Topology-aware Knowledge Preservation for Class-Incremental Learning

Han Zang[1], Yongfeng Dong[1], Linhao Li[1*], Liang Yang[1], Yu Wang[2]

[1]School of Artificial Intelligence, Hebei University of Technology, Tianjin, China
[2]College of Intelligence and Computing, Tianjin University, Tianjin, China
202312802003@stu.hebut.edu.cn, {dongyf, lilinhao}@hebut.edu.cn, yangliang@vip.qq.com, wang.yu@tju.edu.cn

Abstract

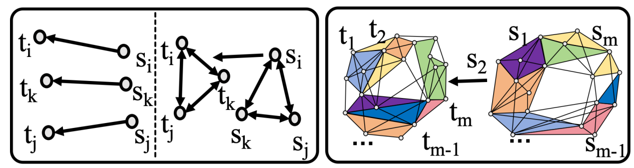
Class Incremental Learning (CIL) aims to enable model to continually learn new classes while retaining previously learned knowledge. The principal challenge in CIL is catastrophic forgetting, which prior approaches typically address by distilling knowledge from previous model. However, such way is often limited to pairwise alignment, failing to preserve the underlying global manifold structure of feature space—ultimately resulting in semantic drift over time. To capture multi-scale structural patterns in the feature space, we propose a topology-aware distillation framework that leverages persistent homology. Specifically, by enforcing topological alignment across incremental stages, our method ensures structure-consistent knowledge transfer and robust preservation of old classes. Furthermore, we still devise a dual-branch architecture with an inverse sampling and dynamic reweighting mechanism that addresses the inherent data imbalance in standard replay-based frameworks. These innovations coalesce into TaKP (Topology-aware Knowledge Preservation), a unified framework designed to enhance knowledge preservation in CIL. Extensive experiments demonstrate that TaKP achieves state-of-the-art performance on multiple benchmarks, significantly improving old-class preservation and average accuracy.

Code — https://github.com/zanghan111/TaKP

Introduction

Class Incremental Learning (CIL) aims to enable model to continuously learn new classes over time while retaining knowledge of previously learned ones (Wang et al. 2024). It has broad applications in real-world scenarios such as autonomous driving, medical diagnostics, and adaptive robotics. However, CIL remains fundamentally challenged by catastrophic forgetting (CF), where learning new classes significantly degrades performance on prior classes.

Many approaches have been proposed to mitigate catastrophic forgetting in CIL (Rebuffi et al. 2017; Kirkpatrick, Pascanu, and Rabinowitz 2017; Wang et al. 2022; Yan, Xie, and He 2021; Zhou et al. 2023). Among these, the example replay strategy stands out as a widely adopted solution, which stores a small set of samples from previously learned classes and incorporates them during training on new classes. Knowledge distillation (KD) methods based on replay (Kang, Park, and Han 2022; Wen et al. 2024; Fan et al. 2024b,a; Kang et al. 2023; Gao et al. 2025) have been proposed to achieve the transfer of old class knowledge by constraining the consistency of output between previous and current models on stored samples, thereby alleviating catastrophic forgetting in CIL.



(a) Vanilla KD. (b) Topology-aware KD.

Figure 1: Comparison between vanilla KD and topology-aware KD. (a) Left: logits or features-based KD; Right: relation-based KD. (b) Topology-aware KD. t_i and s_i denote the outputs of the teacher (previous) model and student (current) model respectively.

Vanilla KD methods transfer information by aligning output logits or intermediate features (Kang, Park, and Han 2022; Wen et al. 2024) between current and previous models. More recent work has explored relational distillation methods that attempt to retain inter-instance structure—such as class similarity (Kang et al. 2023) or local structure similarity (Fan et al. 2024b,a)—offering richer supervision beyond pairwise matching, as illustrated in Fig.1(a). However, despite these advances, existing KD methods still lack global structural awareness. They remain limited in their ability to capture the global topology of the feature space. **By focusing primarily on local consistency, these methods overlook the underlying global manifold structure, resulting in topological distortion and accelerating the erosion of old-class knowledge over time.**

To address the aforementioned challenge, we explore extracting multi-scale shape and structural information from stored samples, which is more stable for knowledge preservation. Inspired by the manifold assumption and the homomorphism between manifolds and simplicial complexes (Melas-Kyriazi 2020), we abstract high-dimensional visual

[*]Corresponding authors.

features into simplicial complexes and explore the preservation of multi-scale topological structures, as illustrated in Fig.1(b). Persistent Homology (PH) is utilized to efficiently capture global topological structures across multiple scales, revealing distributional shapes, multi-scale connectivity, and structural patterns. These stable topological signatures are invariant to homeomorphisms and robust to perturbations, enabling structure-aware knowledge preservation in CIL.

In addition, we still notice the influence of the inherent data imbalance in CIL - as new classes arrive sequentially and old-class data diminishes. To address this challenge, we explore whether decoupling representation learning and classifier adaptation through a dual-branch design can be effectively applied to the CIL setting. Building on this idea, we propose a unified dual-branch framework with adaptive rebalancing. Our method explicitly models data distribution shifts, employs inverse sampling to highlight underrepresented old classes, and introduces a dynamic weighting mechanism to gradually shift focus from general representation learning to old-class preservation.

To this end, we propose Topology-aware Knowledge Preservation (TaKP), a unified CIL framework that incorporates topology-aware knowledge preservation and dual-branch adaptive rebalancing. At its core, TaKP introduces a topology-aware distillation framework that leverages persistent homology to extract topological signatures from latent features. These signatures encode high-order structural patterns, which are resilient to perturbations and representation shifts. A dedicated topological distillation loss guides the current model to align with the previous structural landscape, enabling globally consistent knowledge transfer and mitigating semantic erosion across incremental sessions. In parallel, TaKP adopts a dual-branch architecture with adaptive rebalancing, where a conventional branch captures generalizable patterns from the original distribution, and a rebalancing branch emphasizes old-class representations via inverse sampling. A dynamic weighting coefficient α balances the two branches throughout training. The primary contributions of our work are as follows:

- To address the lack of global structural constraints in existing knowledge transfer methods, we propose a topology-aware distillation framework, that effectively mitigates the semantic erosion caused by representation drift.
- We propose a topological distillation loss that aligns the structural landscapes between current and previous models, and enforces global consistency for robust knowledge transfer.
- To mitigate the data imbalance inherent in CIL, we propose an adaptively rebalanced dual-branch architecture with a inverse sampling and dynamic reweighting mechanism.
- We validate our method through extensive experiments and visualizations, showing that it effectively mitigates catastrophic forgetting, and outperforms existing methods on several benchmarks.

Related Work

Persistent homology in deep learning

Topological Data Analysis (TDA), particularly persistent homology (PH), has emerged as a powerful tool for extracting multiscale geometric and structural features from discrete data, often visualized through persistence diagrams (PDs) (Zhao et al. 2020; Zhao and Wang 2019; Hofer et al. 2020; Yan et al. 2021; Cang, Mu, and Wei 2018). In computer vision, TDA has been applied to object-level tasks such as image segmentation (Clough et al. 2022; He et al. 2023) and embedding optimization (He et al. 2023), but its potential for category-level representation learning remains underexplored. Although Liu (Liu, Dey, and Gleich 2023) proposed a topological prediction model for image datasets, it lacks integration with deep learning frameworks. Madhu and Chepuri (Madhu and Chepuri 2023) introduced a simplex contrastive loss to enhance local neighbor consistency but overlooked global topological constraints. Recent studies (Du et al. 2022; Jeon et al. 2024) used PH merely as a preprocessing step for knowledge distillation, not taking full advantage of its full potential in intermediate feature learning. In contrast to these works, our approach directly integrates PH into deep neural networks to dynamically capture multiscale topological structures (e.g., inter-class relationships, distributional holes) in feature embeddings, providing geometrically rich priors beyond conventional methods.

Class incremental learning

Class Incremental Learning is the task of enabling models to learn new classes sequentially while retaining knowledge of previously learned classes, with limited memory for old samples. Existing methods can be divided into three classes.

Regularization-based approaches (Rebuffi et al. 2017; Kirkpatrick, Pascanu, and Rabinowitz 2017; Chen, Wang, and Hu 2023; Cha et al. 2023; Dong et al. 2023) work by selectively controlling changes in key network parameters or by maintaining consistency in the output of prediction functions. MGRB (Chen, Wang, and Hu 2023) constructs a knowledge structure for existing classes and is utilized for regularization when learning new classes. (Cha et al. 2023) solves the problem of imbalance more accurately through batch normalization that uses reshaping and repetitive operations for horizontal connections during the training process.

Replay-based methods (Wen et al. 2024; Fan et al. 2024b,a; Gao et al. 2025) store few representative samples for retraining or old knowledge distillation. DSGD (Fan et al. 2024b) proposes a dynamic graph construction and preserves the invariance of the subgraph structure, which maintains instance associations during the CIL process. PsHD (Fan et al. 2024a) preserves intrinsic structural information that is insensitive to noise. (Wen et al. 2024) introduces a multi-teacher knowledge distillation approach for incremental learning, employing weight permutation, feature perturbation, and diversity regularization to ensure distinct mechanisms among teachers.

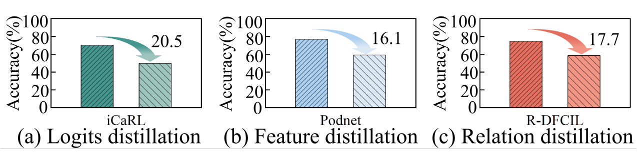
Prototype-based methods (Gao et al. 2025; Douillard et al. 2020; McDonnell et al. 2023; Rypesc et al. 2024) establish a prototype for each task and update the prototypes in subsequent phases, classifying samples into the category of the most similar prototype in the inference phase. CREATE (Chen et al. 2025) assigns each class a lightweight auto-encoder to learn a compact latent space, ensuring only correct-class samples reconstruct well. RanPAC (McDonnell et al. 2023) proposes projecting features to an expanded dimension, which enhances the linear separability of prototypes. SEED (Rypesc et al. 2024) employs one Gaussian distribution for each class and performs an ensemble of Bayes classifiers.

Preliminary Analysis

Problem formulation CIL is defined as follows. Assume that given $T = \{T_1, T_2, ..., T_N\}$ representing T tasks, for a task T_n, its data is represented as $D_n = \{x_i, y_i\}_{i=1}^{M_n}$, where M_n represents the total number of samples for T_n, x_i denotes a sample and y_i denotes the corresponding label. The goal of CIL is to classify all classes as $D = \bigcup_{i=1}^{N} D_i$. To be specific, for any two tasks T_i and T_j where $i \neq j$, $D_i \neq D_j$, which means that each is task independent, and the data sets they possess are unique. After training, model can infer the classes $Y_{T_i} \cup Y_{T_j}$, where Y is a label space.

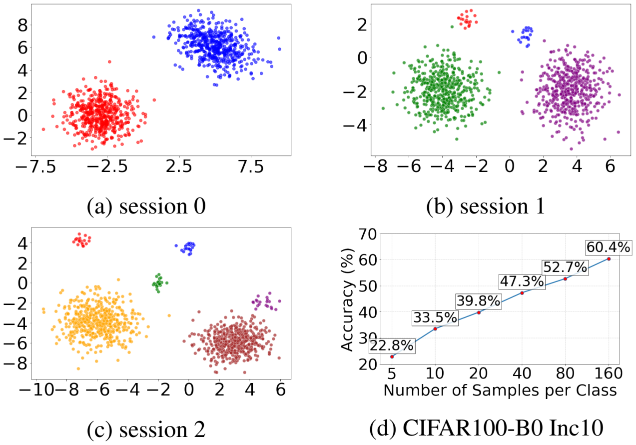
Analyzing knowledge distillation in CIL Current CIL methods primarily rely on pairwise knowledge distillation, where either output logits or intermediate features are directly matched between old and new models. Such approaches inherently ignore the global structural relationships among samples. The pairwise distillation fails to preserve manifold topology: neighborhood relationships in feature space. The empirical evaluation in Fig.2 quantifies old-class performance degradation across distillation strategies: conventional point-to-point methods cause severe catastrophic forgetting of old-class knowledge, as evidenced by the average incremental accuracy metrics.

Analyzing data imbalance in CIL In CIL, the asymmetric data assumption—where limited old-class samples coexist with abundant new-class data—induces severe class imbalance that systematically degrades performance. This occurs through representation bias, where new-class dominance skews the latent space, and catastrophic forgetting amplification, where sparse old-class exemplars fail to preserve prior decision boundaries. Fig.3 (a) – (c) visualize this via t-SNE under CIFAR100-B0 Inc2, showing intensifying imbalance as fixed-memory constraints persist across sessions (Session:0→n). Fig.3 (d) further validates the impact on CIFAR100-B0 Inc10, where larger examples memory reserves improve accuracy of the baseline (Masana et al. 2023), directly linking performance drop to imbalance.



(a) Logits distillation (b) Feature distillation (c) Relation distillation

Figure 2: Quantify the decrease in average incremental performance of old classes using different distillation methods.



(a) session 0 (b) session 1

(c) session 2 (d) CIFAR100-B0 Inc10

Figure 3: Evolution of class distributions and performance analysis in class incremental learning.

Methodology

Dual-Branch with Adaptive Rebalancing

To address data imbalance, we propose a unified dual-branch network that jointly handles representation learning and classifier adaptation. The model consists of two complementary branches: **Conventional Branch**: Learns generalizable patterns from the original imbalanced data distribution. **Rebalancing Branch**: Coupled with an inverse sampler, it specifically models old-class data by adaptive sample reweighting.

Both branches share the backbone network. Input samples are fed into their respective branches to obtain feature vectors f_c (conventional branch) and f_r (rebalancing branch). The adaptive weighting parameter α controls the contribution of each branch, producing weighted feature vectors that are independently processed by their classifiers. The final output is obtained through element-wise summation of both classifiers' outputs:

$$z = \alpha \mathbf{W_c}^\top f_c + (1 - \alpha)\mathbf{W_r}^\top f_r. \tag{1}$$

The adaptive weighting parameter α is automatically generated based on training iterations to dynamically adjust the model's focus. Initially, α prioritizes learning generalizable features from the original data distribution, then gradually shifts attention to old-class data preservation. Crucially, α regulates parameter updates in both branches - when emphasizing old-class data in later stages, it safeguards previously learned universal features from degradation.

$$\alpha = 1 - (\frac{T}{T_{max}})^2. \tag{2}$$

Concretely, the number of total training epochs is denoted as T_{max} and the current epoch is T. The rebalancing branch incorporates an inverse sampler to address extreme class imbalance, particularly enhancing old-class classification accuracy. The sampling probability p_i for class i is inversely proportional to its sample size:

$$p_i = \frac{w_i}{\sum_{j=1}^{C} w_j}. \tag{3}$$

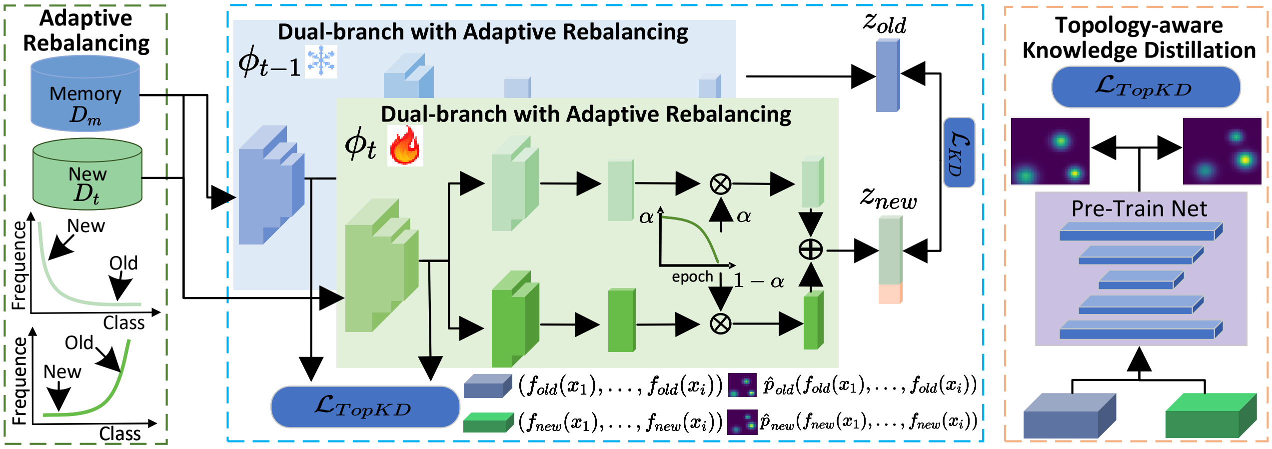


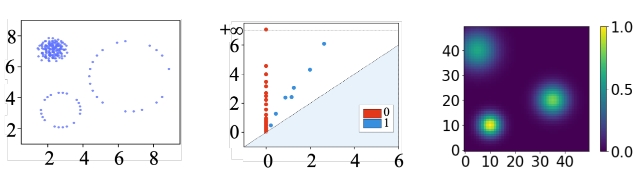
Figure 4: Illustration of our proposed TaKP framework for CIL. Two curves represent the class sampling frequency, corresponding to dual-branch respectively. In the previous model, α was set to 0.5. Right corresponds to our topological distillation loss, serving as $\mathcal{L}_{TopKD}$, employed on the stored samples.

Where $w_i = \frac{N_{max}}{N_i}$. Where N_i denotes the number of samples in class i, and N_{max} represents the maximum sample number of all classes. Randomly sample a class according to p_i, and then uniformly select a sample from class i with replacement. By repeating this reversed sampling process, a mini-batch of training data is obtained.

Distillation of topological knowledge

Background of persistent homology. Persistent homology tracks topological features through barcodes and persistence diagrams. Features emerge at birth time δ_b and vanish at death time δ_d during a filtration of sublevel sets. Their persistence $(\delta_b - \delta_d)$ indicates significance, with longer durations marking stable structures. Diagrams compactly encode multiscale topological patterns.

For point cloud data (PCD) analysis, the Vietoris-Rips filtration is commonly used due to its computational efficiency. It constructs simplicial complexes by connecting points within increasing distance thresholds, revealing the underlying topology at different scales. The filtration process tracks how topological features evolve as the connection radius grows, capturing both local and global structural properties. To integrate persistence diagrams into machine learning frameworks, vectorization methods transform them into finite-dimensional representations. Persistence images (PIs) are a common approach, mapping a diagram to a weighted Gaussian surface over a discretized grid. The diagram is first rotated via $(b, d) \mapsto (b, d-b)$, then convolved with Gaussian kernels whose weights emphasize features with long persistence. The surface is integrated over pixel domains to generate the final vector. As shown in Fig.5, we demonstrate the examples of persistent homology.



(a) Point cloud data
(b) Persistence diagram
(c) Persistence image

Figure 5: Examples of persistent homology (PH).

Topology-aware knowledge distillation. We extract sample embeddings through the backbone network and define their persistent homology representations - specifically manifested as persistence diagrams - as global topological knowledge. These PDs comprehensively record the topological structure of feature space. Through persistence image (PI) vectorization with its tunable weighting mechanism, we effectively enhance the model's capacity to capture global topological features.

We employ RipsNet to approximate the computation of persistence diagrams and persistence images per batch. The old model generates point cloud data from input batch (containing m tuples of samples), where each PCD is formed by the embeddings $f(x_i)$. PDs are computed via Rips filtration.

$$P = (f(x_1), ..., f(x_m)). \tag{4}$$

We consider $f(x_i)$ as a point by passing through the pooling layer. The persistence image construction begins with a coordinate transformation that converts the barcode B from (b, d) to $(b, d-b)$ space through the mapping $T : \mathbb{R}^2 \to \mathbb{R}$, explicitly representing topological persistence as vertical displacement. The persistence surface is constructed by convolving the transformed diagram with a weighted combination of Gaussian kernels, formally expressed as:

$$\beta_B(f) = \sum_{\mu \in T(B)} \omega(\mu) g_\mu(f). \tag{5}$$

Where the Gaussian kernel $g_\mu(f)$ is defined by:

$$g_\mu(f) = \frac{1}{2\pi\sigma^2} exp(-\frac{\|f-\mu\|^2}{2\sigma^2}). \tag{6}$$

Where μ is mean and σ^2 is variance. The PDs are calculated by using the Rips complex. To compute PIs, we set the grid size to 20 × 20 and the weight function $w(x,y) = 10(tanh(y) + ln(100x + 1))$. For the given PCD $(f(x_1), ..., f(x_m))$ in P, the exact PI is represented as $p(f(x_1), ..., f(x_m))$, and the approximated PI is represented as $\hat{p}(f(x_1), ..., f(x_m))$. We approximate the PIs minimizing the L2 loss function defined as follows:

$$\mathcal{L} = \sum_{(f(x_1),...,f(x_m))\in P} \mathcal{L}_2(p(f(x_1), ..., f(x_m)), \hat{p}(f(x_1), ..., f(x_m))). \tag{7}$$

When training is complete, RipsNet is frozen during the training of the new network to approximate PIs of the embedding features from the old and new models.

Objective function

In order to enable new model to imitate the old model's PI, we define the topological distillation loss function $\mathcal{L}_{TopKD}$ as follows:

$$\mathcal{L}_{TopKD} = \sum_{(x_1,...,x_i)\in D_m} \mathcal{L}_2(\hat{p}_{old}(f_{old}(x_1), ..., f_{old}(x_i)), (\hat{p}_{new}(f_{new}(x_1), ..., f_{new}(x_i))). \tag{8}$$

The logits represent high-level task-relevant knowledge; therefore, to use them further, we define the final loss function by combining the vanilla KD loss with the topology distillation loss, as follows:

$$\mathcal{L}_{total} = \mathcal{L}_{CE} + \mathcal{L}_{KD} + \mathcal{L}_{TopKD}. \tag{9}$$

Where $\mathcal{L}_{KD}$ is defined as follows:

$$\mathcal{L}_{KD} = KL(z_{old}^{1:c_{t-1}}, z_{new}^{1:c_{t-1}}). \tag{10}$$

Among them, $1 : c_{t-1}$ is the class learned by the old model.

Experiments

Experiments setup

Datasets We validate our method on the widely used benchmark **CIFAR100**(Krizhevsky and Hinton 2019) and **ImageNet100**(Deng et al. 2009). CIFAR100 is a classification dataset with 60,000 32 × 32 RGB images from 100 classes. Each class contains 500 training images and 100 testing images. ImageNet-100 is composed of 100 classes with 1,300 images per class for training and 500 images per class for validation. ImageNet-100 resembles real-world scenes with a higher resolution of 256 × 256.

Implementation Details. In our study, we utilized two data partitioning methods for the CIFAR100 and ImageNet100 datasets. Taking CIFAR100 as an example: first method involved dividing the dataset evenly among five sessions, with each session handling 20 classes. This setup is denoted as CIFAR100-B0 Inc20; second method involved splitting the dataset equally among ten sessions, with each session handling 10 classes. This setup is represented as CIFAR100-B0 Inc10. We select 20 samples for each old class as a memory buffer. We adopt resnet-18 as the model. We adopt an SGD optimizer with a weight decay of 2e-4 and a momentum of 0.9. The initial learning rate is set to 0.001 and decays with cosine annealing. We train each task for 80 epochs with a batch size of 64 using Nvidia 4090 GPUs with 24GB of RAM.

Metrics We evaluate the performance of different methods considering the average accuracy over all the classes seen after the last task: $ACC = \frac{1}{N}\sum_{i=1}^{N} A_{N,i}$ (Wang et al. 2024), where $A_{N,i}$ is the accuracy of the i-th task. Analysis with backward transfer, which presents the degree of forgetting $BWT = \frac{1}{N-1}\sum_{i=1}^{N-1}(A_{N,i} - A_{i,i})$ (Fan et al. 2024a), where $A_{i,j}$ represents the test accuracy in j-th task after completely learning i-th task.

Baselines We mainly considered two types of methods: **distillation based methods** iCaRl (Rebuffi et al. 2017), PODNet (Douillard et al. 2020), Foster (Wang et al. 2022), MTD (Wen et al. 2024), DSGD (Fan et al. 2024b), PsHD (Fan et al. 2024a), LKD (Gao et al. 2025), CREATE (Chen et al. 2025) and **class rebalancing methods** CLAD (Xu et al. 2024), RNKS (Song, Chen, and Du 2024), DBL (Zhi et al. 2025).

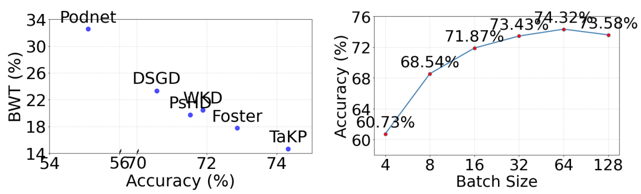
Experimental Results

Comparative performance. The experimental results in Table 1 demonstrate the superior performance of our proposed method across both benchmark datasets. Comparative analysis reveals that topology-based approaches (TaKP, PsHD) consistently outperform traditional distillation-based methods, validating that global topological feature modeling achieves more efficient feature representation than local sample-level similarity distillation. Notably, pure knowledge distillation methods show suboptimal performance in class rebalancing tasks, indicating their evident deficiency in preserving prior-class knowledge. These comprehensive experimental results prove that our strategy combining topological knowledge distillation with explicit class rebalancing mechanisms can more effectively address two core challenges in class incremental learning: the catastrophic forgetting problem and the class imbalance issue.

In addition, to clearly demonstrate the performance of each learning session, Fig.7 (a) - (c) show the accuracy curves of three protocols for the CIFAR100 dataset. Fig.7 (a) presents the performance of each method in the long-term incremental learning process. Compared to other methods, our approach exhibits less performance degradation, clearly demonstrating its superiority in mitigating the forgetting of outdated models during the multi-stage learning process. Unlike Fig.7 (a), Fig.7 (c) shows another extreme performance scenario: retaining only 1000 samples from 50 learned old classes (20 samples per class), while having to learn a large amount of new data from 50 new classes in a single session (500 samples per class), resulting in a severe imbalance between old and new data. Under these conditions, our method still exhibits the lowest performance degradation, demonstrating its ability to solve class imbalance problems.

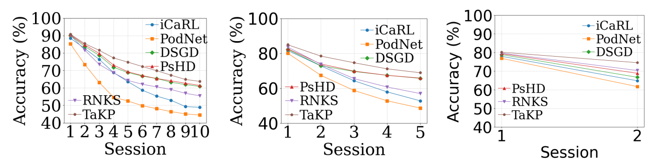
Method	Type	CIFAR100				ImageNet100			
		B0 Inc10		B0 Inc20		B0 Inc10		B0 Inc20	
		last	avg	last	avg	last	avg	last	avg
iCaRL	logits	48.87	64.05	52.78	65.69	50.64	65.33	60.25	71.49
Podnet	feature	44.43	55.10	48.64	61.76	44.62	62.85	57.66	70.58
Foster	logits	62.58	72.86	63.55	71.37	60.18	68.96	67.58	74.37
MTD	feature	46.97	57.94	49.52	62.93	46.85	65.22	58.74	73.12
DSGD	relation	60.87	70.56	65.66	71.08	64.28	74.04	71.06	75.83
PsHD	topology	61.43	71.52	65.81	71.59	64.17	73.94	67.93	74.08
LKD	logits	45.35	56.68	50.62	62.87	46.10	64.13	58.54	71.37
CREATE	logits	62.33	73.84	68.46	77.32	64.56	74.95	72.16	78.89
CLAD	-	52.38	63.41	54.69	64.98	47.44	64.27	59.29	73.18
RNKS	-	55.49	66.85	56.94	67.89	-	-	-	-
DBL	-	59.46	68.73	54.28	65.45	60.87	71.53	55.67	66.24
TaKP	topology	**63.76**	**74.32**	**68.86**	**77.71**	**64.73**	**75.26**	**72.46**	**79.15**

Table 1: Performance comparison on CIFAR-100 and ImageNet-100 datasets (Accuracy %).



(a) Distillation CIL methods (b) Batch sizes

Figure 6: Performance comparison of different distillation CIL methods and different batch sizes



(a) B0 Inc10 (b) B0 Inc20 (c) B0 Inc50

Figure 7: Accuracy of different methods on CIFAR-100.

Different distillation methods applied on CIL. The methods shown in Fig.6(a) use different distillation methods in CIL. TaKP performs the best, with the highest accuracy (ACC=74.32) and lowest backward transfer (BWT=14.67), showing that topological distillation effectively preserves old knowledge while maintaining high performance. Podnet has the lowest performance (ACC=55.1, BWT=32.58), highlighting the limitations of feature distillation. WKD (Lv, Yang, and Li 2024), using class relation distillation, achieves ACC=71.89 and BWT=20.43, showing promise in preserving class relations. Overall, topological distillation methods like TaKP and PsHD strike the best balance, effectively retaining old knowledge while achieving high accuracy, making them the most effective at mitigating CF.

Ablation study

Number of points in PCD. Fig.6(b) demonstrates that batch size setting plays a crucial role in model performance for topological feature learning based on point cloud data (PCD). As the batch size progressively increases from 4 to 128, we observe significant variations in the model's learning efficiency of the underlying data manifold. Specifically, smaller batch sizes (e.g., 4) prevent the model from adequately capturing global topological structures, causing topological distillation loss to interfere with the training process. When the batch size exceeds 8, however, the model demonstrates more comprehensive learning capability for topological embedding features, including both distributional and geometric characteristics. This phenomenon confirms the importance of sufficient data volume for accurately approximating data manifold structures—larger batches can effectively reduce noise interference and provide more complete topological representations.

Dimension of PD. The k-dimensional PD is based on the generation and disappearance of k-dimensional holes (such as connected components, loops, etc.) in PCD. Each dimension PD exhibits unique topological features. Therefore, we conducted ablation experiments on homology dimension of PDs. As 0-dim and 1-dim PD are generally used due to the computational cost, we considered only those. Specifically, 0-dim PDs identify connected components, and 1-dim PDs detect loops (Kim et al. 2024). As shown in Table 2, matching 0-dim PDs achieves the best performance, which can be attributed to the clustering information embedded in 0-dim PDs, making them more suitable for classification tasks. Although combining 0-dim and 1-dim PDs outperforms using 1-dim PDs alone, it still lags behind using 0-dim PDs exclusively. This is because when generating PIs from both 0-dim and 1-dim PDs, the birth times of all 0-dim homology classes are set to 0, causing their features to be concentrated in a small region of the image. As a result, the extraction of clustering information in the latent space is hindered.

Dimension	CIFAR100		ImageNet100	
	last	avg	last	avg
0	**63.76**	**74.32**	**64.73**	**75.26**
1	63.21	73.67	64.35	74.68
0&1	63.42	73.83	64.51	74.84

Table 2: Ablation results on the homology dimension of persistent diagrams on B0 Inc10 protocol.

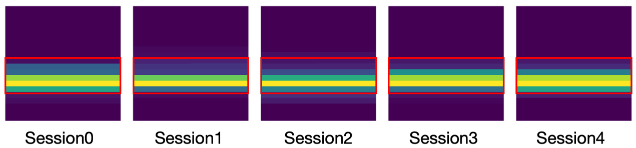


Figure 8: Visualization of PIs generated from the same memory samples (batch size=64) at five incremental sessions.

Quantitative analysis on component effectiveness. As shown in Table 3, we quantitatively evaluated the contributions of each component in persistent homology distillation on two datasets. The benchmark model is based on iCaR and PodNet, where iCaR is the traditional CIL method based on logits distillation, while PodNet is the CIL method based on feature distillation. $\mathcal{L}_{TopKD}$ and rebalance represent topological information distillation loss and dual-branch class rebalancing, respectively. The consistent improvements of TaKP demonstrate the efficacy of the proposed components.

Visualization of the Effectiveness of Topology-aware Distillation. To analyze the stability of topological representations during incremental learning, we conduct a systematic visualization of PIs generated at different sessions. As demonstrated in Fig.8, the PIs derived from identical memory samples exhibit remarkable structural consistency across incremental phases. This preservation of topological features confirms the effectiveness of our method in maintaining stable latent representations. Furthermore, Fig.9 presents a comparative visualization between initial and final session PIs, revealing high structural similarity that underscores the model's robustness against CF. The PH patterns remain largely invariant despite the sequential learning of new classes, suggesting successful knowledge consolidation in the topological embedding space.

Efficiency Analysis

Computation Efficiency. A newly topology-based distillation method PsHD (Fan et al. 2024a) which is similar to our method. A limitation of PsHD is that the computation of the persistence diagram relies on the Guidh package, which operates on the CPU. We evaluate the model parameters , training time and Memory Size between our methods and PsHD in Table 4. Compared to PsHD, our total training time is similar, but it includes an additional 1.4 hours for RipsNet training. However, RipsNet only needs to be pretrained once and does not require retraining during subsequent tasks. Although our method requires loading RipsNet,

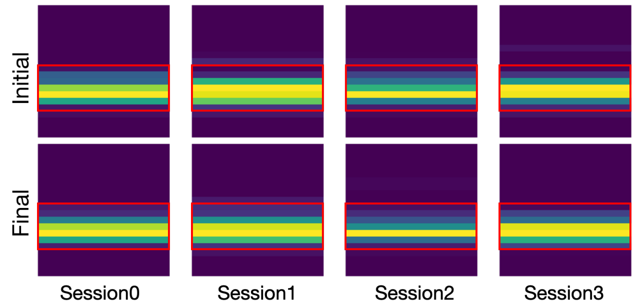


Figure 9: Comparative visualization of initial vs final session PIS. Inital: PIs generated by the initial model. Final: PIs from the model after 4 incremental sessions.

Method	$\mathcal{L}_{TopKD}$	Rebalance	CIFAR100		ImageNet100	
			last	avg	last	avg
iCaRL			48.87	64.05	50.64	65.33
	✓		61.45	71.83	62.36	73.28
	✓	✓	**62.90**	**73.59**	**64.58**	**75.02**
PodNet			44.43	55.10	44.62	62.85
	✓		56.16	67.25	55.31	65.57
	✓	✓	**57.31**	**68.47**	**56.78**	**66.82**

Table 3: Ablation study on B0 Inc10 protocol.

Methods	#P	Train	Memory Size	Acc(%)
PsHD	4.6	8.1	22.1	71.52
TaKP	4.8	6.4(+1.3)	46.7	74.32

Table 4: Effectiveness and efficiency comparison on CIFAR100 B0 Inc10 protocol.

which results in a higher memory size, it excels in both minimizing training time and maximizing accuracy. We adopt the learnable network RipsNet to approximate the persistence diagram and further accelerate the computation. But our approach requires an additional network to compute persistence images, which introduces extra computational overhead. We plan to explore acceleration algorithms in future work to further improve computational efficiency.

Conclusion

In this work, we present TaKP, a unified framework for class incremental learning that addresses the lack of structural awareness in knowledge preservation. At the heart of TaKP is a topology-aware distillation strategy based on persistent homology, which captures and preserves the global manifold structure of the feature space across incremental stages. To further support knowledge preservation under data imbalance, TaKP incorporates a dual-branch architecture with adaptive rebalancing, decoupling general representation learning from old-class preservation. Through extensive experiments on standard CIL benchmarks, we demonstrate that TaKP significantly improves performance and old-class preservation over state-of-the-art methods.

Acknowledgments

This work is supported in part by the Natural Science Foundation of Hebei Province under Grant F2025202039 and F2024202047, in part by the Natural Science Foundation of Tianjin (No. 25JCYBJC01020), in part by the National Natural Science Foundation of China under Grants 62476195, 92570118 and 62376088, and in part by the Hebei Yanzhao Golden Platform Talent Gathering Programme Core Talent Project (Education Platform) (HJZD202509).

References

Cang, Z.; Mu, L.; and Wei, G. 2018. Representability of algebraic topology for biomolecules in machine learning based scoring and virtual screening. *PLoS Comput. Biol.*, 14(1).

Cha, S.; Cho, S.; Hwang, D.; Hong, S.; Lee, M.; and Moon, T. 2023. Rebalancing Batch Normalization for Exemplar-Based Class-Incremental Learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, 20127–20136. IEEE.

Chen, H.; Wang, Y.; Fan, Y.; Jiang, G.; and Hu, Q. 2025. Reducing Class-wise Confusion for Incremental Learning with Disentangled Manifolds. In *CVPR-25,*. CVPR Press.

Chen, H.; Wang, Y.; and Hu, Q. 2023. Multi-Granularity Regularized Re-Balancing for Class Incremental Learning. *IEEE Trans. Knowl. Data Eng.*, 35(7): 7263–7277.

Clough, J. R.; Byrne, N.; Öksüz, I.; Zimmer, V. A.; Schnabel, J. A.; and King, A. P. 2022. A Topological Loss Function for Deep-Learning Based Image Segmentation Using Persistent Homology. *IEEE Trans. Pattern Anal. Mach. Intell.*, 44(12): 8766–8778.

Deng, J.; Dong, W.; Socher, R.; Li, L.; Li, K.; and Fei-Fei, L. 2009. ImageNet: A large-scale hierarchical image database. In *2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2009), 20-25 June 2009, Miami, Florida, USA*, 248–255. IEEE Computer Society.

Dong, J.; Liang, W.; Cong, Y.; and Sun, G. 2023. Heterogeneous Forgetting Compensation for Class-Incremental Learning. In *IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023*, 11708–11717. IEEE.

Douillard, A.; Cord, M.; Ollion, C.; Robert, T.; and Valle, E. 2020. PODNet: Pooled Outputs Distillation for Small-Tasks Incremental Learning. In Vedaldi, A.; Bischof, H.; Brox, T.; and Frahm, J., eds., *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part XX*, volume 12365 of *Lecture Notes in Computer Science*, 86–102. Springer.

Du, S.; Lao, Q.; Kang, Q.; Li, Y.; Jiang, Z.; Zhao, Y.; and Li, K. 2022. Distilling Knowledge from Topological Representations for Pathological Complete Response Prediction. In Wang, L.; Dou, Q.; Fletcher, P. T.; Speidel, S.; and Li, S., eds., *Medical Image Computing and Computer Assisted Intervention - MICCAI 2022 - 25th International Conference, Singapore, September 18-22, 2022, Proceedings, Part II*, volume 13432 of *Lecture Notes in Computer Science*, 56–65. Springer.

Fan, Y.; Wang, Y.; Zhu, P.; Chen, D.; and Hu, Q. 2024a. Persistence Homology Distillation for Semi-supervised Continual Learning. In Globersons, A.; Mackey, L.; Belgrave, D.; Fan, A.; Paquet, U.; Tomczak, J. M.; and Zhang, C., eds., *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*.

Fan, Y.; Wang, Y.; Zhu, P.; and Hu, Q. 2024b. Dynamic Sub-graph Distillation for Robust Semi-supervised Continual Learning. In Wooldridge, M. J.; Dy, J. G.; and Natarajan, S., eds., *Thirty-Eighth AAAI Conference on Artificial Intelligence, AAAI 2024, Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence, IAAI 2024, Fourteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2014, February 20-27, 2024, Vancouver, Canada*, 11927–11935. AAAI Press.

Gao, Z.; Han, S.; Zhang, X.; Xu, K.; Zhou, D.; Mao, X.; Dou, Y.; and Wang, H. 2025. Maintaining Fairness in Logit-based Knowledge Distillation for Class-Incremental Learning. In Walsh, T.; Shah, J.; and Kolter, Z., eds., *AAAI-25, Sponsored by the Association for the Advancement of Artificial Intelligence, February 25 - March 4, 2025, Philadelphia, PA, USA*, 16763–16771. AAAI Press.

He, H.; Wang, J.; Wei, P.; Xu, F.; Ji, X.; Liu, C.; and Chen, J. 2023. TopoSeg: Topology-Aware Nuclear Instance Segmentation. In *IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023*, 21250–21259. IEEE.

Hofer, C. D.; Graf, F.; Rieck, B.; Niethammer, M.; and Kwitt, R. 2020. Graph Filtration Learning. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, 4314–4323. PMLR.

Jeon, E. S.; Choi, H.; Shukla, A.; Wang, Y.; Lee, H.; Buman, M. P.; and Turaga, P. K. 2024. Topological persistence guided knowledge distillation for wearable sensor data. *Eng. Appl. Artif. Intell.*, 130: 107719.

Kang, M.; Park, J.; and Han, B. 2022. Class-Incremental Learning by Knowledge Distillation with Adaptive Feature Consolidation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, 16050–16059. IEEE.

Kang, Z.; Fini, E.; Nabi, M.; Ricci, E.; and Alahari, K. 2023. A soft nearest-neighbor framework for continual semi-supervised learning. In *IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023*, 11834–11843. IEEE.

Kim, J.; You, J.; Lee, D.; Kim, H. Y.; and Jung, J. 2024. Do Topological Characteristics Help in Knowledge Distillation? In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*. OpenReview.net.

Kirkpatrick, J.; Pascanu, R.; and Rabinowitz, N. 2017. Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences*, 114(13): 3521–3526.

Krizhevsky, A.; and Hinton, G. 2019. Learning multiple layers of features from tiny images.

Liu, M.; Dey, T. K.; and Gleich, D. F. 2023. Topological structure of complex predictions. *Nat. Mac. Intell.*, 5(12): 1382–1389.

Lv, J.; Yang, H.; and Li, P. 2024. Wasserstein Distance Rivals Kullback-Leibler Divergence for Knowledge Distillation. In Globersons, A.; Mackey, L.; Belgrave, D.; Fan, A.; Paquet, U.; Tomczak, J. M.; and Zhang, C., eds., *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*.

Madhu, H.; and Chepuri, S. P. 2023. TopoSRL: Topology preserving self-supervised Simplicial Representation Learning. In Oh, A.; Naumann, T.; Globerson, A.; Saenko, K.; Hardt, M.; and Levine, S., eds., *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Masana, M.; Liu, X.; Twardowski, B.; Menta, M.; Bagdanov, A. D.; and van de Weijer, J. 2023. Class-Incremental Learning: Survey and Performance Evaluation on Image Classification. *IEEE Trans. Pattern Anal. Mach. Intell.*, 45(5): 5513–5533.

McDonnell, M. D.; Gong, D.; Parvaneh, A.; Abbasnejad, E.; and van den Hengel, A. 2023. RanPAC: Random Projections and Pre-trained Models for Continual Learning. In Oh, A.; Naumann, T.; Globerson, A.; Saenko, K.; Hardt, M.; and Levine, S., eds., *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Melas-Kyriazi, L. 2020. The Mathematical Foundations of Manifold Learning. *CoRR*, abs/2011.01307.

Rebuffi, S.-A.; Kolesnikov, A.; Sperl, G.; and Lampert, C. H. 2017. iCaRL: Incremental Classifier and Representation Learning. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 5533–5542.

Rypesc, G.; Cygert, S.; Khan, V.; Trzcinski, T.; Zielinski, B.; and Twardowski, B. 2024. Divide and not forget: Ensemble of selectively trained experts in Continual Learning. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

Song, J.; Chen, J.; and Du, L. 2024. Rebalancing network with knowledge stability for class incremental learning. *Pattern Recognit.*, 153: 110506.

Wang, F.-Y.; Zhou, D.-W.; Ye, H.-J.; and Zhan, D.-C. 2022. FOSTER: Feature Boosting and Compression for Class-Incremental Learning. In Avidan, S.; Brostow, G.; Cissé, M.; Farinella, G. M.; and Hassner, T., eds., *Computer Vision – ECCV 2022*, 398–414. Cham: Springer Nature Switzerland. ISBN 978-3-031-19806-9.

Wang, L.; Zhang, X.; Su, H.; and Zhu, J. 2024. A Comprehensive Survey of Continual Learning: Theory, Method and Application. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(8): 5362–5383.

Wen, H.; Pan, L.; Dai, Y.; Qiu, H.; Wang, L.; Wu, Q.; and Li, H. 2024. Class Incremental Learning with Multi-Teacher Distillation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, 28443–28452. IEEE.

Xu, S.; Meng, G.; Nie, X.; Ni, B.; Fan, B.; and Xiang, S. 2024. Defying Imbalanced Forgetting in Class Incremental Learning. In Wooldridge, M. J.; Dy, J. G.; and Natarajan, S., eds., *Thirty-Eighth AAAI Conference on Artificial Intelligence, AAAI 2024, Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence, IAAI 2024, Fourteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2014, February 20-27, 2024, Vancouver, Canada*, 16211–16219. AAAI Press.

Yan, S.; Xie, J.; and He, X. 2021. DER: Dynamically Expandable Representation for Class Incremental Learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 3014–3023.

Yan, Z.; Ma, T.; Gao, L.; Tang, Z.; and Chen, C. 2021. Link Prediction with Persistent Homology: An Interactive View. In Meila, M.; and Zhang, T., eds., *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, 11659–11669. PMLR.

Zhao, Q.; and Wang, Y. 2019. Learning metrics for persistence-based summaries and applications for graph classification. In Wallach, H. M.; Larochelle, H.; Beygelzimer, A.; d'Alché-Buc, F.; Fox, E. B.; and Garnett, R., eds., *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, 9855–9866.

Zhao, Q.; Ye, Z.; Chen, C.; and Wang, Y. 2020. Persistence Enhanced Graph Neural Network. In Chiappa, S.; and Calandra, R., eds., *The 23rd International Conference on Artificial Intelligence and Statistics, AISTATS 2020, 26-28 August 2020, Online [Palermo, Sicily, Italy]*, volume 108 of *Proceedings of Machine Learning Research*, 2896–2906. PMLR.

Zhi, R.; Meng, Y.; Hou, J.; and Wan, J. 2025. Dual Balanced Class-Incremental Learning With im-Softmax and Angular Rectification. *IEEE Trans. Neural Networks Learn. Syst.*, 36(3): 4437–4447.

Zhou, D.; Wang, Q.; Ye, H.; and Zhan, D. 2023. A Model or 603 Exemplars: Towards Memory-Efficient Class-Incremental Learning. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.