BEYOND EXPERIENCE REPLAY: LEARNING STRUCTURAL MEMORY FOR ONLINE CONTINUAL LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Experience replay is a widely adopted strategy for online continual learning (OCL), yet conventional replay memories mainly preserve historical observations and do not explicitly retain how previously learned concepts are organized. We show that continual updates can alter the relational organization of historical representations, leading to structural drift even when replay is employed. Motivated by this observation, we propose **Structural Memory Consolidation (SMC)**, a memory-efficient framework that augments episodic replay with an additional structural memory state. SMC constructs compact structural snapshots from representative anchors at task transitions and consolidates these historical relational states during subsequent online adaptation. By preserving the organization among learned concepts rather than directly storing high-dimensional representations or constraining individual features, SMC enables more stable knowledge evolution while maintaining model plasticity. Extensive experiments on multiple online Class-IL benchmarks demonstrate that SMC consistently improves knowledge retention and representation stability across diverse memory budgets, with particularly strong benefits under memory-constrained settings.

1 INTRODUCTION

Online continual learning (OCL) (Lopez-Paz & Ranzato, 2017; Rebuffi et al., 2017; Soutif-Cormerais et al., 2023; Jung et al., 2023) studies the problem of continuously adapting neural networks from non-stationary data streams under limited memory and computation budgets. Unlike conventional batch learning, OCL requires models to process each instance in a single pass while retaining knowledge from previously observed data. Among existing approaches, experience replay (ER) (Chaudhry et al., 2019) has become a widely adopted paradigm by maintaining a bounded buffer of historical samples to approximate joint training. However, this sample-centric memory raises a fundamental question: beyond storing past observations, what information should be preserved to maintain previously acquired knowledge?

Existing replay-based methods have improved continual learning from different perspectives, including memory construction, exemplar selection, and knowledge preservation. Early approaches maintain episodic memories of historical examples to mitigate catastrophic forgetting (Lopez-Paz & Ranzato, 2017; Rolnick et al., 2019), while subsequent methods improve replay quality through representative exemplar selection and importance-aware retrieval (Rebuffi et al., 2017; Aljundi et al., 2019). More recent approaches extend memory beyond raw observations by preserving additional information, such as historical logits, feature representations, and higher-order statistics (Buzzega et al., 2020b; Douillard et al., 2020; Guo et al., 2022). Beyond point-wise information, relational objectives have also been explored to preserve relationships among representations. Relational knowledge distillation demonstrates that inter-sample relationships provide complementary information beyond individual predictions (Park et al., 2019), while Co^{2}L (Cha et al., 2021) and prototype-sample relation distillation (Asadi et al., 2023) further show the benefits of relational constraints for knowledge retention. However, these methods mainly utilize relational information as an optimization objective during training, where relationships are computed to constrain current model updates. The historical relational organization itself is not explicitly stored as a persistent memory state for future adaptation. This motivates a broader question: can continual memory preserve not only which observations were encountered, but also how previously learned concepts are organized?

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

We argue that continual knowledge is encoded not only in individual examples, but also in the structural organization among learned concepts. During online adaptation, replayed samples provide historical supervision but do not explicitly specify which relational properties among previously learned concepts should remain stable. As illustrated in Figure 1, continual updates can alter the organization of historical representations even when replay is employed. Such structural changes differ from simple feature displacement: representations may undergo global transformations while preserving semantic relationships, whereas changes in relational organization can modify the neighborhood structure among previously learned concepts. Therefore, preserving continual knowledge requires maintaining not only sample-level observations or feature-level information, but also compact descriptions of historical relational organization.

Motivated by this observation, we introduce *structural memory* as an additional memory representation for online continual learning. Unlike episodic memory, which stores historical samples for rehearsal, structural memory explicitly stores compact relational states describing how previously acquired concepts are organized. Unlike relational objectives that transiently constrain optimization, structural memory treats relational organization as a persistent state that can be maintained and reused across learning stages. This perspective extends the role of continual memory from preserving historical observations toward preserving the organization of learned knowledge.

Based on this perspective, we propose **Structural Memory Consolidation (SMC)**, a memory-efficient framework that augments experience replay with structural memory. At task transitions, SMC constructs compact structural snapshots from representative anchors stored in episodic memory, capturing cross-task relational organization without retaining high-dimensional historical features or additional model checkpoints. During subsequent online updates, these structural snapshots provide relational references that guide representation adaptation while preserving previously learned organization. In this way, SMC enables a continual learner to maintain structural knowledge beyond sample-level replay while retaining the flexibility required for learning new concepts.

Our contributions are summarized as follows:

- We introduce **structural memory** as a new memory representation for OCL, which explicitly stores historical relational organization as a persistent memory state rather than using relational information only as an optimization constraint.
- We propose **SMC**, which constructs compact structural snapshots at task transitions and consolidates historical relational states during subsequent online updates.
- We design lightweight structural snapshot construction strategies based on representative anchors, enabling structural information preservation without storing historical feature embeddings or additional model checkpoints.
- Extensive experiments on multiple online Class-IL benchmarks demonstrate that SMC improves knowledge retention and representation stability across different memory budgets, with consistent benefits under limited replay capacity.

2 Related Work

2.1 Continual Learning

Continual Learning (CL) aims to acquire knowledge from sequential data while retaining previously learned capabilities, commonly framed through the stability-plasticity dilemma (Wang et al., 2024a;b; De Lange et al., 2021). Existing approaches mainly include regularization-based methods that constrain changes to previously learned knowledge (Li & Hoiem, 2017; Gao et al., 2024; 2025), parameter isolation methods that reduce interference through task-specific parameter allocation (Konishi et al., 2023), and experience replay methods that retain historical observations for rehearsal (Lopez-Paz & Ranzato, 2017; Buzzega et al., 2020a; Chaudhry et al., 2019). Replay is particularly suitable for online settings, where historical data access is restricted and adaptation must proceed from a non-stationary stream (Shim et al., 2021). However, conventional episodic memory primarily retains observations rather than explicitly encoding the relational organization acquired from them.

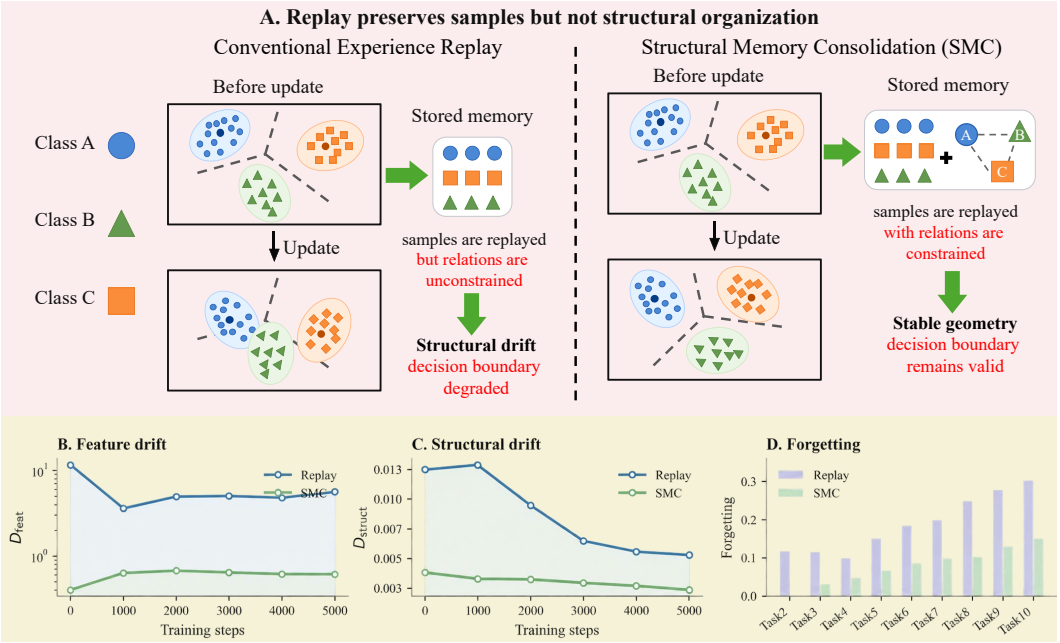


Figure 1: From experience replay to structural memory. Replay preserves historical observations, but continual updates can still change the relational organization among learned concepts. SMC augments episodic memory with compact structural snapshots, explicitly preserving historical relational organization and providing additional guidance for future representation updates.

2.2 Online Continual Learning

Online Continual Learning (OCL) further restricts learning to single-pass data streams under limited memory and computation budgets (Soutif-Cormerais et al., 2023; Liu et al., 2025). Existing replay-based methods improve continual adaptation through sample selection, representation learning, and additional supervision. OCM employs mutual-information-based objectives to improve representation learning, while GSA (Guo et al., 2023) exploits gradient-sensitive replay and OnPro (Wei et al., 2023) maintains online prototypes. More recent approaches, such as MOSE (Yan et al., 2024) and S6MOD (Liu et al., 2025), enhance online learning through multi-level supervision or specialized architectures. These methods improve how historical experience is selected, represented, or utilized during training. Our work investigates a complementary question: how can a bounded memory explicitly retain the relational organization of previously learned representations and use it to guide subsequent online adaptation?

2.3 Memory Representations for Knowledge Preservation

Beyond raw replay samples, continual learning methods have explored preserving richer forms of historical information, including logits (Buzzega et al., 2020a; Michel et al., 2023), feature representations (Li & Hoiem, 2017), class prototypes (Yu et al., 2020; Goswami et al., 2024), and higher-order representation statistics (Wu et al., 2025). Recent studies further investigate representation evolution and structural preservation in continual learning (Kim et al., 2025; Zang et al., 2026), highlighting the importance of retaining knowledge beyond individual observations.

Relational knowledge distillation provides another perspective by preserving relationships among representations rather than matching individual outputs (Park et al., 2019). In continual learning, Co^{2}L (Cha et al., 2021) preserves instance-level similarities through relational objectives, while prototype-sample relation distillation maintains prototype-sample relationships without requiring historical replay (Asadi et al., 2023). These approaches demonstrate the value of relational information for mitigating forgetting. However, relational information is typically introduced as an optimization constraint, where relationships guide current updates rather than being explicitly maintained as a memory component. Table 1 summarizes the distinction between different forms of continual memory representation. Building on this perspective, SMC explicitly stores compact relational snapshots as structural memory and reuses them during subsequent online adaptation.

Table 1: Comparison of different continual memory representations. The comparison focuses on the type of information explicitly maintained for future continual updates.

Method	Stored Information	Training Usage	Explicit Relational State
ER	Historical samples	Replay rehearsal	No
DER++	Samples + logits	Logit distillation	No
Prototype-based CL	Class statistics	Prototype matching	No
Relational KD	Representation relations	Relational loss	No
Co^2L	Instance similarities	Contrastive learning	No
SMC	Structural snapshots	Memory consolidation	Yes

3 METHODOLOGY

3.1 PROBLEM SETUP AND MEMORY STATE

We consider online continual learning in the class-incremental learning (Class-IL) setting. A learner sequentially encounters tasks $\{\mathcal{T}_1, \ldots, \mathcal{T}_T\}$, where each task $\mathcal{T}_t$ provides a data stream $\mathcal{S}_t = \{(x_{t,i}, y_{t,i})\}_{i=1}^{N_t}$, with $x_{t,i} \in \mathcal{X}$ and $y_{t,i} \in C_t$. The class sets are disjoint across tasks, i.e., $C_t \cap C_{t'} = \varnothing$ for $t \neq t'$, and inference is performed over all classes observed so far without task identities. Each incoming sample is encountered only once, while historical data are accessible through a bounded replay memory $\mathcal{M}_t$, with $|\mathcal{M}_t| \leq B$.

The model consists of a feature extractor g_ϕ and a classifier h_θ, with representation $z(x;\phi) = g_\phi(x)$ and prediction logits $\ell(x;\phi,\theta) = h_\theta(z(x;\phi))$. Conventional replay maintains historical observations through episodic memory, but does not explicitly represent the relational organization induced by these observations. We therefore extend the memory state as

$$\mathcal{H}_t = (\mathcal{M}_t, \mathcal{P}_t), \tag{1}$$

where $\mathcal{M}_t$ denotes episodic memory for sample-level rehearsal, and $\mathcal{P}_t$ denotes structural memory that stores compact summaries of historical relational organization. Unlike episodic memory, which provides historical samples for replay, $\mathcal{P}_t$ maintains relational states that can be reused to guide future representation updates. Here, B specifies the episodic replay capacity. The structural memory P_t is an auxiliary relational state containing compact similarity snapshots, while episodic memory preserves historical observations, it does not explicitly maintain the relational organization among learned concepts. Motivated by the structural drift observed during online adaptation (Figure 3), we introduce structural snapshots as a persistent memory state to preserve historical relational organization.

3.2 STRUCTURAL SNAPSHOT CONSTRUCTION

Motivated by the observed structural drift, we propose *Structural Memory Consolidation* (SMC), which extends episodic replay with an explicit memory of historical relational organization. As illustrated in Figure 2, SMC consists of two complementary operations: structural snapshot construction and structural memory consolidation. At each task transition, representative anchors from episodic memory are used to construct compact relational snapshots, which subsequently serve as fixed references for preserving historical relationships during online adaptation.

Representative anchor selection. Before learning task $\mathcal{T}_t$, we select K representative anchors for each historical task $i < t$, assuming sufficient samples are available in the replay memory:

$$R_i = \{x_{i,1}, \ldots, x_{i,K}\} \subseteq \mathcal{M}_{t-1}. \tag{2}$$

We consider two lightweight selection strategies compatible with online learning. *Class-balanced sampling* distributes the anchor budget across historical classes to promote structural coverage, whereas *typical-center sampling* prioritizes samples close to their class centers in normalized representation space to reduce the influence of atypical observations. Both strategies operate on samples already available in episodic memory without revisiting the original historical data streams. The selected anchors are cached with a fixed ordering throughout the current task, ensuring consistent correspondence between stored and reconstructed relational distributions even as the replay memory is updated. Importantly, these anchors are not additional samples beyond the episodic memory budget; they are references to samples already stored in $\mathcal{M}_t$. SMC does not maintain an additional exemplar buffer or store their historical feature embeddings.

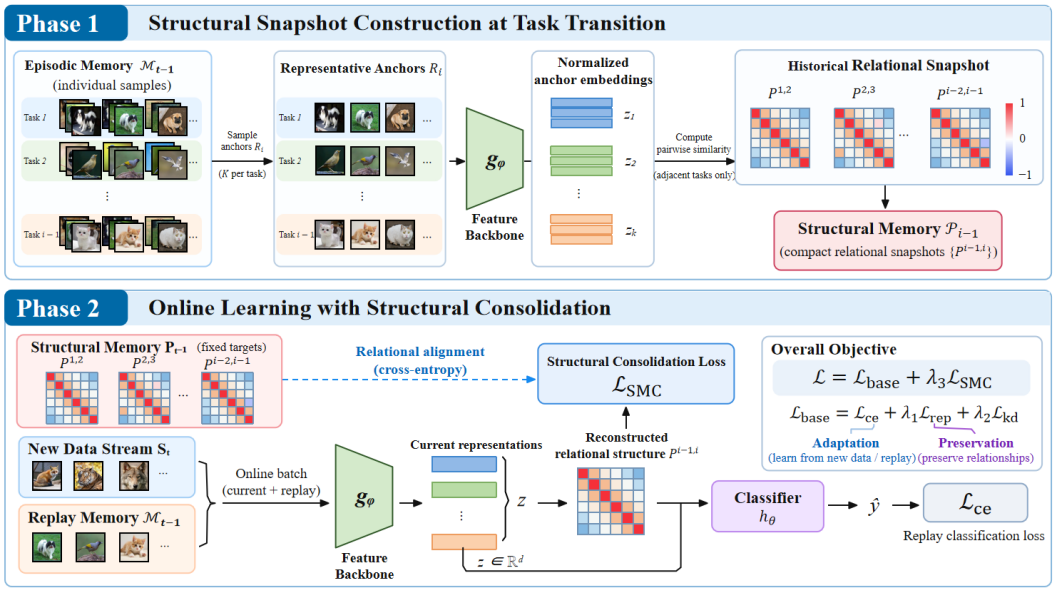


Figure 2: Overview of Structural Memory Consolidation (SMC). Representative anchors from episodic memory are used to construct historical relational snapshots, which are subsequently reconstructed and consolidated during online learning alongside sample-level replay.

Relational snapshot construction. Let $\bar{\phi}_{t-1}$ denote the encoder state at the transition to task $\mathcal{T}_t$. Given the selected anchors, we compute normalized representations $\widetilde{z}(x;\phi) = g_\phi(x)/\|g_\phi(x)\|_2$ and stack them into an embedding matrix $\widetilde{Z}_i(\phi) \in \mathbb{R}^{K\times d}$, where each row corresponds to an anchor from historical task i and d denotes the representation dimension.

Rather than storing historical feature vectors, SMC encodes their relative organization through pairwise similarities between anchors from adjacent historical tasks. This adjacent-pair construction captures cross-task relationships without requiring snapshots for every possible historical task pair. Specifically, the structural snapshot for each pair $(i-1, i)$ is constructed as

$$P_{t-1}^{i-1,i} = \text{softmax}_{row}\left(\frac{\widetilde{Z}_{i-1}(\bar{\phi}_{t-1})\widetilde{Z}_i(\bar{\phi}_{t-1})^\top}{\tau_{\text{mem}}}\right), \tag{3}$$

where $\tau_{\text{mem}} > 0$ controls the sharpness of the stored relational distribution. Each row of $P_{t-1}^{i-1,i}$ represents the relative affinities of one anchor from task $i-1$ to the anchors of task i.

The structural memory available during task $\mathcal{T}_t$ is defined as

$$\mathcal{P}_{t-1} = \left\{P_{t-1}^{i-1,i}\right\}_{i=2}^{t-1}. \tag{4}$$

At each task transition, all adjacent historical task-pair snapshots are refreshed using the current boundary encoder and newly selected representative anchors. The resulting structural memory remains fixed throughout the subsequent task. With K anchors per historical task, the snapshots require $\mathcal{O}((t-2)K^2)$ scalar values without retaining historical feature vectors or additional encoder checkpoints.

3.3 Structural Memory Consolidation

During online learning of task $\mathcal{T}_t$, the structural memory $\mathcal{P}_{t-1}$ remains fixed. At each online update, the current encoder reconstructs the relational state of each historical task pair using the same cached representative anchors:

$$\widehat{P}_t^{i-1,i}(\phi) = \text{softmax}_{row}\left(\frac{\widetilde{Z}_{i-1}(\phi)\widetilde{Z}_i(\phi)^\top}{\tau_{\text{cur}}}\right), \tag{5}$$

where $\tau_{\text{cur}} > 0$ controls the sharpness of the reconstructed relational distribution. Unlike the stored snapshot, the reconstructed state evolves with the current encoder as new observations are learned.

To preserve historical organization, we align the reconstructed relational distributions with their stored counterparts through the structural consolidation objective:

$$\mathcal{L}_{\mathrm{SMC}} = \sum_{i=2}^{t-1} \frac{1}{K} \sum_{p=1}^{K} \mathrm{CE}\left(P_{t-1}^{i-1,i}(p,:), \widehat{P}_{t}^{i-1,i}(p,:)\right), \tag{6}$$

where $\mathrm{CE}(\cdot,\cdot)$ denotes cross-entropy between the stored and reconstructed relational distributions. The stored snapshots are treated as fixed targets, and gradients are propagated only through the reconstructed relational states. The consolidation objective is inactive during the first two tasks and is applied from task $\mathcal{T}_3$ onward.

By aligning historical and current relational states, SMC constrains changes in the relative affinities among historical anchors without directly penalizing individual feature displacement. This allows representations to evolve during online adaptation while encouraging the preservation of previously learned relational organization.

3.4 Overall Learning Objective

SMC is incorporated into a standard replay-based objective to balance adaptation to incoming data and preservation of historical knowledge. Given an incoming mini-batch $\mathcal{B}_t$ and a replay mini-batch $\mathcal{B}_M$, the base learning objective is

$$\mathcal{L}_{\mathrm{base}} = \mathcal{L}_{\mathrm{ce}} + \lambda_1 \mathcal{L}_{\mathrm{rep}} + \lambda_2 \mathcal{L}_{\mathrm{kd}}, \tag{7}$$

where $\mathcal{L}_{\mathrm{ce}}$ provides cross-entropy supervision on incoming samples, $\mathcal{L}_{\mathrm{rep}}$ performs hard-label rehearsal on replay samples, and $\mathcal{L}_{\mathrm{kd}}$ preserves point-wise predictive information through cached soft targets when available. The overall objective is defined as

$$\mathcal{L} = \mathcal{L}_{\mathrm{base}} + \lambda_3 \mathcal{L}_{\mathrm{SMC}}, \tag{8}$$

where λ_1, λ_2, and λ_3 control the contributions of replay, knowledge distillation, and structural consolidation, respectively.

4 Experiments

4.1 Experimental Setup

Datasets and Metrics. We evaluate SMC on CIFAR-10 (10 classes), CIFAR-100 (100 classes) (Krizhevsky et al., 2009), Tiny-ImageNet (200 classes) (Krizhevsky et al., 2012), and ImageNet-Subset (ImageNet-100) (Hou et al., 2019). We follow the online Class-IL protocol: each training example is observed once, and inference is performed jointly over all seen classes without task identities. CIFAR-10 is split into five tasks with two classes each; CIFAR-100 into ten tasks with ten classes each; and Tiny-ImageNet and ImageNet-Subset into ten tasks with 20 and 10 classes, respectively. We report final average accuracy and average forgetting. The latter is computed as the mean, across tasks, of the difference between each task's best historical accuracy and its final accuracy.

Implementation Details. All methods use the same randomly initialized ResNet-18 backbone (He et al., 2016). Replay buffers are constructed with reservoir sampling (Vitter, 1985) and randomly sampled unless otherwise specified by the baseline. Hyperparameters are tuned on CIFAR-100 with a buffer size of $|\mathcal{M}| = 5$k and reused for all other datasets. Each update processes 10 incoming samples and retrieves at most 64 replay samples. We use random horizontal flips, grayscale conversion, color jitter, and random cropping for data augmentation. For blurry datasets, we follow the public implementations with blur scale 500; task-boundary inference is described in the appendix.

We evaluate two variants of SMC: class-balanced sampling (SMC-CB) and typical-center sampling (SMC-TC), both introduced in the structural snapshot construction procedure. The reported buffer size denotes the episodic replay capacity $|\mathcal{M}|$. Structural snapshots are stored separately in $\mathcal{P}$ as compact relational matrices. Their reconstruction anchors are selected from $\mathcal{M}$ and therefore require no additional exemplar storage.

Table 2: Final average accuracy (%) under different episodic memory budgets. Higher is better; bold indicates the highest reported value in each column. SMC-CB and SMC-TC denote Structural Memory Consolidation with class-balanced and typical-center representative selection, respectively.

Method	CIFAR-10			CIFAR-100			ImageNet-Subset			Tiny-ImageNet		
	200	500	1000	200	500	1000	1000	2000	5000	1000	2000	5000
ER	45.50	52.28	65.19	12.35	16.69	24.96	32.84	40.95	44.72	41.73	48.65	55.14
ER + SDP	49.05	57.90	67.79	14.24	20.38	28.31	34.24	42.57	46.45	43.98	50.74	57.82
DER++	50.81	56.74	69.78	15.16	19.80	27.88	36.14	44.30	48.30	45.62	52.31	58.79
DVC	49.72	60.18	70.35	15.90	21.22	29.99	35.01	42.66	46.53	44.01	50.26	56.38
ERACE	48.58	59.79	67.75	13.74	20.02	29.20	37.84	45.86	50.07	47.58	54.41	60.92
GSA	53.15	66.28	71.69	18.61	24.06	32.04	38.20	46.82	50.84	48.66	56.29	62.84
OCM	55.01	67.79	72.17	19.80	25.90	35.99	41.59	50.42	55.30	51.81	57.52	**64.60**
PCR	56.97	63.08	76.39	19.99	29.66	36.28	39.84	48.17	52.38	49.20	54.41	61.36
ER + MKD	54.74	67.77	73.98	21.16	29.02	36.37	40.95	46.05	48.05	52.20	56.50	62.55
S6MOD	51.79	63.72	71.42	19.69	26.64	35.76	38.69	46.98	50.93	50.19	55.92	63.18
SMC-CB	**57.78**	**68.10**	75.31	**22.76**	29.72	**36.99**	42.32	51.27	54.43	52.53	58.29	63.37
SMC-TC	57.10	67.21	**76.52**	20.12	**29.88**	36.80	**43.22**	**52.17**	**56.20**	**53.43**	**59.19**	64.27

Table 3: Ablation study on the contributions of structural memory and structural consolidation. Results are reported on CIFAR-100 with a replay buffer of 200 samples.

SM	SC	Final Acc. (%) ↑	Avg. Forgetting (%) ↓
×	×	12.35	47.11
✓	×	19.64	35.27
×	✓	16.05	43.78
✓	✓	**22.76**	**32.13**

4.2 Overall Performance

Table 2 compares final average accuracy across four benchmarks and different episodic memory budgets. SMC achieves the highest accuracy in 11 of the 12 evaluated settings when considering both representative-selection variants, demonstrating that augmenting episodic replay with structural memory is effective across different dataset scales and memory capacities. The improvements are particularly visible under restricted replay budgets. On CIFAR-100 with a buffer of 200 samples, SMC-CB achieves 22.76% accuracy, compared with 12.35% for ER and 21.16% for ER + MKD. On ImageNet-Subset with 1,000 replay samples, SMC-TC reaches 43.22%, exceeding OCM at 41.59% and ER + MKD at 40.95%. These results are consistent with the motivation of SMC: when episodic memory provides sparse historical coverage, stored relational information offers an additional source of supervision during subsequent online updates. Both representative-selection variants exhibit strong performance, although their relative behavior varies across datasets and memory budgets. SMC-CB achieves higher accuracy in several CIFAR settings, while SMC-TC generally performs better on ImageNet-Subset and Tiny-ImageNet. At the largest Tiny-ImageNet buffer, OCM achieves 64.60%, compared with 64.27% for SMC-TC. Overall, the results show that structural memory can complement sample-level replay across a broad range of online continual-learning settings, without requiring a single representative-selection strategy to dominate in every configuration.

4.3 Ablation Study

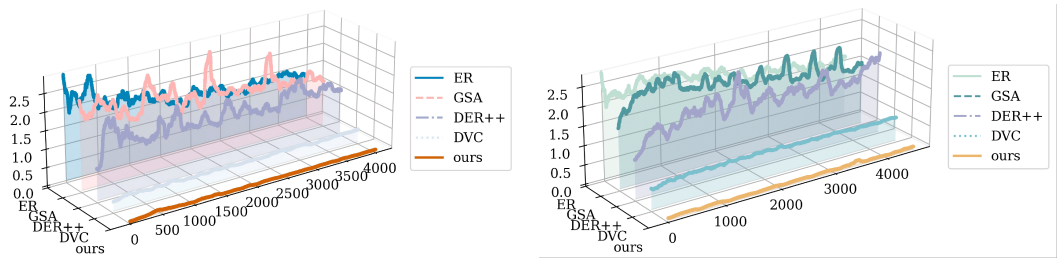
To understand the contribution of each design in SMC, we conduct an ablation study by separately removing structural memory and structural consolidation. SMC w/o Structural Memory removes the persistent relational snapshots while retaining the consolidation mechanism, whereas SMC w/o Structural Consolidation preserves the snapshots but removes the structural alignment objective. The results demonstrate that both components contribute to knowledge retention, validating the necessity of maintaining historical relational states and consolidating them during online adaptation.

4.4 Knowledge Retention

Final accuracy captures both the acquisition of new classes and the retention of previously learned classes. To assess retention directly, Table 4 reports average forgetting across datasets and episodic memory budgets. SMC achieves the lowest forgetting in 9 of 12 settings, including several memory-constrained regimes.

Table 4: Average forgetting (%) under different episodic memory budgets. Lower values indicate better retention relative to each task's historical peak accuracy; bold indicates the lowest reported value in each column.

Method	CIFAR-10			CIFAR-100			ImageNet-Subset			Tiny-ImageNet		
	200	500	1000	200	500	1000	1000	2000	5000	1000	2000	5000
ER	51.12	44.28	22.76	47.11	41.03	31.64	36.43	24.42	16.66	67.15	57.12	46.92
ER + SDP	50.47	33.58	17.35	**31.91**	27.60	20.73	22.10	15.88	10.12	36.21	31.04	24.79
DER++	**31.11**	27.90	28.18	45.84	36.91	31.66	26.78	34.49	37.56	28.94	28.74	25.74
DVC	44.82	32.41	33.43	52.44	45.41	41.38	31.36	29.64	46.35	73.05	70.58	72.54
GSA	39.98	24.16	15.59	48.71	42.38	28.63	47.07	33.09	16.86	29.58	25.80	21.33
OCM	44.89	20.86	12.24	39.96	35.24	19.85	22.38	16.33	13.24	29.14	24.31	19.88
PCR	36.36	23.58	22.82	49.97	42.52	34.70	38.17	27.76	17.76	52.61	39.54	45.34
MKD	44.89	20.68	12.27	43.34	24.28	17.22	**14.85**	9.40	5.23	28.94	25.13	18.20
SMC	42.48	**19.76**	**11.47**	32.13	**20.38**	**16.94**	15.33	**9.20**	**4.88**	**25.66**	**23.69**	**16.33**



(a) Feature drift.

(b) Structural drift.

Figure 3: Smoothed drift of ER, GSA, DER++, DVC, and SMC on CIFAR-10. SMC exhibits distinct feature and structural drift dynamics throughout the data stream.

For example, on CIFAR-100 with a 500-sample buffer, SMC reduces forgetting to 20.38%, compared with 24.28% for MKD and 35.24% for OCM. On Tiny-ImageNet with 5,000 replay samples, it achieves 16.33%, compared with 18.20% and 19.88%, respectively. Forgetting should be interpreted jointly with final accuracy: it measures degradation from each task's historical peak rather than absolute recognition performance. Thus, although DER++ has lower forgetting than SMC under the smallest CIFAR-10 buffer, its final accuracy is substantially lower. Together, Tables 2 and 4 show that SMC improves retention while preserving effective learning of newly introduced classes.

4.5 Representation Evolution and Structural Stability

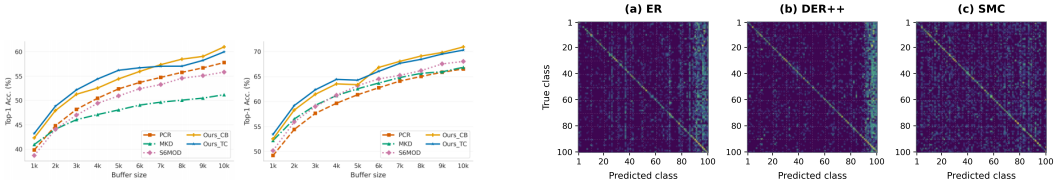
We next investigate whether the improved knowledge retention is accompanied by more stable historical representations. Following the definitions in Section 3, we evaluate feature drift and structural drift to distinguish absolute feature displacement from changes in relational organization. Importantly, structural drift is measured using class-prototype similarity matrices rather than the anchor-level distributions optimized by SMC, providing a separate diagnostic of historical representation structure.

Figure 3 compares representation drift on CIFAR-10 throughout online learning. SMC exhibits lower and smoother structural drift than the compared replay-based methods, indicating that the relational organization among historical classes changes less during subsequent updates. SMC also reduces feature displacement despite not directly constraining individual feature coordinates. These observations are consistent with the role of structural consolidation: aligning current anchor relationships with stored snapshots can constrain representation changes beyond those addressed by sample-level replay alone.

Together with the forgetting results in Table 4, the drift analysis provides complementary evidence that improved retention is accompanied by more stable historical representation organization. We do not interpret this association as establishing a causal relationship between structural drift and forgetting. For visualization, the curves are smoothed using a moving average with window $k = 11$ over 4K iterations; unsmoothed trajectories are provided in Appendix F.

Table 5: Effect of representative selection on CIFAR-100 under different episodic memory budgets. Mean and ΔMean summarize the three buffer settings, with ΔMean measured in percentage points relative to uniform sampling. Higher is better; bold indicates the best reported value in each column. Summary statistics are recomputed from the displayed per-budget values.

Sampling	200	500	1000	Mean	ΔMean
OCM	19.80	25.90	35.99	–	–
PCR	19.99	29.66	36.28	–	–
ER + MKD	21.16	29.02	36.37	–	–
S6MOD	19.69	26.64	35.76	–	–
Uniform	22.73	29.69	36.87	29.76	+0.00
Class-balanced (CB)	**22.76**	29.72	**36.99**	**29.82**	**+0.06**
Typical-center (TC)	20.12	**29.88**	36.80	28.93	-0.83
k-means	20.96	27.58	35.10	27.88	-1.88



(a) Memory scaling on ImageNet-Subset and Tiny-ImageNet.

(b) Final row-normalized confusion matrix on CIFAR-100.

Figure 4: Additional analysis of SMC. (a) Final accuracy under increasing episodic replay capacities on ImageNet-Subset and Tiny-ImageNet. SMC remains competitive across memory budgets, with larger relative gains in low-memory regimes. (b) Final row-normalized confusion matrix on CIFAR-100 after the complete task sequence (seed 0). SMC shows weaker concentration of errors toward recently introduced classes.

4.6 Structural Memory Construction

Table 5 shows that class-balanced sampling performs best overall, improving the mean accuracy over uniform sampling by 0.06 percentage points. Typical-center and k-means sampling are less stable, particularly under smaller buffers, indicating that representative selection remains memory-dependent.

4.7 Other Analysis

Memory scaling. As shown in Figure 4a, SMC remains competitive as the replay capacity increases from 1k to 10k, with larger relative gains in low-memory regimes where episodic replay provides sparse historical coverage.

Task-recency bias. Figure 4b shows that SMC produces a clearer diagonal and weaker off-diagonal concentration toward recently introduced classes, complementing the forgetting results in Table 4.

Computational overhead. The latency-accuracy trade-off and parameter-count comparison are reported in Appendix E. SMC incurs modest overhead while maintaining competitive accuracy.

5 Conclusion

We introduced **Structural Memory Consolidation (SMC)**, a memory-efficient framework that augments episodic replay with compact structural memories for online continual learning. By preserving structural snapshots of previously learned representations and aligning subsequent updates with these snapshots, SMC provides an additional mechanism for maintaining historical knowledge beyond stored examples. Extensive experiments across multiple online class-incremental benchmarks demonstrate that SMC consistently improves performance and retention under diverse memory budgets. Further analyses show that these improvements are accompanied by more stable representation dynamics, highlighting the potential of structural information as a complementary component of continual memory.

ETHICS STATEMENT

We acknowledge that all co-authors of this work have read and commit toadhering to the ICLR Code of Ethics.

REPRODUCIBILITY STATEMENT

We have included the source code with clear instructions, and will release them upon acceptance.

AI USE STATEMENT

AI-assisted tools were used only for improving writing clarity, grammar checking, and minor editing during manuscript preparation. The authors are solely responsible for the content, including the research methodology, experimental results, and conclusions.

REFERENCES

Rahaf Aljundi, Eugene Belilovsky, Tinne Tuytelaars, Laurent Charlin, Massimo Caccia, Min Lin, and Lucas Page-Caccia. Online continual learning with maximal interfered retrieval. *Advances in neural information processing systems*, 32, 2019.

Nader Asadi, MohammadReza Davari, Sudhir Mudur, Rahaf Aljundi, and Eugene Belilovsky. Prototype-sample relation distillation: towards replay-free continual learning. In *International conference on machine learning*, pp. 1093–1106. PMLR, 2023.

Pietro Buzzega, Matteo Boschini, Angelo Porrello, Davide Abati, and Simone Calderara. Dark experience for general continual learning: a strong, simple baseline. *Advances in neural information processing systems*, 33:15920–15930, 2020a.

Pietro Buzzega, Matteo Boschini, Angelo Porrello, Davide Abati, and Simone Calderara. Dark experience for general continual learning: a strong, simple baseline. *Advances in neural information processing systems*, 33:15920–15930, 2020b.

Hyuntak Cha, Jaeho Lee, and Jinwoo Shin. Co 2 l: Contrastive continual learning. In *2021 IEEE/CVF international conference on computer vision (ICCV)*, pp. 9496–9505. IEEE, 2021.

Arslan Chaudhry, Marcus Rohrbach, Mohamed Elhoseiny, Thalaiyasingam Ajanthan, Puneet K Dokania, Philip HS Torr, and Marc'Aurelio Ranzato. On tiny episodic memories in continual learning. *arXiv preprint arXiv:1902.10486*, 2019.

Matthias De Lange, Rahaf Aljundi, Marc Masana, Sarah Parisot, Xu Jia, Aleš Leonardis, Gregory Slabaugh, and Tinne Tuytelaars. A continual learning survey: Defying forgetting in classification tasks. *IEEE transactions on pattern analysis and machine intelligence*, 44(7):3366–3385, 2021.

Arthur Douillard, Matthieu Cord, Charles Ollion, Thomas Robert, and Eduardo Valle. Podnet: Pooled outputs distillation for small-tasks incremental learning. In *European conference on computer vision*, pp. 86–102. Springer, 2020.

Zijian Gao, Xingxing Zhang, Kele Xu, Xinjun Mao, and Huaimin Wang. Stabilizing zero-shot prediction: A novel antidote to forgetting in continual vision-language tasks. *Advances in Neural Information Processing Systems*, 37:128462–128488, 2024.

Zijian Gao, Shanhao Han, Xingxing Zhang, Kele Xu, Dulan Zhou, Xinjun Mao, Yong Dou, and Huaimin Wang. Maintaining fairness in logit-based knowledge distillation for class-incremental learning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 39, pp. 16763–16771, 2025.

Dipam Goswami, Albin Soutif-Cormerais, Yuyang Liu, Sandesh Kamath, Bart Twardowski, Joost Van De Weijer, et al. Resurrecting old classes with new data for exemplar-free continual learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 28525–28534, 2024.

Yiduo Guo, Bing Liu, and Dongyan Zhao. Online continual learning through mutual information maximization. In *International conference on machine learning*, pp. 8109–8126. PMLR, 2022.

Yiduo Guo, Bing Liu, and Dongyan Zhao. Dealing with cross-task class discrimination in online continual learning. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11878–11887. IEEE, 2023.

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

Saihui Hou, Xinyu Pan, Chen Change Loy, Zilei Wang, and Dahua Lin. Learning a unified classifier incrementally via rebalancing. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 831–839. IEEE, 2019.

Dahuin Jung, Dongjin Lee, Sunwon Hong, Hyemi Jang, Ho Bae, and Sungroh Yoon. New insights for the stability-plasticity dilemma in online continual learning. *arXiv preprint arXiv:2302.08741*, 2023.

Joonkyu Kim, Yejin Kim, and Jy-yong Sohn. Measuring representational shifts in continual learning: A linear transformation perspective. *arXiv preprint arXiv:2505.20970*, 2025.

Tatsuya Konishi, Mori Kurokawa, Chihiro Ono, Zixuan Ke, Gyuhak Kim, and Bing Liu. Parameter-level soft-masking for continual learning. In *International conference on machine learning*, pp. 17492–17505. PMLR, 2023.

Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25, 2012.

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

Sihao Liu, Yibo Yang, Xiaojie Li, David A Clifton, and Bernard Ghanem. Enhancing online continual learning with plug-and-play state space model and class-conditional mixture of discretization. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 20502–20511. IEEE, 2025.

David Lopez-Paz and Marc'Aurelio Ranzato. Gradient episodic memory for continual learning. *Advances in neural information processing systems*, 30, 2017.

Nicolas Michel, Maorong Wang, Ling Xiao, and Toshihiko Yamasaki. Rethinking momentum knowledge distillation in online continual learning. *arXiv preprint arXiv:2309.02870*, 2023.

Wonpyo Park, Dongju Kim, Yan Lu, and Minsu Cho. Relational knowledge distillation. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3962–3971. IEEE, 2019.

Sylvestre-Alvise Rebuffi, Alexander Kolesnikov, Georg Sperl, and Christoph H Lampert. icarl: Incremental classifier and representation learning. In *2017 IEEE conference on computer vision and pattern recognition (CVPR)*, pp. 5533–5542. IEEE, 2017.

David Rolnick, Arun Ahuja, Jonathan Schwarz, Timothy Lillicrap, and Gregory Wayne. Experience replay for continual learning. *Advances in neural information processing systems*, 32, 2019.

Dongsub Shim, Zheda Mai, Jihwan Jeong, Scott Sanner, Hyunwoo Kim, and Jongseong Jang. Online class-incremental continual learning with adversarial shapley value. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pp. 9630–9638, 2021.

Albin Soutif-Cormerais, Antonio Carta, Andrea Cossu, Julio Hurtado, Vincenzo Lomonaco, Joost Van de Weijer, and Hamed Hemati. A comprehensive empirical evaluation on online continual learning. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 3518–3528, 2023.

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

Liyuan Wang, Xingxing Zhang, Hang Su, and Jun Zhu. A comprehensive survey of continual learning: Theory, method and application. *IEEE transactions on pattern analysis and machine intelligence*, 46(8):5362–5383, 2024a.

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

Yujie Wei, Jiaxin Ye, Zhizhong Huang, Junping Zhang, and Hongming Shan. Online prototype learning for online continual learning. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 18718–18728. IEEE, 2023.

Fangwen Wu, Lechao Cheng, Shengeng Tang, Xiaofeng Zhu, Chaowei Fang, Dingwen Zhang, and Meng Wang. Navigating semantic drift in task-agnostic class-incremental learning. *arXiv preprint arXiv:2502.07560*, 2025.

Hongwei Yan, Liyuan Wang, Kaisheng Ma, and Yi Zhong. Orchestrate latent expertise: Advancing online continual learning with multi-level supervision and reverse self-distillation. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 23670–23680. IEEE, 2024.

Lu Yu, Bartlomiej Twardowski, Xialei Liu, Luis Herranz, Kai Wang, Yongmei Cheng, Shangling Jui, and Joost van de Weijer. Semantic drift compensation for class-incremental learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 6982–6991, 2020.

Han Zang, Yongfeng Dong, Linhao Li, Liang Yang, and Yu Wang. Topology-aware knowledge preservation for class-incremental learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 28059–28067, 2026.

A Choice of Representative Sampling Strategies

Cross-task structural distillation in OCL relies on a replay buffer to approximate the historical data distribution. Therefore, the representativeness of buffered samples directly affects the stability of the distillation signal and generalization across tasks. We evaluate several strategies on CIFAR-100 under buffers of 200/500/1000; results are reported in Table 5.

Compared strategies.

- **BaseSample**: default sampling without explicitly enforcing representativeness.
- **EqualSample**: balanced sampling at the class/cluster level to prevent frequent or recent samples from dominating the buffer.
- **CenterDist**: samples closer to the feature-space center, retaining typical representatives and reducing noisy or boundary samples.
- k**-means**: a clustering-based prototype strategy that keeps nearest neighbors of cluster centers.

Results and analysis. EqualSample (class-balanced sampling) provides modest gains over BaseSample (uniform sampling) at all three buffer sizes, improving mean accuracy from 29.76% to 29.82% (ΔMean $= +0.06$ percentage points). EqualSample achieves the highest reported accuracy at buffers of 200 and 1,000. CenterDist (typical-center sampling) performs best at 500, but its accuracy of 20.12% at 200 lowers its mean to 28.93% (ΔMean $= -0.83$ percentage points). k-means yields the lowest mean accuracy, 27.88%. Thus, class-balanced sampling has the highest reported mean in this comparison, while the benefit of typical-center sampling varies with the memory budget.

Algorithm 1 Online Training with Cross-task Structure-wise Distillation (CTR-KD)

Require: Streamed tasks $\{\mathcal{S}_1, \ldots, \mathcal{S}_T\}$ with class sets $\{C_t\}$; memory budget B; representatives per past task K; temperatures $\tau_{\text{tea}}, \tau_{\text{stu}}$; sampling strategy CB or TC
Ensure: Trained parameters (ϕ, θ)

Initialize model (g_ϕ, h_θ); global memory $\mathcal{M} \leftarrow \varnothing$
Initialize cached relations $\mathcal{P} \leftarrow \varnothing$ and cached representatives $\mathcal{R} \leftarrow \varnothing$
for $t = 1$ to T **do**
  if $t > 1$ **then** ▷ Task-boundary snapshot and structure caching
    $(\phi_{t-1}^-, \theta_{t-1}^-) \leftarrow (\phi, \theta)$
    for $i = 1$ to $t-1$ **do**
      $\mathcal{M}_i \leftarrow \{(x, y) \in \mathcal{M} : y \in C_i\}$
      $R_i \leftarrow \textsc{SelectReps}(()\mathcal{M}_i, K, \phi_{t-1}^-, \text{CB/TC})$
      Cache R_i into $\mathcal{R}$
      $\widetilde{Z}_i^- \leftarrow \textsc{Embed}(()R_i, \phi_{t-1}^-)$ ▷ $K \times d$
    end for
    for $i = 2$ to $t-1$ **do**
      $S_{\text{tea}}^{i-1,i} \leftarrow \widetilde{Z}_{i-1}^- (\widetilde{Z}_i^-)^\top$
      $P_{\text{tea}}^{i-1,i} \leftarrow \text{softmax}_{\text{row}}(S_{\text{tea}}^{i-1,i} / \tau_{\text{tea}})$
      Cache $P_{\text{tea}}^{i-1,i}$ into $\mathcal{P}$
    end for
  end if
  for all mini-batches $\mathcal{B} \subset \mathcal{S}_t$ in arrival order **do**
    Sample replay batch $\mathcal{B}_M \subseteq \mathcal{M}$ (if $\mathcal{M} \neq \varnothing$)
    $\mathcal{B}_{\text{train}} \leftarrow \mathcal{B} \cup \mathcal{B}_M$
    Compute $\mathcal{L}_{\text{cls}}$ on $\mathcal{B}_{\text{train}}$ using logits $\ell(x; \phi, \theta)$
    $\mathcal{L}_{\text{CTR}} \leftarrow 0$
    if $t > 2$ **then**
      for $i = 1$ to $t-1$ **do**
        $\widetilde{Z}_i \leftarrow \textsc{Embed}(()R_i, \phi)$ ▷ Current student embeddings
      end for
      for $i = 2$ to $t-1$ **do**
        $S_{\text{stu}}^{i-1,i} \leftarrow \widetilde{Z}_{i-1} \widetilde{Z}_i^\top$
        $P_{\text{stu}}^{i-1,i} \leftarrow \text{softmax}_{\text{row}}(S_{\text{stu}}^{i-1,i} / \tau_{\text{stu}})$
        $\mathcal{L}_{\text{CTR}} \leftarrow \mathcal{L}_{\text{CTR}} + \frac{1}{K} \sum_{p=1}^{K} \text{CE}(P_{\text{tea}}^{i-1,i}(p,:), P_{\text{stu}}^{i-1,i}(p,:))$
      end for
    end if
    $\mathcal{L} \leftarrow \mathcal{L}_{\text{cls}} + \lambda \mathcal{L}_{\text{CTR}}$
    Update memory $\mathcal{M}$ with incoming samples in $\mathcal{B}$ under $|\mathcal{M}| \leq B$
  end for
end for

B Buffer Scaling

CenterDist versus EqualSample reflects a typicality-coverage trade-off under distribution shift. Across both benchmarks, CenterDist tends to outperform at smaller buffers, while EqualSample becomes competitive (and occasionally slightly better) when memory is large. Once coverage is sufficient, enforcing balance helps preserve long-tail classes and mitigate recency bias, matching the objective of replay as an approximate distributional support.

C Pseudocode

Algorithm 1 gives the online training procedure for CTR-KD.

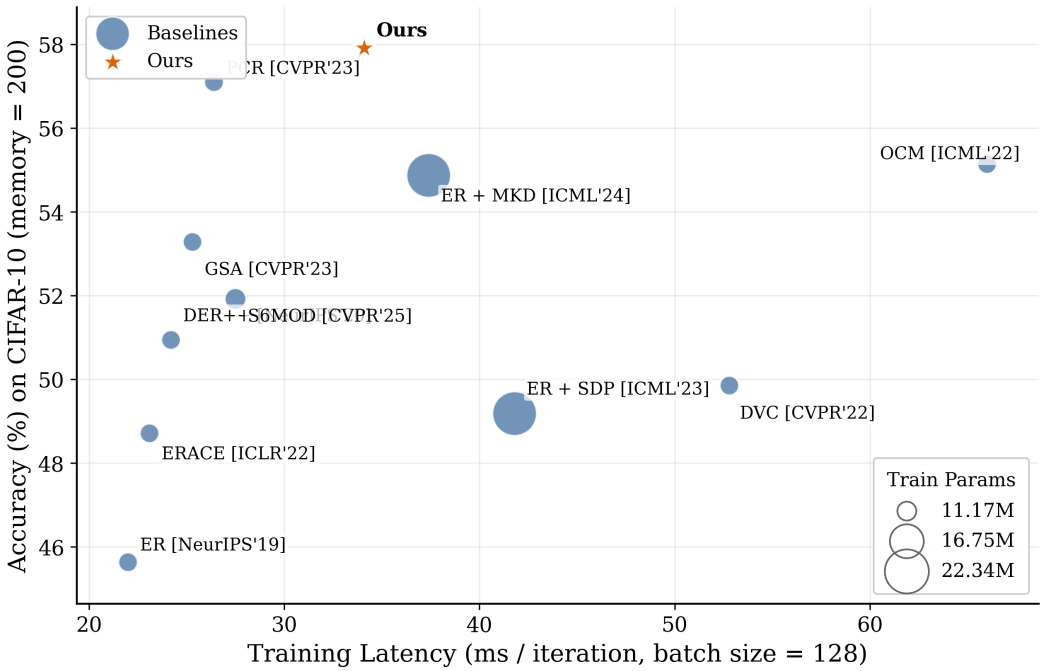


Figure 5: Accuracy-latency trade-off with parameter-aware visualization. Training latency is measured in ms/iteration with batch size 128 on CIFAR-10 under memory budget $|\mathcal{M}| = 200$; marker size denotes trainable parameters.

D Experimental Details

D.1 Datasets and Selection Rationale

To evaluate OCL across image resolutions, scales, and category granularities, we adopt CIFAR-10, CIFAR-100, Tiny-ImageNet, and ImageNet-Split. Together they span low-resolution recognition, fine-grained recognition, increased resolution and class count, and greater visual/distributional variety. Unless otherwise specified, experiments use the class-incremental protocol with identical normalization and data augmentation across datasets.

D.1.1 Dataset summaries

CIFAR-10 and CIFAR-100 contain natural images at 32×32 resolution; CIFAR-10 has 10 classes, while CIFAR-100 has 100 classes organized into 20 superclasses. Tiny-ImageNet increases both class count and resolution, with 200 classes, 500 training images and 50 validation images per class at 64×64. ImageNetSplit (ImageNet-100) is a 100-class subset of ImageNet-1k; images are resized/cropped to 224×224 during training and evaluation, with approximate train/validation counts of 130k/5k.

E Computational Overhead

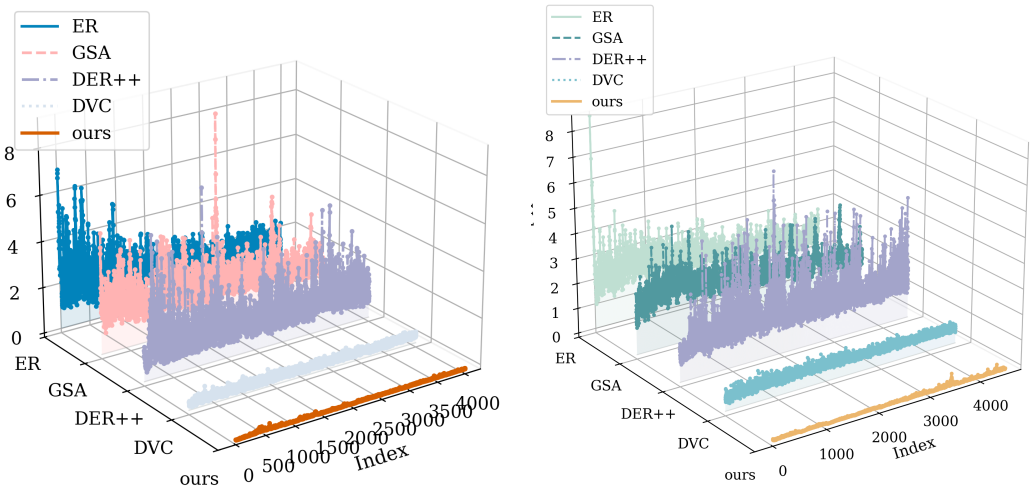
As shown in Figure 5, CTR-KD lies in the upper-left region of the accuracy-latency plane, indicating a favorable efficiency-performance trade-off. It achieves the highest or near-highest accuracy while maintaining competitive training latency comparable to lightweight replay baselines. The gains are not obtained by trading off excessive computation, but by using a more effective geometric learning signal.

Table 6: Hyperparameter search spaces and settings used for each method.

Method	Parameter	Values
ER	optimizer	SGD, Adam
	weight decay	0, 10^{-4}
	learning rate	10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}
	momentum	0, 0.9
	augmentation	full, partial
ER-ACE	optimizer	SGD, Adam
	weight decay	0, 10^{-4}
	learning rate	10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}
	momentum	0, 0.9
	augmentation	full, partial
DER++	optimizer	SGD, Adam
	weight decay	0, 10^{-4}
	learning rate	10^{-4}, 10^{-3}, 10^{-2}, 0.03
	momentum	0, 0.9
	augmentation	full, partial
	α	0.1, 0.2, 0.5, 1.0
	β	0.5, 1.0
DVC	optimizer	SGD, Adam
	weight decay	0, 10^{-4}
	learning rate	10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}
	momentum	0, 0.9
	augmentation	full, partial
GSA	optimizer	SGD, Adam
	weight decay	0
	learning rate	10^{-4}, 5×10^{-4}, 10^{-2}, 0.05, 0.1
	momentum	0, 0.9
	augmentation	full, partial
ER+SDP	optimizer	Adam
	weight decay	0
	learning rate	3×10^{-4}
	momentum	0
	μ_e	10, 100, 1000, 10000
	c_2	0.5, 0.75, 0.9
PCR	optimizer	Adam
	weight decay	0
	learning rate	5×10^{-4}
	momentum	0.9
	augmentation	full
OCM	optimizer	Adam
	weight decay	0
	learning rate	5×10^{-4}
	momentum	0.9
	augmentation	full
	α	0.01, 0.05, 0.1, 0.5, 0.9, 0.95, 0.99
CTR-KD	optimizer	Adam
	weight decay	0
	learning rate	5×10^{-4}
	momentum	0.9
	augmentation	full
	$\lambda_1, \lambda_2, \lambda_3$	5.5, 0.0005, 0.005

F UNSMOOTHED FEATURE DRIFT AND STRUCTURE DRIFT

In the main text, we report smoothed feature and structure drift curves for clearer visualization. Figure 6 provides the unsmoothed results, which preserve the original fluctuations and are complementary to the figures in the main paper.



(a) Feature drift.

(b) Structure drift.

Figure 6: Unsmoothed feature and structural drift comparison on CIFAR-10.