

Distill Once, Adapt Life-Long: Exploring Dataset Distillation for Continual Test-Time Adaptation

Hyun-Kurl Jang¹^{*}, Jihun Kim¹^{*}, Hyeokjun Kwon²^{*}, and
Kuk-Jin Yoon¹

¹ KAIST, Visual Intelligence Lab
{jhg0001, jihun1998, kjyoon}@kaist.ac.kr

² Chung-Ang University, FoVLab
hyeokunkwon@cau.ac.kr

Abstract. Continual Test-Time Adaptation (CTTA) aims to maintain model performance under evolving target domains by adapting online without labeled data. However, practical deployments often cannot retain the source dataset due to privacy or licensing constraints, and purely source-free CTTA methods tend to become unstable under long-term distribution shift, suffering from compounding self-training errors and catastrophic forgetting. We introduce **DO-ALL** (Distill Once, Adapt Life-Long), a plug-and-play framework that revisits source information in a compact and privacy-conscious form via Dataset Distillation (DD). Before deployment, DO-ALL performs DD to produce a small set of synthetic distilled anchors that summarize the source distribution. During adaptation, each target sample is matched with its most semantically aligned anchor, which provides a stable reference for various CTTA via source replay, representation alignment, and manifold-smoothing regularization. DO-ALL can be seamlessly integrated into existing CTTA algorithms, consistently improving long-term robustness across CIFAR100-C, ImageNet-C, and the CCC benchmark. This demonstrates the potential of leveraging DD to enable stable and continuous adaptation without retaining raw source data. The code is available at <https://github.com/blue-531/DOALL>.

Keywords: Dataset Distillation · Continual Test-Time Adaptation

1 Introduction

Perception systems rarely operate in a static environment. After deployment, external factors such as weather can vary, sensors may degrade, and operational contexts often shift. Test-Time Adaptation (TTA) [34, 36, 37, 57] addresses this by updating a source-trained model with unlabeled target data encountered during inference. Further, Continual Test-Time Adaptation (CTTA) [3, 13, 28, 41, 43, 46, 47, 53, 59, 67, 69, 81] considers a more dynamic scenario in which the model must

^{*} Equal contribution.

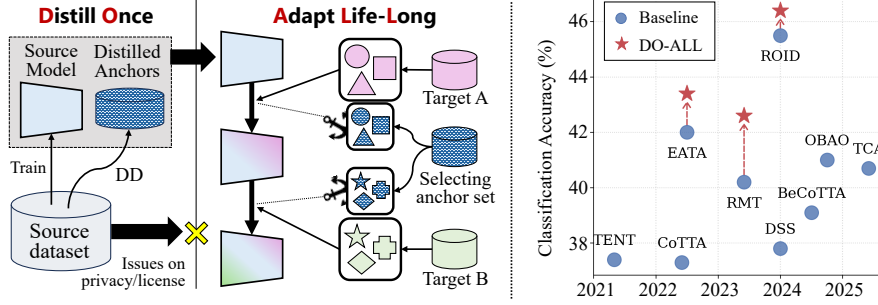


Fig. 1: Left: Overview of our **DO-ALL** framework. Before deployment, we distill the source dataset into a compact set of synthetic anchors. After deployment, each target sample is matched to the anchor that best aligns with its representation. The selected anchors serve as stable reference points that regularize model updates and prevent drift throughout CTTA. **Right:** Chronological comparison of accuracy on ImageNet-C for existing CTTA methods (blue circles) and DO-ALL-augmented variants (red stars).

adapt online to a sequence of evolving domains while maintaining prediction quality.

Since distributing or retaining the full source set is often infeasible (*e.g.*, due to privacy/license issues), various CTTA works adopt a source-free setting. While this direction has clear practical advantages, fully discarding source information fundamentally limits stability under long-term, non-stationary target streams. As distributions drift, self-training signals become increasingly unreliable, and repeated updates exacerbate catastrophic forgetting. In such cases, relying solely on the test stream makes long-term stabilization inherently challenging, as the model lacks a persistent anchor to keep its representation grounded [7, 13].

Several works [1, 2, 13, 22, 23, 55] have explored using source information in ways that avoid distributing the raw dataset. For example, they have proposed to retain only source statistics [22, 37], feature prototypes [7, 13], or lightweight proxy representations [1, 13, 23]. These strategies aim to provide a stable reference that prevents the model from drifting excessively during adaptation, while still respecting the practical constraints of the source-free setting. In a similar spirit, we revisit source information in a compact and privacy-conscious manner.

Our **DO-ALL** (Distill Once, Adapt Life-Long) framework is shown in Fig. 1. It performs Dataset Distillation (DD) on the source dataset once before deployment, summarizing the source domain into a tiny set of synthetic samples. DD is known to compress a large training set into a small number of learnable synthetic examples that preserve class-discriminative structure and training behavior, while requiring orders of magnitude less storage and exposing far less raw data content [61, 75]. Thus, the distilled samples provide a compact and persistent representation of the source distribution that can be carried into deployment.

A key advantage of DO-ALL is that it is designed to integrate this distilled source knowledge into existing CTTA algorithms without altering their objectives or architectures. Therefore, the distilled samples function as an **optional**

stability buffer that can be plugged into a broad range of CTTA methods to counteract drift and forgetting, rather than replacing their adaptation strategy.

To achieve this, we introduce two core components in DO-ALL. First, before deployment, we construct source-distilled anchors via DD (left of Fig. 1). Each anchor consists of (1) a synthetic sample, (2) its soft label under the source model, and (3) its latent feature in the source feature space. These anchors provide compact but semantically meaningful reference points for the model to preserve the source representation, even after deployment.

Second, during CTTA, we associate each target sample with its most semantically related distilled anchor (right of Fig. 1). This correspondence then guides adaptation through source replay, MixUp-style regularization [72], and feature-space alignment that preserves local neighborhood structure. To further prevent long-term drift, we introduce a harm-adaptive blending mechanism that selectively restores unstable parameter groups toward the source initialization based on their gradient magnitudes.

Together, these enable **DO-ALL** to deliver stable and robust adaptation, while preserving **plug-and-play** compatibility with existing CTTA methods. We experimentally verify that DO-ALL consistently improves a wide range of CTTA methods when paired with various DD strategies, yielding notable gains across CIFAR100-C [19], ImageNet-C [19], and the CCC benchmark [48].

Beyond the main results, we conduct extensive ablations on the proposed components, anchor association, and cross-architecture settings. Interestingly, we observe a clear positive correlation between the quality of the distilled anchors and the magnitude of CTTA improvement in DO-ALL: as DD methods produce more informative synthetic samples, DO-ALL can translate this semantic richness directly into stronger stabilization during adaptation. We hope this work encourages further exploration at the intersection of DD and CTTA toward more robust deployment-time learning.

2 Related work

2.1 Dataset Distillation (DD)

Scaling model capacity has intensified the demand for massive datasets, which in turn raises storage and compute costs. DD [8, 21, 35, 44, 75, 80] tackles this by synthesizing a compact set of images whose training effect approximates that of the full dataset, but early approaches [61, 75] incurred substantial computation and slow convergence. Later methods improved scalability and stability via trajectory matching [5, 15–17, 29, 30, 39, 52, 66, 77, 79], aligning optimization paths between real and synthetic training to reduce parameter-update mismatch and accelerate convergence. Distribution matching approaches [10, 12, 32, 33, 40, 49, 51, 58, 60, 70, 71, 76, 78] minimize cross-layer feature discrepancies to improve generalization. These advances have expanded DD to large-scale settings [9, 40, 42, 50, 51, 56, 68, 74] with stronger memory/time profiles. Beyond compression, DD has been adopted in continual learning [65] and privacy-sensitive

regimes [14, 54, 64] where raw data cannot be shared. Our approach leverages a distilled source set as a compact, privacy-aware anchor within continual test-time adaptation, using it as an explicit anti-forgetting prior that stabilizes adaptation over evolving target distributions.

2.2 Continual Test-Time Adaptation (CTTA)

Source-trained models typically degrade under distribution shift at test time. TTA [20, 24–26, 34, 36, 37, 57] addresses this by updating the trained model on unlabeled target samples without accessing source data. Common strategies include entropy minimization [57], consistency across stochastic augmentations [73], and normalization re-estimation to align activation statistics to the target data [34, 37]. In practice, however, deployments rarely face a static target; the test distribution often drifts over time. CTTA [3, 13, 28, 41, 43, 46, 47, 53, 59, 67, 69, 81] formalizes this setting, aiming to adapt over long horizons while avoiding error accumulation and forgetting. CoTTA [59] employs a student–teacher framework and stochastic parameter restoration to prevent error accumulation over long streams. RMT [13] reinforces teacher consistency with contrastive objectives, keeping features close to the source space. EcoTTA [53] focuses on low-memory updates and self-distilled regularization to keep adaptation lightweight and stable. CTTA methods primarily adapt online from unlabeled target streams, but some explicitly carry over source information using cached prototypes, source proxy or feature statistics [1, 7, 13, 22, 23, 37]. These priors are prepared before test time and then combined with standard TTA objectives to improve stability. We take this idea a step further by using a distilled source dataset as the carried-over structure: it preserves class-discriminative information in a compact form, enabling regularization when needed.

3 Method

In this paper, we explore the potential of DD within the context of CTTA. Following the conventional formulation of CTTA [43, 45, 69], let $D_s \sim P_s$ denote the source dataset and target inputs are drawn from time-varying distributions P_t . Here, no target labels are available at test time, and the stream is observed only once. The goal of CTTA is to adapt a source-trained model f_{θ_0} , parameterized by θ_0 , to the non-stationary target stream.

A major challenge in CTTA is to adapt to continuously shifting target distributions while preserving the knowledge acquired from the source domain, *i.e.*, avoiding catastrophic forgetting [48, 59]. The key idea behind our DO-ALL is to leverage dataset distillation to compactly preserve essential source information, allowing it to serve as a stable **anchor** during adaptation. We illustrate the overall framework in Fig. 2, describing (1) what information is distilled and stored from the source domain, and (2) how the stored distilled knowledge is effectively utilized to regularize adaptation throughout the evolving target stream.

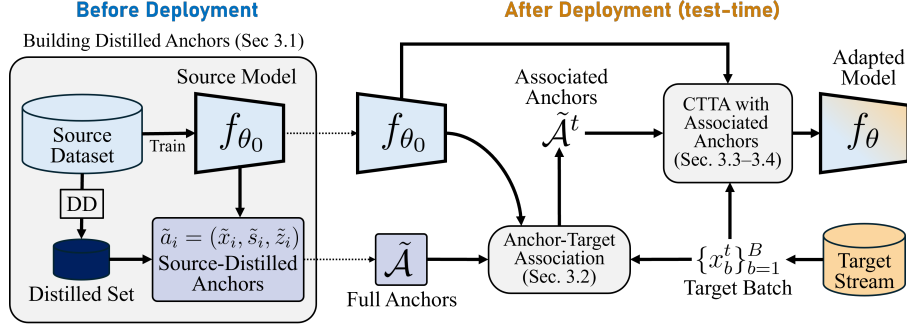


Fig. 2: Before deployment, we perform DD once on the source dataset to obtain a compact set of source-distilled anchors $\tilde{\mathcal{A}}$, where each anchor stores a synthetic input, its soft label under the source model, and its latent feature representation (Sec. 3.1). After deployment, during CTTA, each incoming target sample is matched to its most semantically aligned anchor in feature space, forming the associated anchor set $\tilde{\mathcal{A}}^t$ for the current batch (Sec. 3.2). The model is then updated with anchor-guided objectives and stabilized via harm-adaptive blending (Sec. 3.3–3.4). Together, these components enable stable and robust CTTA with the distilled summary of the source dataset.

3.1 Building Source-Distilled Anchors through DD

Dataset distillation (DD) [5, 40, 51, 61, 68] compresses a large dataset into a small set of informative synthetic samples that preserve the training dynamics of the original data. We apply DD to the source dataset D_s once before deployment, obtaining a compact synthetic set $\tilde{D}_s = \{\tilde{x}_i\}_{i=1}^{N_a}$ with $N_a \ll |D_s|$. This provides a memory- and privacy-efficient summary of the source distribution during CTTA. Throughout the paper, we use the tilde notation (*e.g.*, $\tilde{x}$) to denote quantities derived from the distilled source set.

Each distilled sample $\tilde{x}_i$ is assigned a soft pseudo-label by source model:

$$\tilde{s}_i = f_{\theta_0}(\tilde{x}_i). \quad (1)$$

We use soft rather than hard labels because they preserve the inter-class relations between distilled samples and the class-wise confidence structure encoded by the source model, offering richer semantic guidance for more robust CTTA.

In addition to the soft label, we extract and store the latent representation of each synthetic sample. We decompose the source model into $f_{\theta_0} = h_{\theta_0} \circ g_{\theta_0}$, where g_{θ_0} is the feature extractor and h_{θ_0} the classifier head. The latent representation of each distilled sample is:

$$\tilde{z}_i = g_{\theta_0}(\tilde{x}_i). \quad (2)$$

These latent features $\{\tilde{z}_i\}$ capture the source feature-space geometry and serve as stable reference points that preserve representational structure during CTTA.

Finally, we package the synthetic input, its soft label, and its latent feature into a single unit, which we refer to as a source-distilled anchor:

$$\tilde{a}_i = (\tilde{x}_i, \tilde{s}_i, \tilde{z}_i), \quad (3)$$

and denote the full anchor set as $\tilde{\mathcal{A}} = \{\tilde{a}_i\}_{i=1}^{N_a}$.

This anchor set $\tilde{\mathcal{A}}$ is constructed once before deployment and carried with the model throughout CTTA, providing a persistent reference to the source distribution during adaptation. Compared to storing the full source dataset, $\tilde{\mathcal{A}}$ is dramatically more compact and privacy-preserving, yet remains sufficiently informative to serve as an effective guide.

3.2 Anchor-Target Association at Test Time

After constructing and storing the anchors, we deploy the model and perform continual adaptation on the incoming target stream. As the target distribution drifts gradually, each test-time batch typically reflects only a local region of the shift. Therefore, rather than referencing all anchors in $\tilde{\mathcal{A}}$ uniformly, we retrieve the most relevant source information for each target sample based on feature similarity, enabling context-aware and stable adaptation.

Formally, at test time t , a mini-batch of B target samples $\{x_b^t\}_{b=1}^B$ arrives. For each target sample x_b^t , we compute its latent representation $z_b^t = g_\theta(x_b^t)$ and compare it with the anchor features $\{\tilde{z}_i\}_{i=1}^{N_a}$ using cosine similarity:

$$\text{sim}(b, i) = \frac{z_b^t \cdot \tilde{z}_i}{\|z_b^t\|_2 \|\tilde{z}_i\|_2}. \quad (4)$$

We then assign to each target sample its nearest anchor:

$$j = \arg \max_i \text{sim}(b, i), \quad \tilde{a}_b^t = (\tilde{x}_j, \tilde{s}_j, \tilde{z}_j). \quad (5)$$

To put it simply, each target sample x_b^t is paired with its nearest anchor $\tilde{a}_b^t$. Collecting these assignments over the batch yields the associated anchor set $\tilde{\mathcal{A}}^t = \{\tilde{a}_b^t\}_{b=1}^B$ for CTTA at this timestep t . This provides each target sample with a semantically aligned source reference. By grounding adaptation updates in latent-space similarity, the model adapts within semantically consistent regions of the representation space, substantially improving stability under continual distribution shift.

3.3 Test-Time Adaptation with Associated Anchors

Given the associated anchor set $\tilde{\mathcal{A}}^t$ for the current test-time batch, we adapt the model using two complementary objectives. These respectively (1) preserve source knowledge while imposing local smoothness around the anchor–target manifold and (2) align the evolving representation with the source feature geometry. All objectives operate on the paired correspondences between target samples and their matched anchors.

(1) Anchor-based replay loss. To prevent catastrophic forgetting during continual adaptation, we replay source knowledge through the source-distilled anchors and constrain the model to remain consistent with them. Compared to relying on the full source dataset, distilled anchors require no access to D_s at

test time, making them suitable for deployment settings with memory or privacy constraints. Moreover, unlike randomly sampled real images or coreset-style subsets, each distilled synthetic sample is optimized to serve as a compact and information-dense proxy of the source distribution [8,31,35], capturing class-level structure more effectively than uniform subsampling or clustering-based selection. As a result, anchor replay provides stronger and more targeted guidance for stabilizing the model under distribution shift.

Our replay objective consists of two terms. The first term performs *direct replay* on anchors, preserving the source model’s predictive behavior. The second term extends replay to the *local target-anchor manifold* by mixing each target sample with its matched anchor, which encourages smoother predictions in the region where adaptation is actually taking place.

(i) *Direct replay*. For each matched anchor $\tilde{x}_b$ with stored soft label $\tilde{s}_b$, we enforce consistency via temperature-scaled KL:

$$\mathcal{L}_{\text{dir}} = \beta_s^2 \cdot \text{KL}(\text{softmax}(f_\theta(\tilde{x}_b)/\beta_s) \parallel \text{softmax}(\tilde{s}_b/\beta_s)), \quad (6)$$

where $\beta_s > 0$ is a scale factor. This term explicitly regularizes the model toward source semantics and prevents the decision boundary from drifting away from the source class structure during test-time adaptation.

(ii) *Mixup replay*. MixUp [72] and its variants are known to improve robustness by encouraging approximately linear behavior between samples, which effectively smooths decision boundaries. We tailor this idea to our anchor-based setting by mixing each target sample with its *matched* anchor, rather than mixing arbitrary pairs. This yields a semantically meaningful interpolation path that connects a shifted target input to its corresponding source reference, providing targeted regularization where adaptation is most required.

Concretely, given a target-anchor pair $(x_b^t, \tilde{x}_b)$, we form the mixed input

$$x_b^m = \lambda x_b^t + (1 - \lambda)\tilde{x}_b, \quad \lambda \sim \text{Beta}(\alpha, \alpha), \quad (7)$$

and the corresponding mixed soft target

$$p_b^m = \lambda \text{softmax}(f_\theta(x_b^t)) + (1 - \lambda) \text{softmax}(\tilde{s}_b). \quad (8)$$

We then enforce prediction consistency on the mixed sample:

$$\mathcal{L}_{\text{mix}} = \text{KL}(\text{softmax}(f_\theta(x_b^m)) \parallel p_b^m). \quad (9)$$

This manifold replay encourages the decision boundary to vary smoothly along the target-anchor direction, mitigating local overfitting and improving robustness under continuous shift. Finally, we combine the two components:

$$\mathcal{L}_{\text{replay}} = \mathcal{L}_{\text{dir}} + \mathcal{L}_{\text{mix}}. \quad (10)$$

To validate the importance of mixing each target sample with its matched anchor, Table 1 compares three mixing variants. While Target-Target and Anchor-Anchor mixing provide modest gains over the baseline, our Target-Anchor mixing yields the largest improvement, supporting the intuition that replay and

Table 1: Experiments on ROID [43] evaluate MixUp-style variants (**T-T**: Target-Target; **A-A**: Anchor-Anchor; **T-A**: Target-Anchor), reporting classification error (%).

Baseline	DO-ALL without MixUp	T-T	A-A	T-A
54.40	53.77	53.68	53.69	53.55

smoothing are most effective when applied along the semantically meaningful path connecting each target to its distilled source anchor.

(2) Feature-level alignment: While anchor soft labels preserve decision-level semantics, they do not explicitly constrain the feature-space geometry during adaptation. As the target stream evolves, the feature extractor g_θ may drift even if predictions remain consistent, leading to representation collapse or overfitting to target domains. To maintain the source manifold structure, we align anchor and target features in the latent space using a layer-wise MMD objective:

$$\mathcal{L}_{\text{mmd}} = \sum_{\ell \in L} \left\| \mathbb{E}_{x^t} [\phi(g_\theta^\ell(x^t))] - \mathbb{E}_{\tilde{x}} [\phi(g_\theta^\ell(\tilde{x}))] \right\|_{\mathcal{H}}^2, \quad (11)$$

where $\phi(\cdot)$ denotes the feature mapping to the reproducing kernel Hilbert space. This encourages the evolving representation to remain aligned with the anchor space and reduces representation drift across layers.

Total loss: To sum up, the total adaptation loss is:

$$\mathcal{L}_{\text{anchor}} = \mathcal{L}_{\text{replay}} + \lambda_{\text{mmd}} \mathcal{L}_{\text{mmd}}, \quad (12)$$

where λ_{mmd} balances alignment and smoothness. In short, $\mathcal{L}_{\text{replay}}$ preserves source competence, $\mathcal{L}_{\text{mmd}}$ maintains representational structure, and $\mathcal{L}_{\text{mix}}$ stabilizes local decision geometry—together enabling stable and robust CTTA.

3.4 Harm-Adaptive Blending with Anchors

While the associated anchors guide the adaptation direction, the model parameters can still accumulate harmful drift over time, especially under long, noisy test streams. To preserve the source-aligned representation encoded by the distilled anchors, we introduce a harm-adaptive blending mechanism that selectively pulls unstable parameter groups back toward the source model, while keeping beneficial updates intact.

After each anchor-based update, we compute the parameter change $\Delta\theta = \theta_{\text{new}} - \theta_{\text{prev}}$ and the anchor-driven gradient $g_S = \nabla_{\theta} \mathcal{L}_{\text{anchor}}$. Parameters are divided into groups $\{G\}$ (*e.g.*, by layer or parameter), and each group’s harm score is measured as:

$$\text{score}(G) = \sum_{p \in G} \left(\langle g_S^{(p)}, \Delta\theta^{(p)} \rangle + \frac{1}{2} \hat{h}^{(p)} (\Delta\theta^{(p)})^2 \right), \quad (13)$$

where $\hat{h}$ is the running average of squared gradients.

Groups whose scores exceed a threshold ρ are partially reverted to the initial source model f_{θ_0} :

$$\theta^{(p)} \leftarrow (1 - \beta_G) \theta^{(p)} + \beta_G \theta_0^{(p)}, \quad (14)$$

Table 2: Classification error rate (%) on ImageNet-to-ImageNet-C. All results are evaluated with the largest corruption severity level 5 in an online manner. We report the performance of our method averaged over 5 runs.

Method	Gau.	shot	imp.	def.	glass	mot.	zoom	snow	fro.	fog	bri.	con.	ela.	pix.	jpeg	Avg.
Source	97.8	97.1	98.2	81.7	89.8	85.2	78.0	83.5	77.1	75.9	41.3	94.5	82.5	79.3	68.6	82.0
TENT _{ICLR21} [57]	81.6	74.6	72.7	77.6	73.8	65.5	55.3	61.6	63.0	51.7	38.2	72.1	50.8	47.4	53.3	62.6
CoTTA _{CVPR22} [59]	84.7	82.1	80.6	81.3	79.0	68.6	57.5	60.3	60.5	48.3	36.6	66.1	47.3	41.2	46.0	62.7
DSS _{WACV24} [62]	82.3	78.4	76.7	81.9	77.8	66.9	60.9	50.8	60.9	47.7	35.4	69.0	47.5	40.9	46.2	62.2
BeCoTTA _{ICML24} [28]	84.1	74.3	72.2	77.4	71.9	63.4	55.1	57.2	61.2	50.7	36.4	66.1	49.2	45.6	48.4	60.9
OBAO _{ECCV24} [81]	78.5	75.3	73.0	75.7	73.1	64.5	56.0	55.8	58.1	47.6	38.5	58.5	46.1	42.0	43.4	59.0
TCA _{CVPR25} [45]	78.3	71.8	73.5	74.4	73.5	63.3	56.5	56.9	59.4	48.1	39.6	59.6	47.2	42.9	44.7	59.3
EATA _{ICML22} [46]	76.3	66.5	65.0	73.1	69.1	62.1	53.5	58.9	59.3	48.1	35.9	62.8	47.5	43.9	47.5	58.0±0.18
RMT _{CVPR23} [13]	77.9	73.1	69.9	73.5	71.1	63.1	57.1	57.1	59.2	50.4	42.9	60.1	49.0	45.7	46.9	59.8±0.18
ROID _{WACV24} [43]	71.7	62.2	62.2	69.6	66.5	57.1	49.3	52.3	57.4	43.5	33.4	59.1	45.4	41.8	46.2	54.5±0.10
ASR _{ICLR26} [38]	70.5	61.3	61.0	69.2	66.5	57.3	49.9	52.5	57.3	44.5	34.3	57.4	45.2	42.1	45.2	54.3±0.14
EATA + DO-ALL	73.2	63.3	62.7	73.0	67.8	61.3	53.0	56.4	58.1	46.6	34.9	61.8	47.4	42.6	46.3	56.6±0.22
RMT + DO-ALL	75.6	70.6	69.0	72.7	70.4	61.8	54.4	54.0	57.2	46.0	39.0	56.9	45.9	42.7	44.4	57.4±0.23
ROID + DO-ALL	68.9	60.5	60.5	69.2	65.8	56.6	48.4	51.5	56.9	42.3	32.8	58.9	44.6	41.0	45.4	53.6±0.09
ASR + DO-ALL	69.3	59.9	60.6	69.1	65.6	57.0	49.4	52.1	56.8	43.7	33.7	57.3	44.7	41.2	44.8	53.7±0.13

where $\beta_G = \beta_{\max} \sigma(\beta_s(s_G - 0.5))$ scaled with the normalized harm score $s_G \in [0, 1]$. This blending mitigates harmful updates while retaining benefits from the distilled anchors, resulting in a stable yet flexible CTTA.

4 Experiments

4.1 Settings

Datasets: We conduct extensive experiments to evaluate the robustness of our method under CTTA scenarios. The primary evaluation uses CIFAR100-C and ImageNet-C [19], which extend CIFAR-100 [27] and ImageNet [11] with distribution shifts. Each benchmark comprises fifteen corruption types and five severity levels. Results are reported at severity level five. On these benchmarks, we perform CTTA by adapting online to a sequence of fifteen corruptions, each presented as a stream of 5,000 unlabeled target images, following common practice [13, 43, 45]. We further assess scalability on the CCC benchmark [48]. CCC is designed to probe extreme-length adaptation on a continuously evolving stream of 7.5 million images with three severity levels.

Implementation: Following prior works [13, 43, 45], we employ ResNeXt-29 [63] for CIFAR100-C and ResNet-50 [18] for ImageNet-C and CCC benchmark. Hyperparameters are set to $\lambda_{\text{mmd}} = 10$, $\beta_{\max} = 0.05$, and $\beta_s = 5$ across all experiments. Unless otherwise noted, anchors are distilled with WMDD [40] using the same backbone as the source model to preserve feature compatibility. We also evaluate anchors distilled with DELT [51] and SRe2L [68], observing comparable effectiveness. Interestingly, we find that anchors distilled with a backbone different from the one used for CTTA remain equally effective, implying that DO-ALL can be generalized across architectures.

Table 3: Classification accuracy (%) on CCC benchmark. All results are averaged over 9 runs (3 seeds per each transition speed). $\dagger$ denotes reproduced results. **Bold** indicates the best results.

Method	CCC-Easy	CCC-Medium	CCC-Hard	Avg.
CoTTA [59]	14.9 $\pm$ 0.88	7.7 $\pm$ 0.43	1.1 $\pm$ 0.16	7.9
ETA [46]	41.4 $\pm$ 0.95	1.1 $\pm$ 0.43	0.2 $\pm$ 0.05	14.2
EATA [46]	48.2 $\pm$ 0.60	35.4 $\pm$ 1.02	8.7 $\pm$ 0.80	30.8
SANTA [6]	47.8 $\pm$ 0.46	32.7 $\pm$ 0.80	9.1 $\pm$ 0.60	29.9
RDumb [48]	49.3$\pm$0.88	38.9 $\pm$ 1.40	9.6 $\pm$ 1.60	32.6
TCA [45]	49.1 $\pm$ 0.35	39.5 $\pm$ 0.53	10.1 $\pm$ 0.22	32.9
ROID † [43]	48.5 $\pm$ 0.63	39.2 $\pm$ 1.98	13.2 $\pm$ 2.47	33.6
ROID+DO-ALL	48.9 $\pm$ 0.64	39.6$\pm$1.97	15.5$\pm$1.49	34.7

Table 4: Ablation study for DO-ALL. The results report the classification error rate (%) on ImageNet-to-ImageNet-C. The baseline method is ROID [43]. **Bold** indicates the best results.

Exp.	$\mathcal{L}_{\text{dir}}$	$\mathcal{L}_{\text{mix}}$	$\mathcal{L}_{\text{mmd}}$	Blend	Avg.
Baseline					54.5
A	✓				54.1
B		✓			54.1
C			✓		54.2
D	✓	✓	✓		54.0
E	✓	✓	✓	✓	53.6

4.2 Performance Evaluation

Imagenet-C We first evaluate DO-ALL with three representative CTTA baselines, including EATA [46], RMT [13]³, and ROID [43]. As DO-ALL is designed in a plug-and-play manner, it can be seamlessly integrated into the conventional algorithms, without modifying the adaptation objective or the architecture of these methods. The distilled anchor set is carried with the model and referenced during adaptation. Across all evaluations, we observe consistent improvements regardless of the baseline, indicating that DO-ALL provides complementary advantages for the existing CTTA algorithms.

Table 2 reports results on ImageNet-to-ImageNet-C under the highest corruption severity (level 5) in an online adaptation setting. DO-ALL consistently improves each base method across nearly all corruption types. For example, EATA+DO-ALL reduces the average error from 58.0% to 56.6% and RMT+DO-ALL from 59.8% to 57.4%. Given that severity level 5 induces highly unstable adaptation dynamics, these improvements highlight DO-ALL’s ability to anchor representation updates, preventing drift and ensuring more stable long-horizon adaptation. The Results on CIFAR100-C are in Supplementary Material.

CCC Benchmark We further evaluate DO-ALL under the CCC benchmark, which is designed to measure stability under long-horizon, continuously accumulating distribution shifts. Unlike corruption benchmarks such as ImageNet-C and CIFAR100-C, where shifts occur independently per sample, CCC gradually transitions the test stream across domains, making models highly vulnerable to representation drift. As shown in Table 3, ROID benefits noticeably from the introduction of DO-ALL. While ROID achieves an average accuracy of 33.6%, DO-ALL improves this to 34.7%, with the significant gain in the most important Hard scenario (13.2% to 15.5%). This setting is particularly challenging due to the prolonged exposure to unseen distributions, and the improvement confirms

³ In fact, RMT itself stores source prototypes and leverages them for CTTA. Our experimental results show that incorporating information from DD through the proposed method can still yield meaningful gains even on top of this approach.

Table 5: Ablation study of the anchor association method. The results report the classification error rate (%) on ImageNet-to-ImageNet-C, averaged over 5 runs. All experiments are conducted with IPC=10 setting. **Bold** indicates the best results.

CTTA	Baseline	Random	Farthest	Nearest
EATA [46]	58.0	57.3	58.7	56.6
RMT [13]	59.8	60.2	61.9	57.4
ROID [43]	54.5	54.0	54.0	53.6

Table 6: Accuracy–efficiency trade-off on ImageNet-to-ImageNet-C. We report classification error rate (%) with FPS, peak GPU memory, and anchor storage on CPU. Experiments are conducted using anchors with IPC = 10.

Method	FPS	GPU Mem (GB)	Anchor Size (GB)	Err. (%)
Baseline (ROID [43])	284.9	9.32	N/A	54.5
+DO-ALL (stride-1)	106.3	15.94	1.55	53.6
+DO-ALL (stride-3)	187.5	15.94	1.55	53.7
+DO-ALL (stride-5)	220.4	15.94	1.55	53.8
+DO-ALL (stride-7)	244.0	15.94	1.55	53.9

that the distilled anchors serve as a persistent reference that prevents collapse over long periods.

DO-ALL not only improves short-term adaptation quality but also delivers robust long-term stability in continuously evolving environments, confirming the claim of our **DO-ALL** framework: **Distill Once, Adapt Life-Long**.

4.3 Ablation Study

Components Analysis Table 4 presents the ablation results of the four components in DO-ALL. Introducing each component individually (Rows A~C) yields only small but consistent reductions in error, indicating that Replay, MMD, and MixUp each contribute non-overlapping stabilization effects. Specifically, Replay (A) slightly reduces forgetting, MMD alignment (B) mitigates representation drift, and MixUp smoothing (C) regularizes local adaptation behavior. Because these components operate on different aspects of the adaptation process, their contributions do not conflict.

When combined (Row D), the improvement becomes clearer, reducing the average error from 54.5% to 54.0%. Finally, harm-adaptive Blending (E, DO-ALL) yields the best performance of 53.6%, confirming that preventing harmful parameter drift accumulated over long adaptation trajectories is essential to stable CTTA. The results show that the components are complementary, and their combination is necessary for the full benefit of DO-ALL.

Anchor Association Strategy While DD provides high-quality anchors, their effectiveness depends critically on how they are used during adaptation. Table 5 compares different strategies for assigning anchors to target samples. Assigning the nearest anchor in feature space consistently yields the best accuracy across all baselines. In contrast, selecting anchors at random provides negligible benefit, and selecting the farthest anchor even degrades performance (*e.g.*, RMT: 59.8% $\rightarrow$ 61.9%). This confirms that DO-ALL’s stabilization does not arise merely from replaying DD-based anchors, but specifically from preserving semantic correspondence between shifted target and their counterparts.

Table 7: Classification error rate (%) on ImageNet-to-ImageNet-C using various DD methods. All results are evaluated with the largest corruption severity level 5 in an online manner.

CTTA	Baseline	SRe2L	DELT	WMDD
EATA [46]	58.0	56.6	56.4	56.6
RMT [13]	59.8	57.5	56.9	57.4
ROID [43]	54.5	53.7	53.6	53.6

Table 8: Impact of images per class (IPC) on DO-ALL. The results report the classification error rate (%) on ImageNet-to-ImageNet-C, over 5 runs.

CTTA	Baseline	IPC=1	IPC=5	IPC=10
EATA [46]	58.0	57.8	56.7	56.6
RMT [13]	59.8	57.9	57.6	57.4
ROID [43]	54.5	53.7	53.6	53.6

4.4 Efficiency

DO-ALL augments a base CTTA method with anchor-guided objectives and harm-adaptive blending, while leaving the original CTTA pipeline unchanged. Importantly, the distilled anchor set is generated *once* before deployment (offline distillation) and carried alongside the model at test time. As a result, DO-ALL introduces no additional distillation cost during deployment and can be plugged into existing CTTA algorithms as a lightweight stabilization component.

To further reduce the overhead during CTTA, we test to compute the anchor branch only every k test-time updates (*stride-k*). At step t , if $t \bmod k \neq 0$, we optimize only the base CTTA objective; if $t \bmod k = 0$, we additionally apply L_{anchor} and perform harm-adaptive blending. This amortizes the anchor computation across updates without altering the core CTTA optimization. As shown in Table 6, DO-ALL consistently improves robustness over the ROID baseline even with large strides (e.g., stride-7), demonstrating a favorable accuracy–efficiency trade-off. We additionally report frames per second (FPS) and peak GPU memory (GB) on a single RTX 3090 using identical batch size and data-loading settings across methods. We also include the host-side (CPU) storage required for the anchor bank. Table 6 summarizes the resulting trade-off in terms of error, FPS, GPU memory, and CPU anchor storage.

Overall, DO-ALL incurs only modest test-time overhead and requires a small host-side anchor storage, while delivering consistent robustness gains. This suggests that DO-ALL can be an attractive direction for CTTA.

4.5 In-Depth Analysis on Source-Distilled Anchors

Experiments on Diverse DD Methods To examine whether DO-ALL relies on a specific dataset distillation algorithm, we also evaluate its performance using several recent DD methods: SRe2L [68], DELT [51], and WMDD [40]. Results are shown in Table 7. Across all three CTTA baselines, every DD variant consistently improves performance over the baseline. For example, ROID+DO-ALL improves from 54.5% to 53.7%, 53.6%, and 53.6% when using SRe2L, DELT, and WMDD, respectively. These results show that DO-ALL does not depend on the choice of DD algorithm. While different DD algorithms vary in how they construct

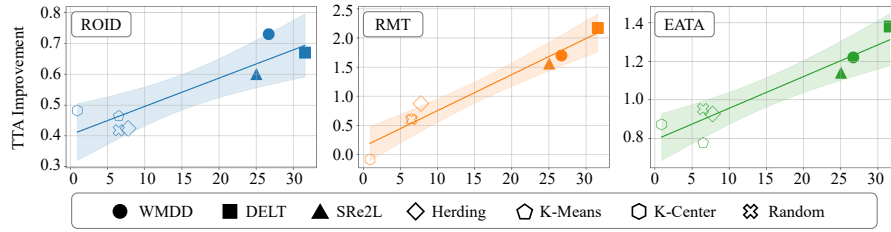


Fig. 3: Correlation between the quality of the anchor set and its effect on CTTA. The x-axis denotes the validation accuracy obtained by training a classifier solely on the small set, while the y-axis shows the corresponding improvement achieved by integrating the set into DO-ALL. A positive trend indicates that higher-quality anchors lead to stronger CTTA stabilization, highlighting DO-ALL’s ability to exploit semantic richness in distilled data.

synthetic samples, DO-ALL is able to effectively utilize the distilled set regardless of the specific DD objective or formulation.

Analysis on IPC We further examine how the number of distilled samples per class (IPC) influences performance (Table 8). Across all CTTA baselines, introducing even a very small anchor set ($IPC = 1$) already yields noticeable improvements over the baseline, indicating that DO-ALL does not require a large number of anchors to be effective—a lightweight distilled set is sufficient to stabilize adaptation. Additionally, increasing IPC provides generally consistent gains, reflecting that richer distilled information can further enhance stability. DO-ALL allows users to flexibly choose the IPC that best fits their application, balancing performance improvements against the size of source-distilled anchors.

Impact of DD Performance on CTTA Gain To further analyze the relationship between the quality of the distilled (or coreset) anchors and their impact within DO-ALL, we plot in Fig. 3 the validation performance of each small set (x-axis) against the resulting TTA improvement (y-axis). Here, the validation performance refers to the accuracy obtained when a classifier is trained only on the small distilled or coreset set and then evaluated on the original source validation data. This metric effectively reflects how well the small set preserves the semantic structure of the source distribution and is widely regarded as the standard evaluation criterion in the DD literature.

Interestingly, we observe a positive correlation across all CTTA baselines: anchor sets that achieve higher standalone validation accuracy also yield larger improvements when used within DO-ALL. Notably, using dataset-distilled anchors (filled markers) consistently delivers larger TTA gains than coreset-selected anchors (hollow markers) across all CTTA methods. This indicates that DO-ALL’s gains are not merely due to retaining a subset of source information, but rather stem from the structured, optimization-driven compression enabled by dataset distillation. Moreover, the result suggests that higher semantic fidelity of the distilled set translates into stronger stabilization during CTTA, leading to more

Table 9: Experiments with data distilled using different architectures (ResNet-18) on ImageNet-to-ImageNet-C over 5 runs.

CTTA	Baseline	SRe2L	DELT	WMDD
ROID [43]	54.5	53.7	53.7	53.7
RMT [13]	59.1	57.8	58.1	57.8
EATA [46]	57.8	56.8	56.6	56.7

Table 10: Experiments on ImageNet-C with a ViT architecture. ROID is used as the baseline CTTA method. For experiments with LGM, we set IPC to 1.

DD Method	TTA Arch.	Baseline	DO-ALL
LGM [4] with ViT-B	RN50	54.5	53.8
LGM [4] with ViT-B	ViT-B	45.0	44.5
WMDD [40] with RN50	ViT-B	45.0	44.4

robust adaptation. In other words, as DD methods continue to advance in capturing richer and more compact source semantics, DO-ALL can readily transfer these gains into more stable and robust CTTA performance by leveraging distilled anchors as high-density semantic cues.

Together with the earlier results demonstrating compatibility across diverse CTTA methods (Sec. 4.2), this correlation strongly suggests DO-ALL’s role as a plug-and-play bridge that seamlessly connects and benefits from progress in both fields of DD and CTTA.

Cross-Architecture Experiments To further assess the flexibility of DO-ALL, we study whether distilled anchors remain effective when the architecture used for distillation differs from that of the CTTA model. As shown in Table 9, we distill anchors with ResNet-18 (RN18) and apply them to CTTA with ResNet-50 (RN50). Across all baselines (ROID, RMT, and EATA), performance is virtually unchanged compared to using anchors distilled from the same backbone, indicating strong cross-backbone transfer. We also evaluate ViT-based anchors and cross-family transfer (Table 10). Specifically, we distill anchors with ViT-B and test them on ResNet-50 CTTA (ViT→CNN), and we further consider both matched (ViT→ViT) and reverse transfer (CNN→ViT) settings. In all cases, DO-ALL consistently improves over the corresponding baseline. Overall, these results suggest that DO-ALL does not require architectural alignment between the distillation and adaptation stages.

5 Conclusion

In this work, we explore the use of dataset distillation as a compact mechanism for retaining source-domain knowledge in continual test-time adaptation. By leveraging a compact set of source-distilled anchors and a lightweight integration strategy, DO-ALL offers a plug-and-play stabilization layer for CTTA that works across different adaptation algorithms and different distillation methods, without modifying the underlying model. Our experiments demonstrate that DO-ALL meaningfully improves long-term adaptation stability across multiple datasets and settings, while offering a flexible performance–memory trade-off. In addition, DO-ALL is orthogonal to most test-time objectives and can be combined with future advances in both CTTA and dataset distillation, making it a practical pathway for continual deployment under strict privacy and storage constraints.

We believe a promising direction is to further improve anchor selection and update schedules under non-stationary streams, and to study robustness when the target shift is abrupt or adversarially corrupted. More broadly, our findings suggest that distilled data can function as a general-purpose, compact memory to stabilize online learning beyond CTTA.

6 Acknowledgement

This work was supported by the National Research Foundation of Korea(NRF) grant funded by the Korea government(MSIT) (RS-2026-25473963, RS-2026-25478915).

References

1. Adachi, K., Yamaguchi, S., Kumagai, A., Hamagami, T.: Test-time adaptation for regression by subspace alignment. arXiv preprint arXiv:2410.03263 (2024)
2. Adachi, K., Yamaguchi, S., Kumagai, A.: Covariance-aware feature alignment with pre-computed source statistics for test-time adaptation to multiple image corruptions. In: 2023 IEEE International Conference on Image Processing (ICIP). pp. 800–804. IEEE (2023)
3. Boudiaf, M., Mueller, R., Ben Ayed, I., Bertinetto, L.: Parameter-free online test-time adaptation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8344–8353 (2022)
4. Cazenavette, G., Torralba, A., Sitzmann, V.: Dataset distillation for pre-trained self-supervised vision models. arXiv preprint arXiv:2511.16674 (2025)
5. Cazenavette, G., Wang, T., Torralba, A., Efros, A.A., Zhu, J.Y.: Dataset distillation by matching training trajectories. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4750–4759 (2022)
6. Chakrabarty, G., Sreenivas, M., Biswas, S.: Santa: Source anchoring network and target alignment for continual test time adaptation. Transactions on Machine Learning Research (2023)
7. Choi, S., Yang, S., Choi, S., Yun, S.: Improving test-time adaptation via shift-agnostic weight regularization and nearest source prototypes. In: European Conference on Computer Vision. pp. 440–458. Springer (2022)
8. Cui, J., Wang, R., Si, S., Hsieh, C.J.: Dc-bench: Dataset condensation benchmark. Advances in Neural Information Processing Systems **35**, 810–822 (2022)
9. Cui, J., Wang, R., Si, S., Hsieh, C.J.: Scaling up dataset distillation to imagenet-1k with constant memory. In: International Conference on Machine Learning. pp. 6565–6590. PMLR (2023)
10. Cui, X., Qin, Y., Zhou, W., Li, H., Li, H.: Optical: Leveraging optimal transport for contribution allocation in dataset distillation. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 15245–15254 (2025)
11. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: 2009 IEEE conference on computer vision and pattern recognition. pp. 248–255. Ieee (2009)
12. Deng, W., Li, W., Ding, T., Wang, L., Zhang, H., Huang, K., Huo, J., Gao, Y.: Exploiting inter-sample and inter-feature relations in dataset distillation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 17057–17066 (2024)

13. Döbler, M., Marsden, R.A., Yang, B.: Robust mean teacher for continual and gradual test-time adaptation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7704–7714 (2023)
14. Dong, T., Zhao, B., Lyu, L.: Privacy for free: How does dataset condensation help privacy? In: International Conference on Machine Learning. pp. 5378–5396. PMLR (2022)
15. Du, J., Jiang, Y., Tan, V.Y., Zhou, J.T., Li, H.: Minimizing the accumulated trajectory error to improve dataset distillation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 3749–3758 (2023)
16. Du, J., Shi, Q., Zhou, J.T.: Sequential subset matching for dataset distillation. *Advances in Neural Information Processing Systems* **36**, 67487–67504 (2023)
17. Guo, Z., Wang, K., Cazenavette, G., Li, H., Zhang, K., You, Y.: Towards loss-less dataset distillation via difficulty-aligned trajectory matching. *arXiv preprint arXiv:2310.05773* (2023)
18. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
19. Hendrycks, D., Dietterich, T.: Benchmarking neural network robustness to common corruptions and perturbations. *arXiv preprint arXiv:1903.12261* (2019)
20. Jang, H.K., Kim, J., Kweon, H., Yoon, K.J.: Talos: Enhancing semantic scene completion via test-time adaptation on the line of sight. *Advances in Neural Information Processing Systems* **37**, 74211–74232 (2024)
21. Jeong, J., Kwon, H., Kim, M., Yoon, K.J.: Multimodal distribution matching for vision-language dataset distillation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 23072–23082 (2026)
22. Jung, S., Lee, J., Kim, N., Shaban, A., Boots, B., Choo, J.: Cafa: Class-aware feature alignment for test-time adaptation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 19060–19071 (2023)
23. Kang, J., Kim, N., Kwon, D., Ok, J., Kwak, S.: Leveraging proxy of training data for test-time adaptation (2023)
24. Kim, H., Jang, H.K., Yoon, K.J.: Test-time training for lidar semantic segmentation under corruption via geometric inlier discrimination. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 24206–24216 (2026)
25. Kim, J., Kwon, H., Kweon, H., Jeong, W., Yoon, K.J.: Dc-tta: Divide-and-conquer framework for test-time adaptation of interactive segmentation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 23279–23289 (2025)
26. Kim, J., Kwon, H., Kweon, H., Yoon, K.J.: Bootstrapping video semantic segmentation model via distillation-assisted test-time adaptation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10766–10777 (2026)
27. Krizhevsky, A., Hinton, G., et al.: Learning multiple layers of features from tiny images (2009)
28. Lee, D., Yoon, J., Hwang, S.J.: Becotta: Input-dependent online blending of experts for continual test-time adaptation. *arXiv preprint arXiv:2402.08712* (2024)
29. Lee, S., Chun, S., Jung, S., Yun, S., Yoon, S.: Dataset condensation with contrastive signals. In: International Conference on Machine Learning. pp. 12352–12364. PMLR (2022)

30. Lee, Y., Chung, H.W.: Selmatch: Effectively scaling up dataset distillation via selection-based initialization and partial updates by trajectory matching. arXiv preprint arXiv:2406.18561 (2024)
31. Lei, S., Tao, D.: A comprehensive survey of dataset distillation. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **46**(1), 17–32 (2023)
32. Li, H., Zhou, Y., Gu, X., Li, B., Wang, W.: Diversity-enhanced distribution alignment for dataset distillation. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. pp. 3747–3756 (2025)
33. Li, W., Li, G., Maeda, K., Ogawa, T., Haseyama, M.: Hyperbolic dataset distillation. arXiv preprint arXiv:2505.24623 (2025)
34. Li, Y., Wang, N., Shi, J., Liu, J., Hou, X.: Revisiting batch normalization for practical domain adaptation. arXiv preprint arXiv:1603.04779 (2016)
35. Li, Z., Zhong, X., Khaki, S., Liang, Z., Zhou, Y., Shi, M., Wang, Z., Zhao, X., Zhao, W., Qin, Z., et al.: Dd-ranking: Rethinking the evaluation of dataset distillation. arXiv preprint arXiv:2505.13300 (2025)
36. 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)
37. Lim, H., Kim, B., Choo, J., Choi, S.: Ttn: A domain-shift aware batch normalization in test-time adaptation. arXiv preprint arXiv:2302.05155 (2023)
38. Lim, T., Hwang, J.W., Lee, K.: When and where to reset matters for long-term test-time adaptation. In: *ICLR* (2026)
39. Liu, D., Gu, J., Cao, H., Trinitis, C., Schulz, M.: Dataset distillation by automatic training trajectories. In: *European Conference on Computer Vision*. pp. 334–351. Springer (2024)
40. Liu, H., Li, Y., Xing, T., Dalal, V., Li, L., He, J., Wang, H.: Dataset distillation via the wasserstein metric. arXiv preprint arXiv:2311.18531 (2023)
41. Liu, J., Xu, R., Yang, S., Zhang, R., Zhang, Q., Chen, Z., Guo, Y., Zhang, S.: Continual-mae: Adaptive distribution masked autoencoders for continual test-time adaptation. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 28653–28663 (2024)
42. Ma, Z., Cao, A., Yang, F., Gong, Y., Wei, X.: Curriculum dataset distillation. *IEEE Transactions on Image Processing* (2025)
43. Marsden, R.A., Döbler, M., Yang, B.: 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)
44. Nguyen, T., Chen, Z., Lee, J.: Dataset meta-learning from kernel ridge-regression. arXiv preprint arXiv:2011.00050 (2020)
45. Ni, C., Lyu, F., Tan, J., Hu, F., Yao, R., Zhou, T.: Maintaining consistent inter-class topology in continual test-time adaptation. In: *Proceedings of the Computer Vision and Pattern Recognition Conference*. pp. 15319–15328 (2025)
46. 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)
47. Niu, S., Wu, J., Zhang, Y., Wen, Z., Chen, Y., Zhao, P., Tan, M.: Towards stable test-time adaptation in dynamic wild world. arXiv preprint arXiv:2302.12400 (2023)
48. Press, O., Schneider, S., Kümmerer, M., Bethge, M.: Rdumb: A simple approach that questions our progress in continual test-time adaptation. *Advances in Neural Information Processing Systems* **36**, 39915–39935 (2023)

49. Sajedi, A., Khaki, S., Amjadian, E., Liu, L.Z., Lawryshyn, Y.A., Plataniotis, K.N.: Datadam: Efficient dataset distillation with attention matching. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 17097–17107 (2023)
50. Shao, S., Zhou, Z., Chen, H., Shen, Z.: Elucidating the design space of dataset condensation. *Advances in neural information processing systems* **37**, 99161–99201 (2024)
51. Shen, Z., Sherif, A., Yin, Z., Shao, S.: Delt: A simple diversity-driven earlylate training for dataset distillation. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 4797–4806 (2025)
52. Shin, S., Bae, H., Shin, D., Joo, W., Moon, I.C.: Loss-curvature matching for dataset selection and condensation. In: International Conference on Artificial Intelligence and Statistics. pp. 8606–8628. PMLR (2023)
53. Song, J., Lee, J., Kweon, I.S., Choi, S.: 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)
54. Song, R., Liu, D., Chen, D.Z., Festag, A., Trinitis, C., Schulz, M., Knoll, A.: Federated learning via decentralized dataset distillation in resource-constrained edge environments. In: 2023 International Joint Conference on Neural Networks (IJCNN). pp. 1–10. IEEE (2023)
55. Su, Z., Guo, J., Yao, K., Yang, X., Wang, Q., Huang, K.: Unraveling batch normalization for realistic test-time adaptation. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 38, pp. 15136–15144 (2024)
56. Sun, P., Shi, B., Shang, X., Lin, T.: Information compensation: A fix for any-scale dataset distillation. In: ICLR 2024 Workshop on Data-centric Machine Learning Research (DMLR) (2024)
57. Wang, D., Shelhamer, E., Liu, S., Olshausen, B., Darrell, T.: Tent: Fully test-time adaptation by entropy minimization. *arXiv preprint arXiv:2006.10726* (2020)
58. Wang, K., Zhao, B., Peng, X., Zhu, Z., Yang, S., Wang, S., Huang, G., Bilen, H., Wang, X., You, Y.: Cafe: Learning to condense dataset by aligning features. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12196–12205 (2022)
59. 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)
60. Wang, S., Yang, Y., Liu, Z., Sun, C., Hu, X., He, C., Zhang, L.: Dataset distillation with neural characteristic function: A minmax perspective. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 25570–25580 (2025)
61. Wang, T., Zhu, J.Y., Torralba, A., Efros, A.A.: Dataset distillation. *arXiv preprint arXiv:1811.10959* (2018)
62. Wang, Y., Hong, J., Cheraghian, A., Rahman, S., Ahmedt-Aristizabal, D., Petersson, L., Harandi, M.: Continual test-time domain adaptation via dynamic sample selection. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 1701–1710 (2024)
63. Xie, S., Girshick, R., Dollár, P., Tu, Z., He, K.: Aggregated residual transformations for deep neural networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1492–1500 (2017)
64. Xiong, Y., Wang, R., Cheng, M., Yu, F., Hsieh, C.J.: Feddm: Iterative distribution matching for communication-efficient federated learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 16323–16332 (2023)

65. Yang, E., Shen, L., Wang, Z., Liu, T., Guo, G.: An efficient dataset condensation plugin and its application to continual learning. *Advances in Neural Information Processing Systems* **36**, 67625–67642 (2023)
66. Yang, S., Cheng, S., Hong, M., Fan, H., Wei, X., Liu, S.: Neural spectral decomposition for dataset distillation. In: *European Conference on Computer Vision*. pp. 275–290. Springer (2024)
67. Yang, X., Chen, X., Li, M., Wei, K., Deng, C.: A versatile framework for continual test-time domain adaptation: Balancing discriminability and generalizability. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 23731–23740 (2024)
68. Yin, Z., Xing, E., Shen, Z.: Squeeze, recover and relabel: Dataset condensation at imagenet scale from a new perspective. *Advances in Neural Information Processing Systems* **36**, 73582–73603 (2023)
69. Yuan, L., Xie, B., Li, S.: Robust test-time adaptation in dynamic scenarios. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 15922–15932 (2023)
70. Zhang, H., Li, S., Lin, F., Wang, W., Qian, Z., Ge, S.: Dance: Dual-view distribution alignment for dataset condensation. *arXiv preprint arXiv:2406.01063* (2024)
71. Zhang, H., Li, S., Wang, P., Zeng, D., Ge, S.: M3d: Dataset condensation by minimizing maximum mean discrepancy. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. vol. 38, pp. 9314–9322 (2024)
72. Zhang, H., Cisse, M., Dauphin, Y.N., Lopez-Paz, D.: mixup: Beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412* (2017)
73. Zhang, M., Levine, S., Finn, C.: Memo: Test time robustness via adaptation and augmentation. *Advances in neural information processing systems* **35**, 38629–38642 (2022)
74. Zhang, X., Du, J., Liu, P., Zhou, J.T.: Breaking class barriers: Efficient dataset distillation via inter-class feature compensator. In: *Proceedings of the International Conference on Learning Representations (ICLR)* (2025)
75. Zhao, B., Bilen, H.: Dataset condensation with differentiable siamese augmentation. In: *International Conference on Machine Learning*. pp. 12674–12685. PMLR (2021)
76. Zhao, B., Bilen, H.: Dataset condensation with distribution matching. In: *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. pp. 6514–6523 (2023)
77. Zhao, B., Mopuri, K.R., Bilen, H.: Dataset condensation with gradient matching. *arXiv preprint arXiv:2006.05929* (2020)
78. Zhao, G., Li, G., Qin, Y., Yu, Y.: Improved distribution matching for dataset condensation. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 7856–7865 (2023)
79. Zhong, W., Tang, H., Zheng, Q., Xu, M., Hu, Y., Guan, W.: Towards stable and storage-efficient dataset distillation: Matching convexified trajectory. In: *Proceedings of the Computer Vision and Pattern Recognition Conference*. pp. 25581–25589 (2025)
80. Zhou, Y., Nezhadarya, E., Ba, J.: Dataset distillation using neural feature regression. *Advances in Neural Information Processing Systems* **35**, 9813–9827 (2022)
81. Zhu, Z., Hong, X., Ma, Z., Zhuang, W., Ma, Y., Dai, Y., Wang, Y.: Reshaping the online data buffering and organizing mechanism for continual test-time adaptation. In: *European Conference on Computer Vision*. pp. 415–433. Springer (2024)