Perturbing to Preserve: Defending Fragile Knowledge in Online Continual Learning

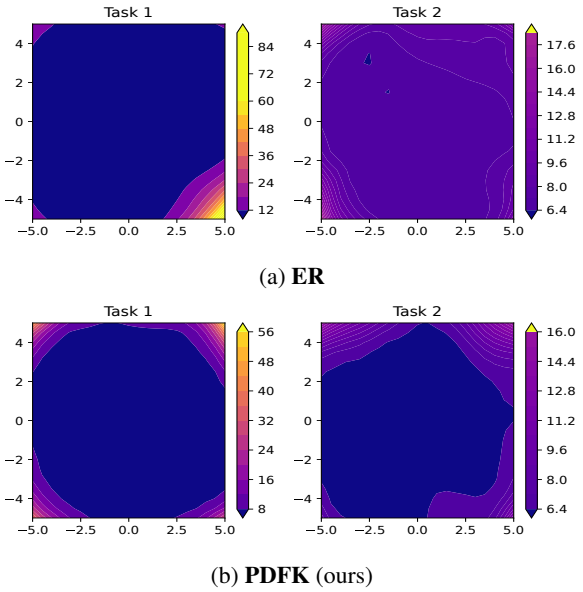
Dulan Zhou[1,2†], **Zijian Gao**[1,2†], **Kele Xu**[1,2*]

[1]College of Computer Science and Technology, National University of Defense Technology
[2]National Key Laboratory of Parallel and Distributed Computing
{dulan_zhou, gaozijian19, xukelele}@nudt.edu.cn

Abstract

Online continual learning mandates the ability to acquire knowledge from non-stationary data streams while preserving previously learned information, yet neural networks often forget. In this work, we uncover and systematically analyze a critical yet underexplored issue in this setting, which we term knowledge fragility: the phenomenon where correctly learned instances are abruptly forgotten following minor parameter updates. We attribute this phenomenon to two factors: (1) temporally, where high-frequency oscillations in parameter space lead to disproportionate forgetting relative to adaptation, and (2) spatially, where fragile instances reside in sharp, high-curvature regions of the loss landscape, making them highly susceptible to optimization noise. To counteract this fragility, we propose PDFK (Perturbing to Defend Fragile Knowledge)—a unified and task-agnostic framework that fortifies fragile knowledge along both temporal and spatial dimensions. Temporally, PDFK stabilizes long-term memory by employing exponential moving average (EMA) to suppress volatile parameter shifts. Spatially, it introduces lightweight, structured perturbations guided by consistency regularization, effectively flattening the local loss surface and enhancing robustness to future updates. Extensive experiments across diverse benchmarks demonstrate that PDFK consistently improves knowledge retention and surpasses state-of-the-art methods in both accuracy and forgetting metrics under challenging streaming settings.

Code — https://github.com/colaudiolab/PDFK



(a) **ER**

(b) **PDFK** (ours)

Figure 1: **Loss landscape across a task switch.**2D projection of the test loss surface around the solution obtained after *Task 1* (left) and *Task 2* (right). (a) ER exhibits sharp ridges, signaling fragile zones that coincide with a sharp accuracy drop. (b) PDFK flattens these regions, enlarging the safe radius and reducing boundary crossings during updates.

1 Introduction

Continual Learning (CL) (Kirkpatrick et al. 2017) aims to enable models to acquire knowledge over time while mitigating *catastrophic forgetting* (French 1999)—the tendency to overwrite previously learned information when exposed to new data. Among its variants, *Online Continual Learning* (OCL) (Lopez-Paz and Ranzato 2017) defines a more realistic yet stringent setting where data arrive as a non-stationary stream and each instance is observed only once (Deng et al. 2023; Soutif-Cormerais et al. 2023). Unlike offline CL, which assumes clear task boundaries and multiple passes, OCL enforces three constraints: (i) single-pass input, (ii) non-i.i.d. data, and (iii) no task annotations—conditions that challenge the plasticity–stability balance and render many offline methods ineffective.

Knowledge distillation (Li and Hoiem 2017; Buzzega et al. 2020; Jung et al. 2023) and feature synthesis (Zhu et al. 2021a,b; Guo, Liu, and Zhao 2022) struggle under OCL constraints: the former suffers from delayed supervision without task boundaries (Deng et al. 2023), while the latter conflicts with tight memory budgets. Although memory-based rehearsal (Chaudhry et al. 2019; Shim et al. 2021) is now standard, we find it insufficient for retaining knowledge reliably. Even with replay, state-of-the-art models remain alarmingly sensitive: confidently predicted instances at time t may be misclassified just a few gradient steps later—a phenomenon

*Corresponding author. † Equal Contribution.
Copyright © 2026, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

we term *knowledge fragility*.

Our analysis shows this fragility arises in **high-curvature regions** of the loss landscape, where predictions become sensitive to small parameter changes (see Fig. 1). Crucially, this fragility is not random but concentrated near decision boundaries, driven by a *temporal–spatial dual mechanism*. **Spatially**, sharp curvature reduces the fragility radius—the minimum displacement needed to flip a prediction. **Temporally**, noisy online updates induce parameter oscillations that traverse these narrow margins, causing rapid forgetting. This perspective reframes the stability–plasticity dilemma: while OCL models must remain adaptive, they lack mechanisms to protect brittle representations.

To tackle this, we propose **PDFK** (*Perturbing to Defend Fragile Knowledge*), targeting both temporal and spatial fragility. **Temporally**, we apply Exponential Moving Average (EMA) (Tarvainen and Valpola 2017)to smooth parameter trajectories and suppress high-frequency noise. **Spatially**, we inject minimal structured perturbations and enforce consistency on fragile instances, flattening the landscape and enlarging their fragility radius. Together, these techniques reduce fragile knowledge and improve retention across OCL benchmarks.

In summary, Our main contributions are as follows:

- We identify and formalize the phenomenon of *knowledge fragility* in OCL, and uncover its roots in a temporal–spatial dual mechanism, explaining the limitations of existing replay-based strategies.
- We propose **PDFK**, a task-agnostic framework that combines temporal smoothing (via EMA) and spatial consolidation (via structured perturbations and consistency) without requiring additional memory buffers.
- We conduct extensive experiments under non-i.i.d. and task-free streaming settings, showing that PDFK significantly improves retention of fragile knowledge and achieves state-of-the-art performance across multiple OCL benchmarks.

2 Related Work

Continual Learning Continual Learning methods can be broadly classified into three categories based on their implementation strategies (Wang et al. 2024b,a; De Lange et al. 2021; Feng et al. 2025a,b): regularization-based, parameter isolation-based, and replay-based approaches. Regularization methods (Buzzega et al. 2020; Li and Hoiem 2017; Gao et al. 2024b, 2025c, 2024a, 2025a) constrains parameter updates to retain prior knowledge. Parameter isolation (Chaudhry et al. 2018; Konishi et al. 2023) assigns task-specific parameters to avoid interference. Replay revisits past data—real or synthetic—to mitigate forgetting. Among them, replay-based approaches (Wan et al. 2025; Lopez-Paz and Ranzato 2017; Shim et al. 2021; Gao et al. 2025b) offer superior scalability and remain central to CL in dynamic environments.

Online Continual Learning Online Continual Learning is a constrained form of CL that emphasizes real-time adaptation, where data arrives sequentially and can be processed only once, typically in single samples or mini-batches. In such strict settings, replay-based methods have become the dominant paradigm. Early approaches like **ER** (Chaudhry et al. 2019) combine cross-entropy loss with a random memory buffer for sample replay. **OCM** (Guo, Liu, and Zhao 2022) leverages mutual information maximization to mitigate feature bias and retain past knowledge. **GSA** (Guo, Liu, and Zhao 2023) introduces a gradient-sensitive optimizer to address dynamic training bias, while **OnPro** (Wei et al. 2023) employs prototype alignment to counteract shortcut learning. **MOSE** (Yan et al. 2024) alleviates forgetting by integrating multi-level supervision and reverse self-distillation. **S6MOD** (Liu et al. 2025) enhances adaptability in OCL by introducing a discretization-based auxiliary branch and class-conditional routing.

Blurry Task Boundaries Conventional CL protocols often assume *known, sharp* task boundaries during training, enabling explicit task identification and task–specific adaptation (Riemer et al. 2019; Chaudhry et al. 2019; Guo, Liu, and Zhao 2022). In real streaming OCL, however, data arrive continuously without segmentation or task labels, making such assumptions unrealistic. To bridge this gap, the *blurry* setting (Michel et al. 2024) models task shifts as *smooth* rather than abrupt: around each transition, samples from adjacent tasks co–occur and their proportions vary gradually. This protocol better reflects real drift and avoids boundary cues that can overstate performance.

3 Methodology

Background and Notation

Problem Setup We consider the setting of OCL where the learner observes a non-stationary stream of labeled data $\{(x_t, y_t)\}_{t\geq 1} \sim \mathcal{P}_t$, without access to future samples. At each time step t, the model receives a mini-batch $\mathcal{B}_t$ and updates its parameters by optimizing on a *mixed batch* that combines $\mathcal{B}_t$ with a replay batch $\mathcal{B}_{\mathcal{M}} \subset \mathcal{M}$ drawn from a fixed-size memory buffer $\mathcal{M}$:

$$\mathcal{L}_{\text{total}} = \mathcal{L}(f_\theta(\mathcal{B}_t), y_t) + \lambda_r \cdot \mathcal{L}(f_\theta(\mathcal{B}_{\mathcal{M}}), y_{\mathcal{M}}),$$

where f_θ denotes the model with parameters θ, and λ_r balances the contribution of the replay term. When explicit task boundaries are available for evaluation, we follow the common assumption that each task has a disjoint label space.

Blurry task boundaries To model the more realistic case where task shifts are gradual rather than discrete, let the global label space be partitioned into latent tasks with distributions $\{\mathcal{P}^{(k)}\}_{k=1}^{K}$. Instead of switching abruptly, the stream is a smooth mixture:

$$(x_t, y_t) \sim \sum_{k=1}^{K} \pi_k(t)\, \mathcal{P}^{(k)}(x, y), \qquad \sum_k \pi_k(t) = 1, \quad (1)$$

where the time-varying weights are given by normalized kernels $\pi_k(t) = \phi\big((t-\mu_k)/\sigma\big) \Big/ \sum_j \phi\big((t-\mu_j)/\sigma\big)$ (Gaussian or HalfNormal). The overlap parameter σ controls the degree of *blurriness*: larger σ yields stronger temporal mixing and thus higher ambiguity. The blurry regime introduces

additional difficulty by obfuscating task boundaries, thereby requiring models to exhibit stronger temporal robustness and memory stability in the absence of explicit shifts.

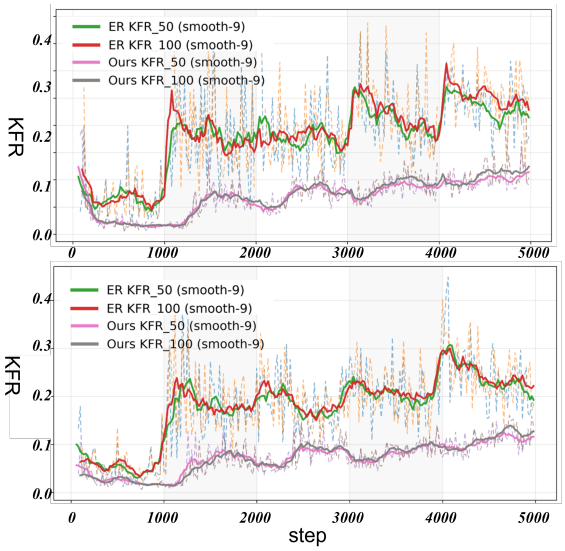


Figure 2: K-Step forgetting rate curves exhibit abrupt forgetting within tens of steps post-learning. *Top:* Clear Task Boundaries. *Bottom:* Blurry Task Boundaries.

Empirical Analysis

Fast distribution shifts in OCL streams lead to rapid parameter drift, which in turn exacerbates short-term forgetting. We observe that some replayed samples, although correctly classified at step t, are misclassified after only a few subsequent gradient updates despite minimal global parameter change. This suggests that newly acquired knowledge is inherently unstable and highly sensitive to the learning dynamics of online settings. To quantify this phenomenon, we introduce the **K-step forgetting rate (KFR)**, defined as the proportion of samples that are correctly predicted at time t but forgotten within the next K optimization steps:

$$\mathrm{KFR}_K = \frac{1}{|\mathcal{S}_t|} \sum_{(x,y)\in\mathcal{S}_t} 1\{f_t(x) = y,\ f_{t+K}(x) \neq y\}. \quad (2)$$

Here, $\mathcal{S}_t$ denotes the evaluation set at step t. Intuitively, KFR_K captures the fragility of model predictions—how easily newly learned representations can be overwritten in the short term. A persistently high KFR signals poor robustness and highlights the transient nature of recent learning. We report KFR_K curves under both clear and blurry task boundaries, using K=50 and 100. As shown in Figure 2, forgetting rates remain consistently above zero, and often spike shortly after encountering new data. This provides empirical evidence for the existence of *fragile knowledge*, referring to representations that are quickly acquired yet highly susceptible to forgetting within a short span of updates.

Theoretical Foundation

Above findings motivate a deeper theoretical inquiry into the origin and behavior of fragile knowledge. Figure 3 illustrates the parameter trajectory $\{\theta^{t-i}\}_{i=0}^{n}$ drifts into a high–curvature ridge; subsequent microscopic updates push θ^t across a nearby decision boundary, instantaneously erasing the prediction.

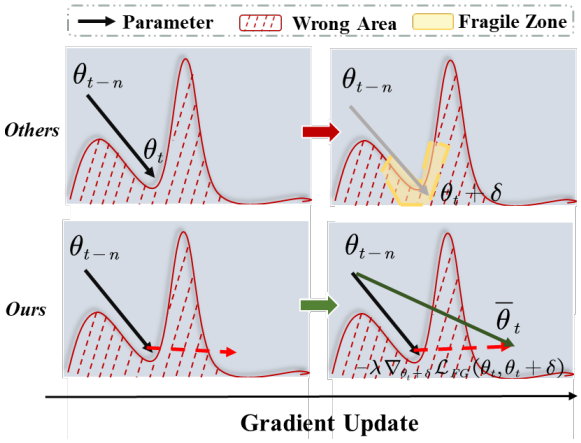


Figure 3: **Knowledge fragility under online parameter shift.** *Top:* A replay baseline updates parameters from θ_{t-n} to θ_t without considering local output consistency, pushing the solution into a FragileZone, where a slight future step δ flips predictions (yellow area). *Bottom:* PDFK guides the update toward a flatter basin $\bar{\theta}_t$ via EMA and perturbation–consistency, shrinking the fragile zone.

Margin geometry Let $z_\theta(x) = softmax(f_\theta(x)) \in \mathbb{R}^C$ denote the logits and define the classification margin

$$\gamma_\theta(x,y) = z_\theta(x)_y - \max_{c\neq y} z_\theta(x)_c. \quad (3)$$

A sample is correctly classified if $\gamma_\theta(x,y) > 0$. For a small parameter perturbation δ, a first–order expansion gives

$$\gamma_{\theta+\delta}(x,y) \approx \gamma_\theta(x,y) + \nabla_\theta \gamma_\theta(x,y)^\top \delta. \quad (4)$$

The *minimal flip displacement* is thus

$$r_\theta(x,y) \ :=\ \frac{|\gamma_\theta(x,y)|}{\|\nabla_\theta \gamma_\theta(x,y)\|_2}, \quad (5)$$

the radius within which an adversarial move can change the label under linear approximation. We call $r_\theta(x,y)$ the *fragility radius*. Small margins and large gradient sensitivity jointly yield a small r_θ, making the instance intrinsically easy to forget.

Spatial Attribution (High Curvature) Let $\mathcal{L}_\theta$ be the average loss over seen data and $H_\theta = \nabla^2_\theta \mathcal{L}_\theta$ its Hessian. Along the update direction δ,

$$\gamma_{\theta+\delta}(x,y) = \gamma_\theta(x,y) + \nabla_\theta \gamma_\theta(x,y)^\top \delta + \tfrac{1}{2}\,\delta^\top \nabla^2_\theta \gamma_\theta(x,y)\,\delta + O(\|\delta\|^3). \quad (6)$$

A large local spectral norm $\|H_\theta\|_2$ (sharp region) amplifies both $\|\nabla_\theta \gamma_\theta(x,y)\|$ and the second–order term, shrinking $r_\theta(x,y)$ and expanding the set of points whose labels can flip under microscopic motion. This constitutes the *spatial* cause: fragile knowledge concentrates in high–curvature basins.

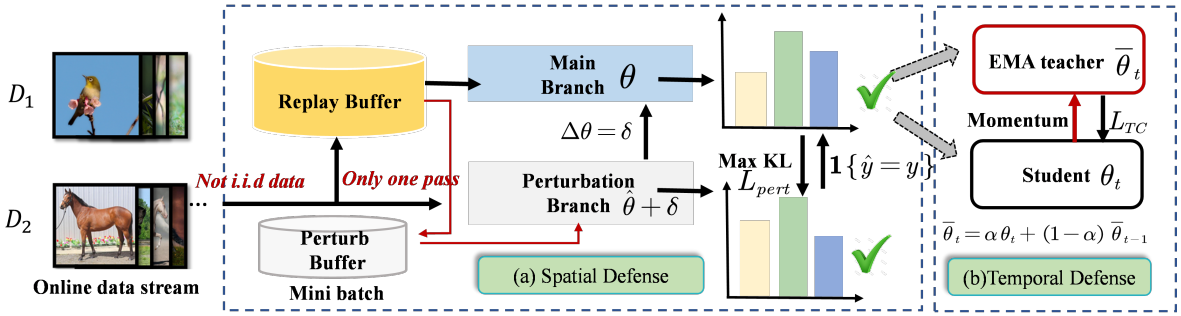


Figure 4: An overview of the proposed PDFK framework. The model receives non-i.i.d. data in a streaming setting with only one pass. A *replay buffer* stores past samples for rehearsal, while a *perturb buffer* caches correctly classified samples for perturbation. (a) A perturbed branch with parameters $\hat{\theta} + \delta_0$ is constructed to identify fragile knowledge via KL divergence. (b) An EMA teacher provides soft supervision to the student model through momentum distillation.

Temporal Attribution (High-frequency Instability) Let the online parameter dynamics be

$$\theta^{t+1} = \theta^t - \eta g_t, \qquad g_t = \underbrace{\bar{g}_t}_{\text{drift}} + \underbrace{\xi_t}_{\text{zero-mean noise}}, \tag{7}$$

where η is the learning rate and ξ_t captures stochastic gradient noise and fast task shifts. Over n steps the cumulative displacement decomposes as

$$\theta^t - \theta^{t-n} = -\eta \sum_{i=0}^{n-1} \bar{g}_{t-i} \; - \; \eta \sum_{i=0}^{n-1} \xi_{t-i}. \tag{8}$$

Even when the *net* norm $\|\theta^t - \theta^{t-n}\|_2$ is small (due to oscillations canceling out), the trajectory may have repeatedly traversed low $r_\theta(\cdot)$ regions, causing label flips. High variance or high-frequency components of $\{\xi_t\}$ accelerate crossings of fragile radii. This constitutes the *temporal* cause: instability in parameter evolution allows fragile examples to be forgotten *before* sufficient consolidation occurs.

Unified definition We formalize fragility over a short temporal window as follows.

Definition 1 (Temporal–Spatial Fragility)**.** *A replay example $(x, y) \in \mathcal{M}$ is (ε, n)-*fragile *at step t if*

$$\begin{cases} \gamma_{\theta^{t-n}}(x, y) > 0, \\ \gamma_{\theta^t}(x, y) \le 0, \\ \|\theta^t - \theta^{t-n}\|_2 \le \varepsilon. \end{cases} \tag{9}$$

Equivalently, its label flips within n online updates under a cumulative displacement no larger than ε.

Condition (9) is satisfied precisely when the time–integrated path intersects the ball of radius $r_{\theta^\tau}(x, y)$ for some $\tau \in [t - n, t]$. Sharp curvature (spatial) reduces r_{θ^τ}, while high-frequency noise (temporal) increases the probability of such intersections.

Perturbing to Defend Fragile Knowledge

Knowledge fragility reframes the stability–plasticity dilemma: effective OCL must *(i)* suppress high-frequency oscillations to expand the temporal window before crossings, and *(ii)* flatten sharp regions to enlarge fragility radii. In this section we introduce **PDFK**, which operationalizes these two perspectives via EMA (temporal smoothing) and minimal structured perturbations with consistency regularization (spatial flattening), jointly reducing the measure of (ε, n)-fragile knowledge.

Spatial Defense via Targeted Perturbation At time t we have parameters θ_t, a current stream example (x_t, y_t), and replay buffer $\mathcal{M}_t$. Let $\mathcal{B}_t^* \subset \mathcal{M}_t$ be a mini–batch of correctly classified replay samples (these are the candidates with small positive margins), $x^* \in \mathcal{B}_t^*$. Denote $q_\theta(x) = \mathrm{softmax}(z_\theta(x))$.

Phase 1: Local worst–case perturbation. We seek a perturbation δ that locally maximizes the divergence between the perturbed and unperturbed predictions on $\mathcal{B}_t^*$:

$$\max_{\|\delta\| \le \rho} \mathrm{KL}\big(q_{\theta_t+\delta}(x^*) \,\big\|\, q_{\theta_t}(x^*)\big), \qquad x^* \in \mathcal{B}_t^*. \tag{10}$$

Because q_θ is non-convex and $\delta = 0$ is the global *minimum* of (10) (zero gradient), we initialize δ away from zero with small per–layer noise proportional to the layer norm:

$$\delta_0 \sim \mathcal{N}\big(0,\ \epsilon^2 \|\theta_t\|_2^2 I\big)\,. \tag{11}$$

where δ_0 is the initial Gaussian perturbation, ϵ is a small, positive scalar, and I is the identity matrix with appropriate dimensions.A single normalized gradient–ascent step provides a first–order approximation to the maximizer:

$$\begin{aligned} \mathbf{v} &:= \nabla_{\theta_t+\delta_0} \mathcal{D}_{\mathrm{KL}}\big(q_{\theta_t}(x^*) \,\|\, q_{\theta_t+\delta_0}(x^*)\big), \\ \delta &:= \delta_0 \;+\; \gamma \frac{\|\theta_t\|_2}{\|\mathbf{v}\|_2}\,\mathbf{v}, \end{aligned} \tag{12}$$

where $\mathbf{v}$ is the gradient of the KL term, and $\gamma > 0$ a step size. The factor $\|\theta_t\|_2 / \|\mathbf{v}\|_2$ normalizes the update so that the effective perturbation magnitude scales with the current parameter norm and is insensitive to raw gradient scale.

Because δ=0 makes the KL divergence zero with vanishing gradient, we first inject small noise δ_0 to obtain a nonzero direction and then take one normalized gradient–ascent step along $\mathbf{v}$. This yields a first–order approximation of the local worst–case perturbation that most strongly distorts the predictions and exposes spatial fragility.

Boundary	Method	CIFAR10			CIFAR100			ImageNet-Subset			Tiny-IN		
		200	500	1000	200	500	1000	1000	2000	5000	1000	2000	5000
Clear	ER [NeurIPS'19]	45.72	52.08	65.59	12.02	16.55	25.12	15.43	24.69	33.87	5.82	11.60	17.98
	ER + SDP [ICML'23]	44.92	57.58	65.8	13.11	21.14	28.70	20.84	22.95	32.17	11.10	15.32	23.22
	DER++ [NeurIPS'20]	49.53	55.64	59.68	10.26	17.58	22.37	13.70	18.76	8.57	4.11	4.22	4.39
	DVC [CVPR'22]	49.04	59.82	59.12	11.16	18.60	21.51	11.00	19.61	24.74	1.79	2.04	1.64
	ERACE [ICLR'22]	48.41	55.97	63.21	16.96	23.79	27.27	23.35	27.23	34.55	13.72	18.85	23.90
	GSA [CVPR'23]	48.51	62.12	65.89	15.43	22.16	29.90	21.38	27.93	37.18	12.36	15.24	22.08
	OCM [ICML'22]	55.58	68.46	71.26	21.55	29.30	36.70	17.63	19.58	27.85	15.49	19.63	27.85
	PCR [CVPR'23]	50.85	62.37	66.31	15.08	23.39	31.07	17.34	19.67	31.46	11.39	12.68	20.41
	ER + MKD [ICML'24]	53.15	68.49	**74.82**	21.29	29.15	37.5	28.28	36.32	43.24	18.60	23.05	31.84
	S6MOD [CVPR'25]	50.22	63.74	68.26	13.17	17.65	26.50	26.35	34.65	38.44	9.87	10.94	19.67
	Ours	**56.77**	**70.43**	74.69	**22.89**	**31.63**	**39.87**	**27.83**	**37.28**	**43.30**	**19.77**	**24.43**	**32.73**
Blurry	ER[NeurIPS'19]	38.28	52.02	63.47	11.45	18.45	24.38	15.69	14.74	15.00	6.73	11.16	16.67
	ER + SDP [ICML'23]	49.43	59.48	65.18	13.64	19.43	27.46	13.48	14.26	5.14	11.10	15.98	23.25
	DER++[NeurIPS'20]	47.02	51.00	62.55	11.70	19.34	25.25	4.47	4.84	5.75	11.33	14.95	20.81
	DVC [CVPR'22]	51.37	57.40	62.87	12.28	16.30	22.30	2.97	2.46	1.40	8.30	11.96	16.15
	ERACE [ICLR'22]	48.41	55.97	63.21	16.96	23.79	27.27	11.79	18.23	29.44	13.72	18.85	23.90
	GSA [CVPR'23]	18.97	19.13	19.26	7.19	7.80	7.46	1.00	1.00	1.00	2.17	2.41	2.49
	OCM [ICML'22]	42.79	46.65	50.24	9.37	16.28	24.44	20.88	26.61	**36.36**	13.85	18.57	26.82
	PCR [CVPR'23]	53.88	60.14	69.94	18.16	22.25	31.28	8.57	15.64	31.47	10.30	14.96	23.57
	ER + MKD [ICML'24]	55.35	67.07	74.35	19.60	30.35	38.91	23.24	29.84	33.83	18.07	24.24	31.80
	S6MOD [CVPR'25]	48.68	57.88	65.24	19.28	21.96	25.50	20.44	23.87	27.52	7.56	10.99	20.36
	Ours	**56.17**	**69.49**	**74.43**	**22.84**	**31.97**	**40.02**	**23.70**	**30.37**	35.92	**20.34**	**25.21**	**34.51**

Table 1: Performance (%) under different memory sizes across datasets under **clear** and **blurry** task boundaries. Best is highlighted in **bold**, second best is underlined.

Phase 2: Outer flattening update. We temporarily add δ to the main network, $\theta_t^+ = \theta_t + \delta$, and evaluate the spatial loss

$$\mathcal{L}_{\text{spatial}}(\theta_t^+) = \frac{\lambda}{\sum_i m_i} \sum_i m_i \, \mathrm{KL}\big(q_{\theta_t}(x_i^*) \,\big\|\, q_{\theta_t^+}(x_i^*)\big) \tag{13}$$

which penalizes the divergence between the original predictions q_{θ_t} and the adversarially perturbed predictions $q_{\theta_t^+}$. We backpropagate $\nabla_{\theta_t^+}\mathcal{L}_{\text{spatial}}$ together with the standard continual–learning loss $\mathcal{L}_{\text{CL}}$ on (X_t, Y_t), then restore the parameters ($\theta_t^+ \mapsto \theta_t$) and perform the SGD update

$$\theta_{t+1} = \theta_t - \alpha\Big(\nabla_{\theta_t^+}\mathcal{L}_{\text{spatial}}(\theta_t^+) + \nabla_{\theta_t}\mathcal{L}_{\text{CL}}(\theta_t)\Big).$$

Using the gradient at the adversarial point θ_t^+ (while updating the original θ_t) implements a first–order min–max step analogous to sharpness–aware methods: it enforces invariance along the most fragile direction, reduces local gradient norm and curvature, enlarges fragility radii, and thereby diminishes the (ε, n)-fragile set.

Temporal Defense via Exponential Moving Average Online updates introduce high-frequency gradient noise that causes the parameter trajectory to oscillate around sharp ridges; even when the local surface is flattened, such oscillations can still flip the prediction of marginally learned samples. To damp this *temporal volatility*, we maintain a set of *slow weights* $\bar{\theta}_t$ obtained through an exponential moving average of the instantaneous parameters θ_t:

$$\bar{\theta}_t = \alpha\,\theta_t + (1-\alpha)\,\bar{\theta}_{t-1}, \qquad \bar{\theta}_0 = \theta_0,\ \alpha \in (0,1]. \tag{14}$$

Equation (14) acts as a first-order low-pass filter, suppressing fluctuations whose temporal frequency exceeds the EMA bandwidth; smaller α yields stronger smoothing and thus greater *memory inertia*.

We exploit $\bar{\theta}_t$ as a self-ensemble teacher and penalize the drift of logits on replay samples, which are most vulnerable to forgetting:

$$\mathcal{L}_{\text{temporal}} = \frac{1}{|\mathcal{B}_{\mathcal{M}}|} \sum_{(x,y)\in\mathcal{B}_{\mathcal{M}}} \Big\| m\big(f_{\theta_t}(x)\big) - m\big(f_{\bar{\theta}_{t-1}}(x)\big) \Big\|_2^2. \tag{15}$$

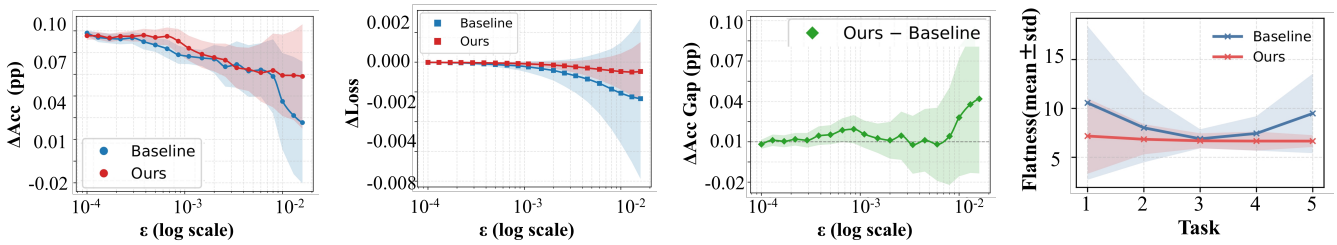
Adding this term yields the overall objective

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{CE}} + \mathcal{L}_{\text{Replay}} + \lambda_c\,\mathcal{L}_{\text{temporal}} + \lambda_p\,\mathcal{L}_{\text{spatial}}, \tag{16}$$

EMA incurs only one extra vector operation per step and no additional gradients; memory overhead is negligible because $\bar{\theta}_t$ is updated in place. Spatial flattening reduces the curvature term $\|\nabla_\theta^2\mathcal{L}\|$, while EMA minimizes the parameter-shift term $\|\theta_t - \bar{\theta}_{t-1}\|$. Jointly, the two defenses tighten the bound on margin change $\Delta\gamma(x,y) \le \|\nabla_\theta\gamma_\theta\|\,\|\theta_t - \bar{\theta}_{t-1}\|$, Together, they contract both the width and height of fragile zones, offering a holistic defense. A detailed pseudo-code of our training algorithm can be found in Appendix.

4 Experiments

Experimental Settings To evaluate the effectiveness of our method, we conducted experiments in both clear boundary setting and blurry boundary setting scenarios on 4 widely used benchmarks: CIFAR-10, CIFAR-100 (Krizhevsky and Hinton 2009), ImageNet-Subset (Hou et al. 2019) and Tiny-ImageNet (Le and Yang 2015). We follow a standard OCL setting with a streaming batch size of 10 and data retrieval from memory is capped at 64 with reservoir sampling (Vitter 1985) for memory management. The memory buffer size



(a) Accuracy Drop ΔAcc-ε (b) Loss Change $\Delta Loss$-ε (c) ΔAcc Gap (d) Flatness Metrics

Figure 5: **Empirical Evaluation of Perturbation Robustness and Flatness.** We assess the sensitivity of trained models to small parameter perturbations $\theta' = \theta + \varepsilon d$ by injecting normalized filter-wise noise and evaluating accuracy/loss changes. **(a)** *Accuracy Drop* ΔAcc under increasing perturbation magnitudes ε; **(b)** *Loss Change* shows that our method yields flatter optima with slower degradation; **(c)** *Accuracy Gap* between Ours and Baseline indicates consistent advantage under perturbations;**(d)** *Flatness Metric* (range-to-mean ratio) across tasks quantifies solution sharpness, with lower values indicating better flatness.Overall, our method exhibits greater robustness and flatter loss landscapes across various metrics.

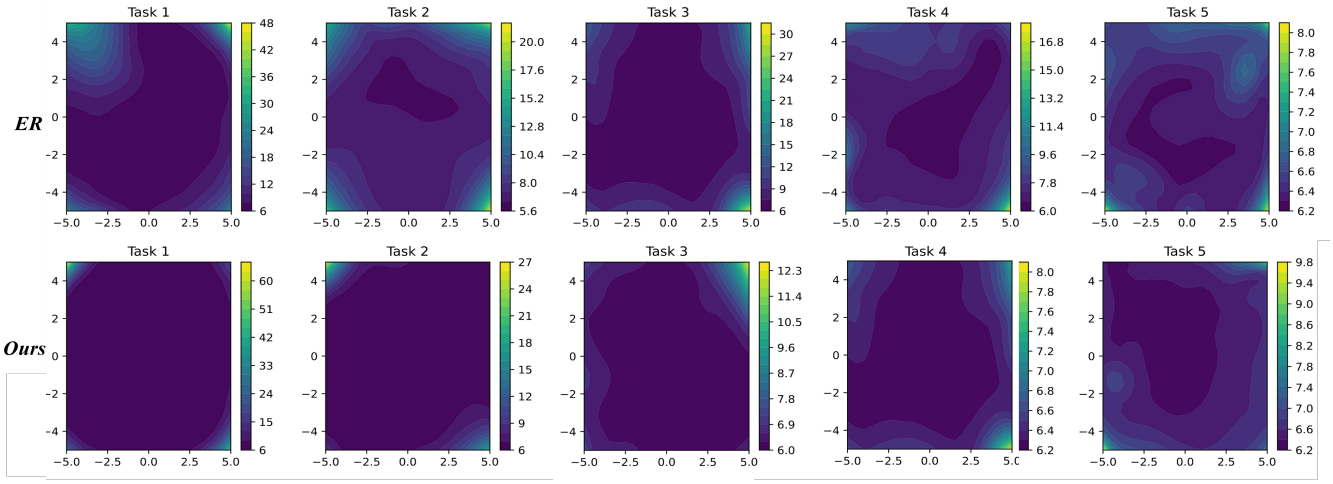


Figure 6: Contour lines of the cross entropy loss values, on the set of tasks seen so far (darker is lower).

M is varied from small (e.g., 200) to large (e.g., 5000) to assess robustness under different memory constraints. We use the accuracy averaged across all tasks after training on the last task to compare the methods under consideration. This metric is commonly known as the final average accuracy (Hsu et al. 2018). All baselines are trained with identical data augmentation strategies, including random horizontal flip, grayscale, color jitter, and random crop. Our proposed method is built upon MKD (Michel et al. 2024) and introduces only two additional hyperparameters: the perturbation weight λ_p and the perturbation radius γ. All models are trained using a single-pass over the data stream.

Evaluation under Clear and Blurry Task Boundaries Table 1 report final accuracies across datasets and memory budgets under **clear** and **blurry** task boundaries. In the clear setting, where task delineations are known, our method consistently achieves the best or second-best performance across all configurations, showing strong compatibility with standard incremental learning. The blurry setting removes task boundary information, simulating real-world non-stationarity via HalfNormal-based class mixing (Caccia et al. 2021). Many strong baselines degrade sharply, GSA collapses due to boundary reliance, and OCM suffers from unstable boundary estimation. In contrast, our method retains top performance across most blurry cases. This demonstrates that our approach is **boundary-agnostic**, maintaining stability and generalization without explicit task signals, making it well-suited for realistic online continual learning.

Loss landscape visualization We probe the geometry of solutions using the 2D loss–surface projection of Li et al. (2018). Given parameters θ after task t, we evaluate cross-entropy on the *cumulative* test set under filter-wise normalized random directions $d_1, d_2 \sim \mathcal{N}(0, I)$:

$$\theta' = \theta + \alpha d_1 + \beta d_2, \quad (\alpha, \beta) \in [-r, r]^2, \tag{17}$$

and render the surface $\{\mathcal{L}_{\alpha,\beta}\}$ as contours over a uniform grid. As shown in Figure 6, ER concentrates around a sharp minimum with a narrow low-loss basin, which implies high curvature and sensitivity to small parameter perturbations and can amplify interference in subsequent online updates. In contrast, PDFK consistently yields broader and smoother valleys that tolerate update noise and help preserve prior knowledge. To summarize these geometries we report simple flatness statistics over the grid. Figure 5d shows that PDFK reduces the reported flatness measures across tasks,

indicating flatter optima and lower perturbation sensitivity and leaving more headroom for future adaptation.

Sensitivity Analysis in Parameter Space To assess the local stability of model solutions in parameter space, we introduce controlled perturbations of the form:

$$\theta' = \theta + \varepsilon \cdot d, \tag{18}$$

where d is a filter-wise normalized random direction and $\varepsilon \in \{10^{-4}, 1.4\times10^{-4}, \ldots, 2.3\times10^{-3}\}$ controls the perturbation magnitude. For each perturbed model θ', we measure the accuracy degradation $\Delta\mathrm{Acc}(\varepsilon)$ and loss increase $\Delta\mathrm{Loss}(\varepsilon)$ relative to the unperturbed model. As shown in Figure 5, we observe that our method exhibits significantly flatter loss landscapes and stronger robustness to small perturbations. Specifically, Figure 5a accuracy drops more slowly, and Figure 5b loss increases less steeply compared to the baseline. To quantify these trends, we report two metrics (Li et al. 2018) in Table 2: **Slope**, capturing the initial degradation rate (sharpness), and **AUC**, measuring cumulative degradation (global sensitivity). We further report the performance gap $\Delta\mathrm{Acc}_{\mathrm{Ours}} - \Delta\mathrm{Acc}_{\mathrm{Baseline}}$ in Figure 5c, where positive values indicate superior robustness. Our method maintains a consistently positive gap across a wide range of ε, with larger benefits emerging under stronger perturbations. This empirically supports that our method not only finds flatter optima, but also generalizes better under parameter shifts.

	ΔAcc		ΔLoss	
	AUC↓	**Slope$_0$ ↓**	**AUC↓**	**Slope$_0$ ↓**
MKD	0.05101±0.0052	6.89±0.18	0.001601±0.0004	0.309±0.07
Ours	**0.03500±0.0044**	**2.97±0.09**	**0.0004845±0.0002**	**0.101±0.05**

Table 2: Sensitivity metrics under parameter perturbations with 95% confidence intervals (CI).

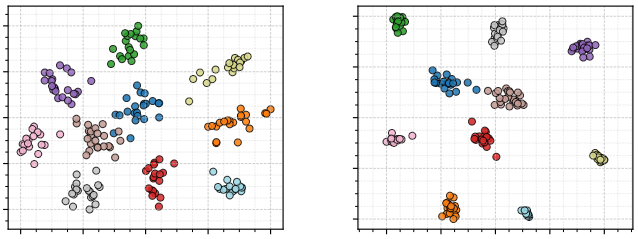
Feature Visualization To qualitatively evaluate the discriminative power of the learned representations, we visualize the features using t-SNE (Maaten and Hinton 2008) on the CIFAR-10 stream. As shown in Figure 7, each point represents a sample and is colored by its corresponding class label. Compared to the ER(Figure 7a), our method (Figure 7b) yields more compact and well-separated clusters. This suggests that the proposed distillation mechanism helps preserve class-specific structure and reduces feature confusion between old and new classes.

Ablation Studies

Module-level analysis. We separately evaluate the effect of temporal perturbation (via EMA teacher), spatial perturbation (via structured noise injection), and their combination. Results in Table 3 show that each module improves performance individually, while their combination yields the best stability–plasticity trade-off.

Perturbation strategy analysis. We analyze the effectiveness of two key components in our perturbation design:

- x vs. x^*: Selecting only correctly classified buffer samples x^* versus using all buffer samples x for perturbation in our loss formulation.



(a) Baseline (ER) (b) Ours

Figure 7: t-SNE visualization of the learned feature representations on the CIFAR-10 stream. Compared to the baseline ER, our method produces more compact and discriminative clusters.

Method / Setting	**Acc. ↑**	**AF ↓**
Module-Level Ablation		
ER	25.12±0.93	12.42±1.39
ER + EMA (Temporal)	37.50±0.87	9.65±1.21
ER + Perturbation (Spatial)	35.45±0.78	10.04±1.17
Ours (Full)	**39.87±0.81**	**8.91±1.32**
Perturbation Strategy		
w/o ($(x, \mathcal{N})$)	36.61±1.04	10.28±1.59
$(x, \mathcal{N})$	37.42±0.88	9.97±1.38
$(x, \mathcal{P})$	38.65±0.91	9.14±1.22
$(x^*, \mathcal{P})$	**39.87±0.81**	**8.91±1.32**

Table 3: Ablation study on CIFAR-100 (M=1k), evaluating both module-level and perturbation strategy design. Accuracy (Acc.↑) and Average Forgetting (AF↓) are reported with 95% confidence intervals though 3 runs. $\mathcal{N}$ = Gaussian noise; $\mathcal{P}$ = maximally allowable perturbation; x^* = correctly classified samples only.

- $\mathcal{P}$ vs. $\mathcal{N}$: Applying the maximally allowable perturbation $\mathcal{M}$, as defined in Eq. 13, versus injecting Gaussian noise $\mathcal{N} \sim \mathcal{N}(0, \sigma^2\mathbf{I})$ of the same magnitude.

We denote each strategy as a pair $(x^\dagger, \delta^\dagger)$, where $x^\dagger \in \{x, x^*\}$ indicates the sample selection criterion, and $\delta^\dagger \in \{\mathcal{N}, \mathcal{P}\}$ represents the perturbation type. In Table 3, both guided sample selection and constrained perturbation significantly contribute to accuracy and forgetting reduction.

5 Conclusion

To mitigate knowledge fragility, we propose **PDFK**, a task-agnostic framework that stabilizes both time and space: (i) EMA smoothing suppresses temporal volatility, while (ii) targeted perturbations with consistency regularization flatten sharp loss regions.

6 Acknowledgments

This work is supported by National Science and Technology Major Project (2023ZD0121101), National University of Defense Technology (ZZCX-ZZGC-01-04) and Major Fundamental Research Project of Hunan Province (2025JC0005).

References

Buzzega, P.; Boschini, M.; Porrello, A.; Abati, D.; and Calderara, S. 2020. Dark experience for general continual learning: a strong, simple baseline. *Advances in neural information processing systems*, 33: 15920–15930.

Caccia, L.; Aljundi, R.; Asadi, N.; Tuytelaars, T.; Pineau, J.; and Belilovsky, E. 2021. New insights on reducing abrupt representation change in online continual learning. *arXiv preprint arXiv:2104.05025*.

Chaudhry, A.; Ranzato, M.; Rohrbach, M.; and Elhoseiny, M. 2018. Efficient lifelong learning with a-gem. *arXiv preprint arXiv:1812.00420*.

Chaudhry, A.; Rohrbach, M.; Elhoseiny, M.; Ajanthan, T.; and Torr, P. H. 2019. Tiny episodic memories in continual learning. In *International Conference on Machine Learning (ICML)*, 1954–1963.

De Lange, M.; Aljundi, R.; Masana, M.; Parisot, S.; Jia, X.; Leonardis, A.; Slabaugh, G.; and Tuytelaars, T. 2021. A continual learning survey: Defying forgetting in classification tasks. *IEEE transactions on pattern analysis and machine intelligence*, 44(7): 3366–3385.

Deng, Y.; Yang, Y.; Yao, Y.; and Hospedales, T. M. 2023. Momentum knowledge distillation for online class-incremental learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 16195–16204.

Feng, Y.; Liang, W.; Wan, X.; Liu, J.; Li, M.; and Liu, X. 2025a. Incremental Multi-View Clustering: Exploring Stream-View Correlations to Learn Consistency and Diversity. *IEEE Transactions on Knowledge and Data Engineering*.

Feng, Y.; Liang, W.; Wan, X.; Liu, J.; Liu, S.; Qu, Q.; Guan, R.; Xu, H.; and Liu, X. 2025b. Incremental Nyström-based Multiple Kernel Clustering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 16613–16621.

French, R. M. 1999. Catastrophic forgetting in connectionist networks. *Trends in cognitive sciences*, 3(4): 128–135.

Gao, Z.; Han, S.; Zhang, X.; Xu, K.; Zhou, D.; Mao, X.; Dou, Y.; and Wang, H. 2025a. Maintaining Fairness in Logit-based Knowledge Distillation for Class-Incremental Learning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(16): 16763–16771.

Gao, Z.; Jia, W.; Zhang, X.; Zhou, D.; Xu, K.; Dawei, F.; Dou, Y.; Mao, X.; and Wang, H. 2025b. Knowledge memorization and rumination for pre-trained model-based class-incremental learning. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 20523–20533.

Gao, Z.; Xu, K.; Zhang, X.; Zhuang, H.; Wan, T.; Ding, B.; Mao, X.; and Huaimin, W. 2025c. Rethinking Obscured Sub-optimality in Analytic Learning for Exemplar-Free Class-Incremental Learning. *IEEE Transactions on Circuits and Systems for Video Technology*, 36(10): 1123–1136.

Gao, Z.; Xu, K.; Zhuang, H.; Liu, L.; Mao, X.; Ding, B.; Feng, D.; and Wang, H. 2024a. Less confidence, less forgetting: Learning with a humbler teacher in exemplar-free Class-Incremental learning. *Neural Networks*, 179: 106513.

Gao, Z.; Zhang, X.; Xu, K.; Mao, X.; and Wang, H. 2024b. Stabilizing Zero-Shot Prediction: A Novel Antidote to Forgetting in Continual Vision-Language Tasks. In Globerson, A.; Mackey, L.; Belgrave, D.; Fan, A.; Paquet, U.; Tomczak, J.; and Zhang, C., eds., *Advances in Neural Information Processing Systems*, volume 37, 128462–128488. Curran Associates, Inc.

Guo, Y.; Liu, B.; and Zhao, D. 2022. Online continual learning through mutual information maximization. In *International conference on machine learning*, 8109–8126. PMLR.

Guo, Y.; Liu, B.; and Zhao, D. 2023. Dealing with cross-task class discrimination in online continual learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 11878–11887.

Hou, S.; Pan, X.; Loy, C. C.; Wang, Z.; and Lin, D. 2019. Learning a Unified Classifier Incrementally via Rebalancing. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 831–839.

Hsu, Y.-C.; Liu, Y.-C.; Ramasamy, A.; and Kira, Z. 2018. Re-evaluating continual learning scenarios: A categorization and case for strong baselines. *arXiv preprint arXiv:1810.12488*.

Jung, D.; Lee, D.; Hong, S.; Jang, H.; Bae, H.; and Yoon, S. 2023. New insights for the stability-plasticity dilemma in online continual learning. *arXiv preprint arXiv:2302.08741*.

Kirkpatrick, J.; Pascanu, R.; Rabinowitz, N.; Veness, J.; Desjardins, G.; Rusu, A. A.; Milan, K.; Quan, J.; Ramalho, T.; Grabska-Barwinska, A.; et al. 2017. Overcoming catastrophic forgetting in neural networks. *Proceedings of the national academy of sciences*, 114(13): 3521–3526.

Konishi, T.; Kurokawa, M.; Ono, C.; Ke, Z.; Kim, G.; and Liu, B. 2023. Parameter-level soft-masking for continual learning. In *International conference on machine learning*, 17492–17505. PMLR.

Krizhevsky, A.; and Hinton, G. 2009. Learning multiple layers of features from tiny images. *Handbook of Systemic Autoimmune Diseases*, 1(4).

Le, Y.; and Yang, X. 2015. Tiny imagenet visual recognition challenge. *CS 231N*, 7(7): 3.

Li, H.; Xu, Z.; Taylor, G.; Studer, C.; and Goldstein, T. 2018. Visualizing the loss landscape of neural nets. *Advances in neural information processing systems*, 31.

Li, Z.; and Hoiem, D. 2017. Learning without forgetting. *IEEE transactions on pattern analysis and machine intelligence*, 40(12): 2935–2947.

Liu, S.; Yang, Y.; Li, X.; Clifton, D. A.; and Ghanem, B. 2025. Enhancing online continual learning with plug-and-play state space model and class-conditional mixture of discretization. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 20502–20511.

Lopez-Paz, D.; and Ranzato, M. 2017. Gradient episodic memory for continual learning. *Advances in neural information processing systems*, 30.

Maaten, L. v. d.; and Hinton, G. 2008. Visualizing data using t-SNE. *Journal of machine learning research*, 9(Nov): 2579–2605.

Michel, N.; Wang, M.; Xiao, L.; and Yamasaki, T. 2024. Rethinking Momentum Knowledge Distillation in Online Continual Learning. In Salakhutdinov, R.; Kolter, Z.; Heller, K.; Weller, A.; Oliver, N.; Scarlett, J.; and Berkenkamp, F., eds., *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, 35607–35622. PMLR.

Riemer, M.; Cases, I.; Ajemian, R.; Liu, M.; Rish, I.; Tu, Y.; and Tesauro, G. 2019. Learning to learn without forgetting by maximizing transfer and minimizing interference. In *International Conference on Learning Representations (ICLR)*.

Shim, D.; Mai, Z.; Jeong, J.; Sanner, S.; Kim, H.; and Jang, J. 2021. Online class-incremental continual learning with adversarial shapley value. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, 9630–9638.

Soutif-Cormerais, A.; Carta, A.; Cossu, A.; Hurtado, J.; Lomonaco, V.; Van de Weijer, J.; and Hemati, H. 2023. A comprehensive empirical evaluation on online continual learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 3518–3528.

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

Vitter, J. S. 1985. Random sampling with a reservoir. *ACM Transactions on Mathematical Software (TOMS)*, 11(1): 37–57.

Wan, H.; Ren, S.; Huang, W.; Zhang, M.; Deng, X.; Bao, Y.; and Nie, L. 2025. Understanding the Forgetting of (Replay-based) Continual Learning via Feature Learning: Angle Matters. In *Forty-second International Conference on Machine Learning*.

Wang, L.; Zhang, X.; Su, H.; and Zhu, J. 2024a. A comprehensive survey of continual learning: Theory, method and application. *IEEE transactions on pattern analysis and machine intelligence*, 46(8): 5362–5383.

Wang, Z.; Yang, E.; Shen, L.; and Huang, H. 2024b. A comprehensive survey of forgetting in deep learning beyond continual learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Wei, Y.; Ye, J.; Huang, Z.; Zhang, J.; and Shan, H. 2023. Online prototype learning for online continual learning. In *Proceedings of the IEEE/CVF international conference on computer vision*, 18764–18774.

Yan, H.; Wang, L.; Ma, K.; and Zhong, Y. 2024. Orchestrate latent expertise: Advancing online continual learning with multi-level supervision and reverse self-distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 23670–23680.

Zhu, F.; Cheng, Z.; Zhang, X.-y.; and Liu, C.-l. 2021a. Class-incremental learning via dual augmentation. *Advances in neural information processing systems*, 34: 14306–14318.

Zhu, F.; Zhang, X.-Y.; Wang, C.; Yin, F.; and Liu, C.-L. 2021b. Prototype augmentation and self-supervision for incremental learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 5871–5880.