	**1.84**	17.02	**1.92**
GFD	3.48	**6.50**	3.58	**3.67**	2.25
Shallow ISD	**1.42**	16.16	4.57	11.71	2.16
ISD	2.05	10.75	5.98	7.12	2.75

Table 15: Analysis of the effect of each Component on ECE (%) under Gaussian noise in ImageNet-C

A.11 Applicability to Diverse DNN Architectures

ISD utilizes a shared classifier across intermediate representations and therefore requires the selected intermediate features to have the same dimensionality as the final representation. This design enables ISD to leverage intermediate representations without introducing additional classifiers or auxiliary modules. We evaluate ISD on MobileNetV3-Small-0.5 and MLP-Mixer-Base to further demonstrate its applicability across diverse DNN architectures. Table 16 shows that ISD improves upon the source model and outperforms ZeroSiam on both architectures. Together with the results on ResNet50-GN, ViT-Base, and ConvNeXt-Tiny in the main paper, these results demonstrate that ISD can be applied across diverse architectural families, including CNNs, Transformers, and MLP-based models, provided that suitable intermediate representations with matching feature dimensionality are available.

Table 16: Comparison results on ImageNet-C with diverse DNN architectures.

Acc. (%)	MobileNetV3-Small-0.5	MLP-Mixer-Base
Source	22.28	26.71
ZeroSiam	21.21	35.01
ISD	**24.21**	**47.85**