Breaking the Model Forgetting Cycle in Long-Incremental 3D Object Detection

Peisheng Qian[1,2], Jie Xu[1], Xulei Yang[2,*], and Na Zhao[1,*]

[1] Singapore University of Technology and Design,
{jie_xu2, na_zhao}@sutd.edu.sg
[2] Institute for Infocomm Research, A*STAR, Singapore,
{qian_peisheng, yang_xulei}@a-star.edu.sg

Abstract. Incremental 3D object detection requires a detector to learn novel object classes while remembering previously learned ones over sequentially arriving data. Previous methods, primarily based on pseudo-labeling, perform reasonably in short-incremental stages but still suffer from severe model forgetting when dealing with long-incremental sequences. We investigate this failure and reveal a detrimental self-reinforcing cycle: data distribution shift of novel classes causes model forgetting on old classes, which further produces accumulated error in pseudo-labeling that exacerbates model degradation. To address this issue, we draw inspiration from the human learning process and propose the *Learning-Dynamics-driven Memory and Review* (LDMR) framework. LDMR monitors per-class detection quality at periodic training checkpoints and uses these learning-dynamics signals to drive two innovative mechanisms, namely (i) *human-like intra-stage review* that divides each incremental stage into multiple sub-stages' training and concentrates on remembering the most-forgotten objects, and (ii) *scene-aware cross-stage memory evolution* that evolves a memory bank to transfer knowledge between two consecutive stages by jointly considering scene learnability and diversity. Extensive experiments across multiple long-incremental protocols on indoor benchmarks SUN RGB-D and ScanNetV2 show that LDMR substantially mitigates the model forgetting and outperforms all baselines by a clear margin. Code is available at `https://github.com/qianpeisheng/LDMR`.

Keywords: 3D object detection · Class-incremental learning · Indoor scene understanding · Model forgetting

1 Introduction

Indoor 3D object detection [4, 7, 8, 30, 39–41, 44] is a core capability for many applications such as augmented reality, embodied agents, and assistive robotics, where a system must localize and recognize objects from RGB-D scans or point

* Corresponding authors.

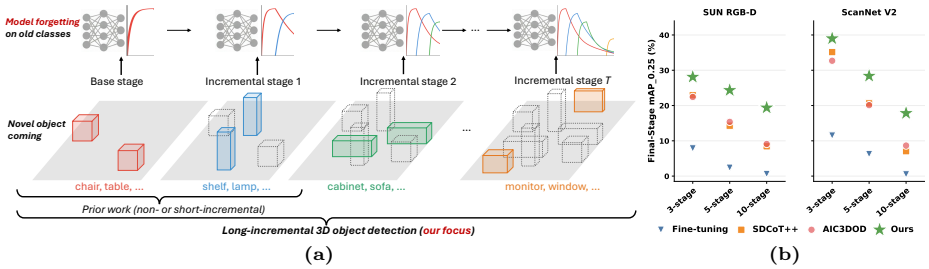


Fig. 1: Task illustration and result comparison. (a) Long-incremental 3D object detection has the continuous incremental learning stages, where novel class annotations are available at the current stage; objects from previously learned old classes are unlabeled in the later stages. (b) Performance comparison over multiple long-incremental 3D detection tasks, showing our method consistently outperforms baseline approaches.

clouds. Recently, modern 3D detectors have advanced rapidly due to deep learning [19, 22, 24], and their training processes typically assume a closed-set label space and a stationary training distribution. In realistic deployments, however, new object classes and new environments usually arrive over time, and repeatedly retraining from scratch is often impractical [33–35, 40, 41].

To address this limitation, *Incremental 3D Object Detection* (I3DOD) has been proposed and has attracted increasing attention [40, 41]. As shown in Fig. 1a, I3DOD updates the detector over a sequence of incremental stages, each introducing novel classes while requiring the model to remember competence on all previously learned classes. To achieve this, previous work mainly leverages pseudo-labeling strategies [4, 40, 41] to incrementally train detector models. For example, SDCoT [40] introduces a teacher-student I3DOD framework in which a static teacher (the previous-stage model) generates pseudo labels for old classes, while a dynamic teacher (an exponential moving average of models) transfers underlying knowledge from the new data to the student network via consistency regularization. Building on this, SDCoT++ [41] consolidates pseudo labels from both teachers and calibrates class probabilities according to object occurrence frequencies, yielding more balanced and complete pseudo annotations that better handle the class imbalance inherent in 3D detection datasets. More recently, AIC3DOD [4] advances the detection backbone with a point transformer architecture for stronger feature representation and higher-quality pseudo labels, and further incorporates room layout constraints across incremental stages.

Although previous I3DOD methods have achieved remarkable progress, they were typically designed for short-incremental stages, *e.g.*, one or two incremental stages [4, 40, 41]. When handling the 3D detection tasks with *long-incremental* stages, these methods still fail to balance learning novel classes and remembering old knowledge, leading to severe model forgetting and performance degradation on old classes (see results in Fig. 1b and analysis in Sec. 3.2). Handling many incremental stages is the more realistic and practically relevant setting: embodied agents or robotic assistants deployed over long periods will continuously encounter new object categories rather than just one or two updates. Motivated by this

real-world requirement, we introduce *Long-Incremental 3D Object Detection* (L-I3DOD), a formulation that studies 3D detection under prolonged streams of category increments and aims to enable lifelong learning for practical 3D perception systems.

In this paper, we investigate the model forgetting by theoretical and empirical analysis, and reveal a self-reinforcing cycle that raises the crucial challenge for L-I3DOD. Specifically, class distribution shifts between novel and old stages produce optimization gradient misalignment, which increases the model's empirical risk on old-stage classes and leads to the model degradation on old-classes. Moreover, the pseudo-labeling strategies conducted on old-classes inevitably introduce error that will accumulate and hinder the model's training over all incremental stages. Motivated by the human's learning process of "learn–evaluate–review", *i.e.*, the learned knowledge is evaluated by tests to identify which parts have been forgotten and then these forgotten parts are reviewed again, we propose an innovative *Learning-Dynamics-driven Memory and Review* (LDMR) framework to mitigate the model forgetting in long-incremental 3D object detection. Concretely, we define the learning dynamics that monitors per-class detection quality at periodic training checkpoints, and LDMR uses these learning-dynamics signals to drive two iterative processes: *intra-stage review* and *cross-stage memorization*. The intra-stage review process divides each incremental stage into multiple sub-stages' training and leverages the human-like learn–evaluate–review mechanism to remember the easily forgotten objects. The cross-stage memorization process focuses on transferring knowledge between two consecutive stages by evolving a memory bank that jointly considers scene-aware learnability and diversity.

In summary, the contributions of this work are as follows:

- This study aims to address the model forgetting challenge for L-I3DOD, where we deeply investigate the crucial issues and propose an innovative L-I3DOD methodology motivated by the natural human learning mechanism.
- We present a learning-dynamics-driven framework that unifies the human-like intra-stage review and scene-aware cross-stage memory evolution, under meaningful measurable retention signals with a limited memory budget.
- Extensive experiments, on SUN RGB-D and ScanNetV2 across different settings of 3/5/10-incremental stages, demonstrate our method can mitigate the model forgetting and achieve consistent improvements over prior methods.

2 Related Work

2.1 3D Object Detection

3D object detection predicts oriented 3D bounding boxes from point clouds reconstructed from RGB-D scans or captured by laser scanners [1, 10–12, 21, 25]. Representative indoor pipelines include vote-based proposal generation [19], transformer-style set prediction with object queries [17], sparse-voxel two-stage detectors that refine proposals with geometric aggregation [29], and proposal-free transformers that directly aggregate global context [16]. Recent work further

focuses on improving 3D model efficiency in feature aggregation for cluttered indoor scenes [22] or exploring 2D foundation model based 3D detection [8]. In this work, we adopt TR3D [22] and VoteNet [19] as backbone detectors.

Since 3D box annotation for indoor scenes is expensive, a growing line of work studies label-efficient indoor detection, including weakly-supervised learning [38], semi-supervised teacher–student training [39], and sparse-supervised learning via class-level prototypes [43]. Active learning has also been introduced to select informative scenes for annotation under a fixed labeling budget [30]. Our cross-stage memory evolution (Sec. 4.3) shares the spirit of selecting high-value scenes under a budget, but targets the incremental learning setting.

2.2 Class-Incremental Learning

Class-incremental learning (CIL) studies continual updates where new classes arrive over time while performance on previously learned classes must be preserved. Common approaches include parameter-importance regularization (e.g., EWC-style constraints) [9], distillation-based retention (e.g., LwF-style feature/logit matching) [13], and rehearsal via exemplar memory [20]. More recent directions emphasize parameter-efficient adaptation such as prompt- and adapter-based continual learning [2,23,27,31,32]; however, these methods are primarily designed for image classification and do not address the localization objectives and missing-annotation issues that arise in detection.

Class-incremental object detection is particularly challenging due to dense foreground-background imbalance, structured localization objectives, and missing annotations for previously learned classes, where unlabeled old objects can be mistakenly optimized as background [14, 15]. Prior incremental 2D detection methods commonly rely on knowledge distillation and replay strategies to address this [6, 14, 15, 18, 26]. These 2D strategies, however, do not transfer directly to 3D indoor detection, where old-class and novel-class objects physically co-exist within the same scenes, intensifying the supervision conflict. Notably, despite rehearsal being well-established in CIL [20, 42], no prior work has explored memory-bank-based replay for incremental 3D object detection.

2.3 Incremental 3D Object Detection

Incremental 3D object detection (I3DOD) incrementally introduces novel categories while old-stage data is unavailable; in practice, old objects may still appear but remain unlabeled, creating stage-wise supervision conflicts [36]. Most existing methods adopt teacher-student training with pseudo labels and distillation. For example, SDCoT proposes static-dynamic co-teaching, where a frozen teacher transfers old-class knowledge and a dynamic teacher stabilizes learning on new-stage data [40]. SDCoT++ improves co-teaching by consolidating teacher pseudo labels and calibrating class scores to mitigate imbalance and missing old objects [41]. CoCo-Teach provides another co-teaching variant [3]. AIC3DOD incorporates transformer-based representations and room-layout constraints to regularize incremental training under incomplete annotations [4]. Across all these

methods, pseudo-labeling is the most effective component for retaining old-class knowledge. In long-incremental settings, however, even pseudo labels become unreliable as errors compound over stages (Sec. 3.2).

Existing I3DOD methods [4, 37, 40, 41] mitigate forgetting through pseudo-labels or regularization techniques, without replaying real annotated old-class scenes. We draw on the fact that humans' learning relies on note-taking and use memory bank mechanism in long-incremental detection tasks. Note that memory bank is a common practice for incremental learning [15, 20], but as of yet, no specific methods have been proposed for I3DOD domain. What is more crucial is how to define the memory bank and the update rules for I3DOD because some information in 3D scenes might be misleading or redundant for incremental learning. In this work, we jointly consider the scene diversity and learnability to establish an innovative scene-aware cross-stage memory evolution method. Moreover, we propose the human-like intra-stage review and show that they mitigate model forgetting over long-incremental 3D detection tasks.

3 Background and Analysis

3.1 Problem Definition

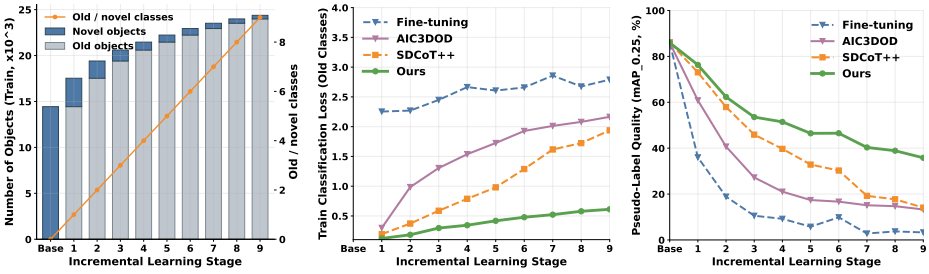
Notation. Let $\mathcal{C}$ be the full class set, partitioned into T disjoint groups $\{\mathcal{C}_t\}_{t=1}^{T}$, *i.e.*, $\mathcal{C}_i \cap \mathcal{C}_j = \emptyset$ for $i \neq j$ and $\bigcup_{t=1}^{T} \mathcal{C}_t = \mathcal{C}$. We use $\mathcal{C}_{\leq t}$ and $\mathcal{C}_{<t}$ to denote the seen classes up to and before stage t, respectively. At each stage t, we call $\mathcal{C}_t$ the *novel* classes and $\mathcal{C}_{<t}$ the *old* classes. The learner receives a stage-specific dataset $\mathcal{D}_t = \{\mathcal{X}_t, \mathcal{Y}_t\}$, where $x \in \mathcal{X}_t$ is an indoor 3D scene and $y_t \in \mathcal{Y}_t$ contains 3D object detection annotations only from the novel classes $\mathcal{C}_t$.

Incremental 3D object detection (I3DOD). Training of I3DOD proceeds over stages $t = 1, \ldots, T$. It trains f_{θ_1} on $\mathcal{D}_1$ and then for $t \geq 2$ updates $f_{\theta_{t-1}}$ to f_{θ_t} using only $\mathcal{D}_t$, where old-class instances may appear but are unlabeled. After stage t, the detector must detect all seen classes $\mathcal{C}_{\leq t}$. Prior work [4, 40, 41] typically consider short-incremental stages' I3DOD tasks (*e.g.*, T=2 or T=3). In this study, we focus on the long-incremental stages' I3DOD tasks (*e.g.*, $T \geq 3$).

3.2 Theoretical and Empirical Analysis of Model Forgetting

Firstly, we conduct a pilot study to diagnose the model forgetting in long-incremental 3D object detection (L-I3DOD). As shown in Fig. 2, we observe that (i) distribution shift and (ii) error accumulation tend to be the crucial factors causing the heavy model forgetting for L-I3DOD tasks.

I) Distribution shift issue: training on the shifted data distribution of novel classes leads to the model degradation on old classes. Firstly, we give the following theoretical analysis to illustrate this. Let stages $t = 1, \ldots, T$ have data distributions P_t on sample space $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$. Let the model parameter be $\theta \in \mathbb{R}^d$ and define the stage-t expected detection loss as $\mathcal{L}_t(\theta) = \mathbb{E}_{z \sim P_t}\big[\ell(f_\theta, z)\big]$, and the expected gradient is $g_t(\theta) = \nabla_\theta \mathcal{L}_t(\theta) = \mathbb{E}_{z \sim P_t}[\nabla_\theta \ell(f_\theta, z)]$. We leverage the



(a) Training data variation. **(b)** Evaluated training loss. **(c)** Pseudo-label quality.

Fig. 2: Empirical analysis of L-I3DOD on SUN RGB-D (40 classes). (a) Training data variation indicates the distribution shift over long-incremental stages. (b) Training loss or empirical risk evaluated on old classes at each stage. (c) The quality of pseudo-labels generated by each stage's model. The results reveal that the vanilla Fine-tuning strategy and the I3DOD methods SDCoT++ [41] and AIC3DOD [4] still face the heavy model forgetting in L-I3DOD tasks, while our method achieves significant advancement.

Wasserstein distribution distance to establish the long-incremental stages' forgetting upper bound as follows (the proof is given in the supplementary material):

Theorem 1 (Long-incremental stages' forgetting upper bound). *Let $i \leq s$ be an index of a previously trained stage. Perform gradient-descent updates at stages $u = s+1, \ldots, t$ with step sizes $\{\eta_u\}$ and we have $\theta_u = \theta_{u-1} - \eta_u\, g_u(\theta_{u-1})$, where the stage-$u$ population gradient is $g_u(\theta_{u-1}) = \mathbb{E}_{z\sim P_u}\left[\nabla_\theta \ell(f_\theta, z)\right]\Big|_{\theta=\theta_{u-1}}$. Under the regularity conditions stated in the proof, the cumulative forgetting (increase of expected loss for stage i) after updates up to stage t satisfies*

$$\mathcal{L}_i(\theta_t) - \mathcal{L}_i(\theta_s) \leq \sum_{u=s+1}^{t} \left(\eta_u K\, W_1(P_i, P_u)\|g_u(\theta_{u-1})\| + \tfrac{L\eta_u^2}{2}\|g_u(\theta_{u-1})\|^2\right) \quad (1)$$

where $W_1(P_i, P_u)$ is the 1-Wasserstein distance on $\mathcal{Z}$, K is the directional Lipschitz constant and L is the smoothness constant in parameter space, and each $\|g_u(\theta_{u-1})\|$ is the Euclidean norm of the population gradient evaluated at θ_{u-1}.

This bound explains why long-incremental stage 3DOD forgets: distribution shifts between two stages (measured by $W_1(P_i, P_u)$) and large new-stage gradients $\|g_u(\theta_{u-1})\|$ produce gradient misalignment that increases old-stage loss, while the input-gradient sensitivity (measured by K), optimization curvature smoothness (measured by L), and step size (η_u) further amplify this effect.

Secondly, the empirical evidence in Fig. 2 aligns our theoretical analysis, which suggests that the difference in quantity between novel-class objects and old-class objects becomes large during the long-incremental training process (Fig. 2a); supervised learning on the minority of novel-class data results in increasing the model empirical risk (training loss) on the majority of old-class data distribution (Fig. 2b), which indicates the occurrence of model forgetting in L-I3DOD.

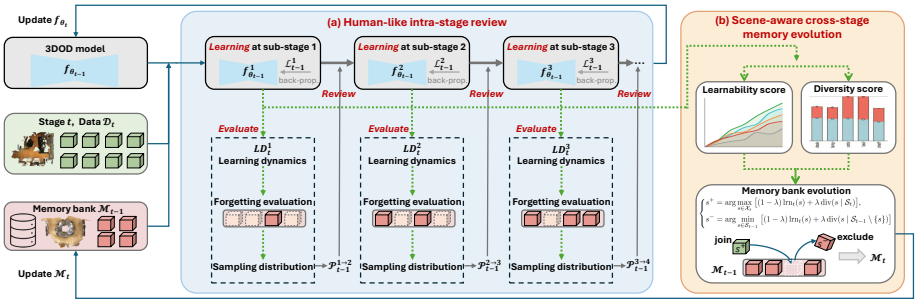


Fig. 3: Method overview. Given the t-th incremental stage data $\mathcal{D}_t$, our LDMR leverages the model $f_{\theta_{t-1}}$ and the memory bank $\mathcal{M}_{t-1}$ inherited from the last $(t-1)$-th stage, to conduct (a) intra-stage human-like review which simulates the learn–evaluate-review process by which humans overcome forgetting, to obtain the new model f_{θ_t}. Then, LDMR performs (b) scene-aware cross-stage memory evolution which establishes the learnability and diversity scores for achieving the cross-stage knowledge transfer, by updating the memory bank $\mathcal{M}_t$.

II) Error accumulation issue: the error from pseudo-labeling strategies will accumulate and hinder the model over long-incremental stages. To enhance remembering old classes, pseudo-labeling strategies [4,40,41] are widely adopted for re-labeling the old classes during incremental 3D object detection. However, the signals generated by pseudo-labeling may inherently contain errors, and as the number of incremental stages increases, the accumulated errors become significant and the pseudo-label quality for the old classes will further deteriorate. Taking the representative pseudo-labeling based I3DOD method SDCoT++ [41] as an example, one can observe that the trained model has the gradually increasing empirical risk on old classes (Fig. 2b) and the decreasing pseudo-label quality (Fig. 2c). As a result, these two issues lead to a self-reinforcing cycle: distribution shift degrades model performance on old classes and wrong pseudo labels accumulate errors over long-incremental stages, making existing methods insufficient to prevent model forgetting. This motivates us to study how to effectively address the model forgetting challenge so as to make I3DOD applicable in practical long-incremental scenarios.

4 Method

4.1 Overview

Motivated by the "learn-evaluate-review" process by which humans overcome forgetting, we propose the Learning-Dynamics-driven Memory and Review (LDMR) framework to address the model forgetting in long-incremental 3D detection. The framework is shown in Fig. 3. It consists of the *human-like intra-stage review* and *scene-aware cross-stage memory evolution*. The technical details are presented in the following sections.

4.2 Human-like Intra-stage Review

As Sec. 3.2 reveals, training on the t-th stage data $\mathcal{D}_t$ focuses on fitting novel classes $\mathcal{C}_t$ and will introduce model forgetting on old classes $\mathcal{C}_{<t}$ due to the distribution shift between $\mathcal{D}_t$ and $\mathcal{D}_{<t}$. To address this issue, we design the human-like intra-stage review that maximally learns novel classes while remembering old ones, given the model parameter $f_{\theta_{t-1}}$ and memory bank $\mathcal{M}_{t-1}$ of the last stage. First, we formally give the definition of memory bank $\mathcal{M}_{t-1}$ as follows:

Memory bank. The memory bank aims to support the knowledge transfer from the previous $t-1$ stages to the current t-th stage. Hence, the memory bank $\mathcal{M}_{t-1}$ preserves a small set of annotated scenes from historical data set $\mathcal{D}_{<t}$, by

$$\mathcal{M}_{t-1} = \{\mathcal{S}_{t-1}, \mathcal{Y}|_{\mathcal{S}_{t-1}}\}, \quad \mathcal{S}_{t-1} \subset \mathcal{X}_{<t}, \quad \mathcal{Y}|_{\mathcal{S}_{t-1}} \subset \mathcal{Y}_{<t}, \quad |\mathcal{S}_{t-1}| = B, \tag{2}$$

where B is a fixed-budget size, $\mathcal{X}_{<t}$ denotes the historical scenes and $\mathcal{Y}_{<t}$ includes the annotations from the previous stages. For scenes $\mathcal{S}_{t-1}$ in the memory bank, $\mathcal{Y}|_{\mathcal{S}_{t-1}}$ is the corresponding annotation set whose classes belong to $\mathcal{C}_{<t}$. $\mathcal{M}_0 = \emptyset$. Then, to simulate the human's learning process, we divide one incremental stage's model training $f_{\theta_{t-1}} \rightarrow f_{\theta_t}$ into I sub-stages as $f^1_{\theta_{t-1}}, f^2_{\theta_{t-1}}, f^3_{\theta_{t-1}}, ..., f^I_{\theta_{t-1}}$ (finally we have $f_{\theta_t} = f^I_{\theta_{t-1}}$), where $f^i_{\theta_{t-1}}$ is a sub-stage model checkpoint. Before and after the sub-stage's training from $f^i_{\theta_{t-1}}$ to $f^{i+1}_{\theta_{t-1}}$, we record the *learning dynamics* which is defined as follows:

Learning dynamics. To dynamically monitor the model's detection performance on old classes, the learning dynamics is defined as a set of the per-class recall performance on the memory bank $\mathcal{M}_{t-1}$:

$$LD^i_t = \{R^i_t(s,c)\}, \quad s \in \mathcal{S}_{t-1}, \quad c \in \mathcal{Y}|_{\mathcal{S}_{t-1}}, \tag{3}$$

where $R^i_t(s,c)$ is the recall of the class c in the scene s in the sub-stage i.

Forgetting evaluation. Furthermore, we evaluate the two learning dynamics LD^i_t and LD^{i+1}_t to check the forgetting knowledge on old classes from model $f^i_{\theta_{t-1}}$ to $f^{i+1}_{\theta_{t-1}}$. The per-class performance drop for each scene s and old class $c \in \mathcal{C}_{<t}$ can be computed by $\max(0,\, LD^i_t(s,c) - LD^{i+1}_t(s,c))$. In this way, we obtain the following scene-level forgetting metric:

$$F^{i \rightarrow i+1}_s = \sum_{c \in \mathcal{C}_{<t}} \log(1+n_{s,c}) \times \max\left(0,\, R^i_t(s,c) - R^{i+1}_t(s,c)\right), \tag{4}$$

where $\log(1+n_{s,c})$ is a log-scaled object-count weight and $n_{s,c}$ denotes the number of ground-truth instances of class c in scene s.

Scenes with larger $F^{i \rightarrow i+1}_s$ indicate that the model has a more severe forgetting of the objects in those scenes during the sub-stage's training. This guides us to carry out the following intra-stage review process.

Intra-stage review with sampling distribution. We leverage the scene-level forgetting metrics $\{F^{i \rightarrow i+1}_s\}$, $s \in \mathcal{M}_{t-1}$ to guide the model to emphasize the forgotten knowledge in the next sub-stage's training (*i.e.*, from $f^{i+1}_{\theta_{t-1}}$ to $f^{i+2}_{\theta_{t-1}}$).

To achieve this, we construct a unified sampling pool from the stage data and the memory bank $\mathcal{X}_t \cup \mathcal{S}_{t-1}$, where each scene is assigned an adaptive sampling weight. To be specific, the sampling weight of scenes in the current stage data $\mathcal{X}_t$ is fixed at $w_s^{i\to i+1} = 1$, and that in the memory bank $\mathcal{S}_{t-1}$ incorporates a forgetting-proportional boost by the following equation:

$$w_s^{i\to i+1} = 1 + \eta \cdot F_s^{i\to i+1} \cdot \mathbb{I}(s \in \mathcal{S}_{t-1}), \tag{5}$$

where $\eta \geq 0$ controls the strength of the review emphasis. $\mathbb{I}(\cdot)$ is an indicator function. Then, the probabilistic sampling distribution for the intra-stage review process of sub-stage i to sub-stage $i+1$ over the unified pool is given by

$$\mathcal{P}_{t-1}^{i\to i+1}(s) = \frac{w_s^{i\to i+1}}{\sum_{j\in\mathcal{D}_t\cup\mathcal{M}_{t-1}} w_j^{i\to i+1}}. \tag{6}$$

At the next sub-stage from $f_{\theta_{t-1}}^{i+1}$ to $f_{\theta_{t-1}}^{i+2}$, training scenes are drawn from $\mathcal{D}_t \cup \mathcal{M}_{t-1}$ according to $\mathcal{P}_{t-1}^{i\to i+1}(s)$ and the detector is optimized with:

$$\mathcal{L}_{t-1}^{i+1} = \mathbb{E}_{s\sim\mathcal{P}_{t-1}^{i\to i+1}(s)}\left[\mathcal{L}_{\text{det}}(s;\ \theta_{t-1})\right], \tag{7}$$

where $\mathcal{L}_{\text{det}} = \mathcal{L}_{\text{cls}} + \mathcal{L}_{\text{bbox}}$ combines the classification and bounding-box regression losses following prior 3D detection work [19, 22].

4.3 Scene-aware Cross-stage Memory Evolution

After stage-t training completes, we compute the learning dynamics $\{LD_t^i\}_{i=1}^I$ for both old and novel classes on the data set $\mathcal{D}_t \cup \mathcal{M}_{t-1}$, generalizing Eq. 3 to all classes:

$$LD_t^i = \{R_t^i(s,c)\}, \quad s \in \mathcal{X}_t \cup \mathcal{S}_{t-1}, \quad c \in \mathcal{C}_t \cup \mathcal{Y}|_{\mathcal{S}_{t-1}}, \tag{8}$$

which is leveraged to establish the scene-aware cross-stage memory evolution for updating the memory bank, *i.e.*, achieving $\mathcal{M}_{t-1} \to \mathcal{M}_t$.

The key insight is that the memory bank should prioritize the learnable and representative scenes, *i.e.*, those where the model has shown it can improve with exposure but has not yet retained that improvement, and the scene objects have the diversity to represent the entire data distribution. Therefore, we define the following *learnability* and *diversity* scores in our memory bank evolution strategy.

Learnability score. To quantify how learnable the scene's objects are, we define the scene-aware learnability score for $s \in \mathcal{X}_t \cup \mathcal{S}_{t-1}$ as follows:

$$\text{lrn}_t(s) = \sum_c \omega_t(c) \cdot \log(1 + n_{s,c}) \cdot G_t(s,c), \tag{9}$$

where $\omega_t(c)$ we introduced is an under-learning metric for class c, and $\log(1+n_{s,c})$ is a log-scaled object-count weight. $G_t(s,c)$ measures the learning gain of scene s on class c to reflect the learnable metric.

To obtain $\omega_t(c)$ for each class $c \in \mathcal{C}_t \cup \mathcal{Y}|_{\mathcal{S}_{t-1}}$, we base on $\{LD_t^i\}_{i=1}^I$ in Eq. (8) to compute the class-level recall of the i-th sub-stage across scenes $R_t^i(c) = 1/S \cdot \sum_s R_t^i(s,c)$, $S = |\mathcal{X}_t \cup \mathcal{S}_{t-1}|$, and record the peak class-level recall $\hat{R}_t^i(c) = \max\{R_t^1(c), R_t^2(c), \ldots, R_t^I(c)\}$ and class-level recall $R_t^I(c)$ for the final sub-stage. A class is under-learnt when: (i) it remains under-detected ($R_t^I(c)$ is low), (ii) its quality has dropped from peak ($\hat{R}_t^i(c) > R_t^I(c)$), and (iii) it is learnable ($\hat{R}_t^i(c)$ is not close to zero). These derive our under-learning metric for class c:

$$\omega_t(c) = \Big[\big(1 - R_t^I(c)\big) + \max\Big(0,\ \hat{R}_t^i(c) - R_t^I(c)\Big)\Big] \cdot \hat{R}_t^i(c). \tag{10}$$

To obtain $G_t(s,c)$, we summarize the positive recall gains of class c in scene s between consecutive sub-stages by:

$$G_t(s,c) = \sum_{i=2}^{I} \max\left(0,\ R_t^i(s,c) - R_t^{i-1}(s,c)\right). \tag{11}$$

This metric reflects that objects of class c in scene s can achieve how much performance gain during the increment stage t.

In this way, our learnability score in Eq. (9) not only reflects class-level under-learning information via $\omega_t(c)$, but also captures scene-level possible learning gain via $G_t(s,c)$.

Diversity score. Beyond learnability, we further consider scene diversity to ensure the memory bank covers a broad range of learning patterns. To this end, we embed each scene by its learning dynamics pattern and favor candidates that differ from scenes already in the memory bank.

Specifically, for each scene $s \in \mathcal{X}_t \cup \mathcal{S}_{t-1}$, we construct a two-dimensional embedding $\mathbf{e}_t(s) \in \mathbb{R}^2$ from its learning dynamics. The first dimension aggregates the class-weighted learning gain $\sum_c \omega_t(c) \cdot G_t(s,c)$, and the second aggregates the class-weighted quality drop $\sum_c \omega_t(c) \cdot D_t(s,c)$, where $D_t(s,c) = \hat{R}_t^i(s,c) - R_t^I(s,c)$ is the drop from peak to final recall. As a result, we have

$$\mathbf{e}_t(s) = \left[\sum_c \omega_t(c) \cdot G_t(s,c); \sum_c \omega_t(c) \cdot D_t(s,c)\right] \in \mathbb{R}^2. \tag{12}$$

Then, the embedding vector is L2-normalized to the unit circle and thus vectors of different scenes become comparable. Scenes with similar $\mathbf{e}_t(s)$ exhibit similar learning dynamics patterns and they usually offer limited information gain.

Let $\mathcal{S}_t$ denote the set of scenes already selected for the t-th increment stage. The diversity score of a candidate s against the memory bank is defined as:

$$\mathrm{div}(s \mid \mathcal{S}_t) = 1 - \frac{1}{|\mathcal{S}_t|} \sum_{j \in \mathcal{S}_t} \max\big(0,\ \mathbf{e}_t(s)^\top \mathbf{e}_t(j)\big), \tag{13}$$

where higher values indicate greater diversity from the existing memory bank.

Memory bank evolution. We combine the scene-aware learnability score from Eq. (9) and diversity score from Eq. (13) into a joint selection objective. Our memory bank evolution consists of the following two update rules:

$$\begin{cases} s^{+} = \arg\max\limits_{s \in \mathcal{X}_t} \left[(1-\lambda)\, \mathrm{lrn}_t(s) + \lambda\, \mathrm{div}(s \mid \mathcal{S}_t) \right], \\ s^{-} = \arg\min\limits_{s \in \mathcal{S}_{t-1}} \left[(1-\lambda)\, \mathrm{lrn}_t(s) + \lambda\, \mathrm{div}(s \mid \mathcal{S}_{t-1} \setminus \{s\}) \right], \end{cases} \tag{14}$$

where $\lambda \in [0, 1]$ balances the learnability and diversity. In Eq. 14, the first rule identifies the most valuable candidate s^{+} from the current stage data $\mathcal{X}_t$ to be added to the memory bank. And the second rule identifies the least valuable scene s^{-} in the existing memory bank $\mathcal{S}_{t-1}$ to be evicted. The evicted scene s^{-} is replaced by the candidate s^{+}. This update process repeats until all $s \in \mathcal{X}_t$ have been checked whether it should be added to $\mathcal{S}_t$, and finally we achieve the memory bank evolution from $\mathcal{M}_{t-1} = \{\mathcal{S}_{t-1}, \mathcal{Y}|_{\mathcal{S}_{t-1}}\}$ to $\mathcal{M}_t = \{\mathcal{S}_t, \mathcal{Y}|_{\mathcal{S}_t}\}$.

5 Experiments

5.1 Experimental Setups

Datasets. We adopt two indoor 3D object detection benchmarks to evaluate the proposed L-I3DOD task. **SUN RGB-D** [28] contains 10,335 RGB-D images captured by multiple sensors, with 5,285 training and 5,050 test samples. **ScanNetV2** [5] contains 1,513 reconstructed RGB-D scans with 1,201 scans for training and 312 for evaluation.

Incremental protocol. We evaluate under three incremental protocols with larger label spaces and more stages than prior work [4, 40, 41]. For SUN RGB-D, we use 40 classes for 3-stage (20+10+10), 5-stage (5 stages of 8 classes each), and 10-stage (10 stages of 4 classes each). For ScanNetV2, we use 35 classes for 3-stage (15+10+10), 5-stage (5 stages of 7 classes each), and 10-stage (10 stages of 3–4 classes each). The classes are arranged in descending frequency order so that later stages introduce progressively rarer categories, making the task increasingly challenging.

Implementation details. We evaluate with two 3D detector backbones, *i.e.*, TR3D [22] and VoteNet [19]. We train the base stage for 90 epochs and each incremental stage for 15 epochs for TR3D, and follow SDCoT [40] for VoteNet. We use the default optimizer and data augmentation settings provided by each backbone. The memory bank maintains a fixed global budget of scenes, set to 10% of the full training set. Each incremental stage is divided into I=5 sub-stages for learning-dynamics monitoring. The learning dynamics recall $R_t^i(s, c)$ is computed at an IoU threshold of 0.5. The review emphasis strength is set to $\eta = 3.0$ (Eq. 5). The learnability-diversity trade-off parameter is $\lambda = 0.5$ (Eq. 14).

Evaluation protocol. We report mean average precision at IoU threshold 0.25 (mAP@0.25) and Average Forgetting (AF) after every training stage. After stage t, mAP is computed over all seen classes $\mathcal{C}_{\leq t}$. To quantify forgetting at stage t, we let $\mathrm{AP}_c^{(t)}$ denote the AP@0.25 of class c using the stage-t model, and let intro(c)

Table 1: Incremental 3D object detection performance (mAP@0.25) over 3-/5-/10-incremental stages on SUN RGB-D and ScanNetV2 benchmarks.

Method	Backbone	SUN RGB-D (40 classes)			ScanNetV2 (35 classes)		
		3-stage	5-stage	10-stage	3-stage	5-stage	10-stage
Fine-tuning	VoteNet [19]	1.12	0.62	0.03	5.09	4.21	0.15
CPDet3D [43]		9.70	7.11	0.39	16.33	9.05	2.54
AIC3DOD [4]		10.73	7.32	0.85	19.21	15.51	1.60
SDCoT [40]		10.94	7.66	0.98	18.39	14.38	2.93
SDCoT++ [41]		11.39	7.86	1.67	18.51	16.19	3.38
SDCoT++ & Random Memory		11.52	7.93	3.53	21.19	19.46	7.25
LDMR (ours)		**11.87**	**8.75**	**6.20**	**24.82**	**23.37**	**9.76**
Joint Training (Upper Bound)		22.39	22.39	22.39	36.00	36.00	36.00
Fine-tuning	TR3D [22]	7.86	2.37	0.59	11.49	6.24	0.50
CPDet3D [43]		8.71	7.55	3.32	18.76	7.26	0.66
AIC3DOD [4]		22.43	15.37	9.11	32.69	20.11	8.66
SDCoT [40]		22.10	13.77	8.72	34.65	20.86	6.93
SDCoT++ [41]		22.92	14.28	8.45	35.15	20.61	7.06
SDCoT++ & Random Memory		25.18	20.02	15.54	36.86	23.67	14.86
LDMR (ours)		**29.12**	**25.10**	**19.38**	**39.00**	**28.38**	**17.82**
Joint Training (Upper Bound)		33.26	33.26	33.26	51.52	51.52	51.52

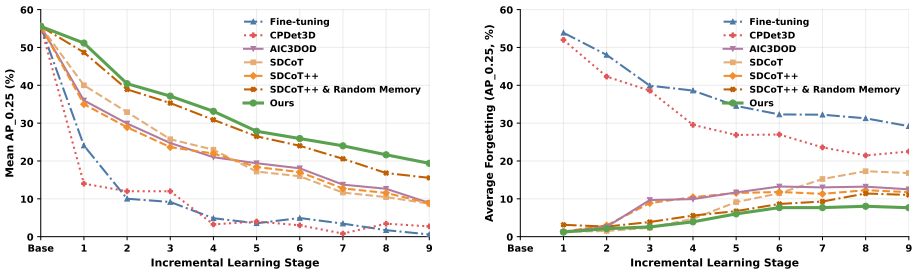


Fig. 4: Stage-wise experimental analysis on SUN RGB-D (with TR3D). *Left:* Overall mAP on all seen classes after each stage. *Right:* Average forgetting computed by Eq. (15) of old classes at each stage. Our method sustains higher seen-class performance and substantially reduces the forgetting issue across stages compared to all baselines.

be the stage at which class c first appears. AF measures the drop from each class's debut performance to its current performance (lower is better):

$$\mathrm{AF}(\mathcal{C}_{<t}) \;=\; \frac{1}{|\mathcal{C}_{<t}|} \sum_{c \in \mathcal{C}_{<t}} \Big(\mathrm{AP}_c^{(\mathrm{intro}(c))} - \mathrm{AP}_c^{(t)} \Big). \tag{15}$$

5.2 Baselines

We compare with the following baselines under the same protocol. **Fine-tuning** is a fundamental baseline that trains only on current-stage ground-truth labels. **CPDet3D** [43] is one of the state-of-the-art 3D detection methods which mines

unlabeled objects via class prototypes and pseudo-label refinement; we adapt it to our incremental setting by treating old-class objects in new-stage scenes as the unlabeled targets. **SDCoT** [40], **SDCoT++** [41], and **AIC3DOD** [4] are the representative incremental 3D object detection methods that rely on teacher-student training with pseudo labels and distillation to retain old-class knowledge. No I3DOD methods are memory-bank based so far. Therefore we include **SDCoT++ & Random Memory** which augments SDCoT++ with a randomly selected memory bank of the same budget as ours.

5.3 Main Results

Overall comparison. Tab. 1 reports the final-stage's mAP@0.25 for all protocols. Among pseudo-labeling based baselines, SDCoT++ is the strongest, but its performance degrades as the number of stages increases. For example, on SUN RGB-D (TR3D), it achieves 22.92 at 3 stages but only 8.45 at 10 stages. Augmenting SDCoT++ with a memory bank leads to a clear improvement, reaching 15.54 at 10 stages. Overall, our method LDMR consistently outperforms this strongest baseline across all settings. On SUN RGB-D (TR3D), we achieve 29.12, 25.10 and 19.38 at 3, 5 and 10 stages, improving over SDCoT++ & Random Memory by 3.94, 5.08, and 3.84 mAP respectively. A similar pattern holds on ScanNetV2 (TR3D), where we reach 39.00, 28.38 and 17.82 compared to 36.86, 23.67, 14.86 by SDCoT++ & Random Memory. These results suggest that our learning-dynamics-driven memory and review provide clear benefits beyond simply having a memory bank in pseudo-labeling methods. For reference, joint training on all classes reaches 33.26/51.52 mAP@0.25 (SUN RGB-D/ScanNetV2, TR3D), serving as a non-incremental upper bound.

Stage-wise trajectory and forgetting. Fig. 4 visualises the per-stage model performance on SUN RGB-D. The left panel shows that our method maintains the highest overall mAP after every stage, while baseline methods show increasingly larger drops as stages progress. The right panel shows that average forgetting of old classes remains consistently lower for our method across all stages, suggesting that our intra-stage review and cross-stage memory evolution effectively preserve previously learned knowledge.

5.4 Ablation Studies

We further verify effectiveness of our method by detailed ablation studies.

Intra-stage review *vs.* cross-stage memory evolution. Tab. 2 demonstrates the individual contribution of key components in our LDMR framework. On SUN RGB-D (TR3D), removing human-like intra-stage review (w/o HR) reduces 10-stage mAP from 19.38 to 17.51 and removing scene-aware memory evolution (w/o ME) reduces it to 18.70. Removing both further drops the 10-stage mAP to 15.54, indicating that the two components are complementary, and their impact holds under different settings. At 3 stages the full model improves over the no-component baseline by 3.94 mAP, and at 5 stages the gain remains 5.08 mAP.

Table 2: Ablation study on our key modules of human-like intra-stage review (HR) and scene-aware cross-stage memory evolution (ME) (SUN RGB-D, TR3D, 3-/5-/10-stages' mAP@0.25).

Method	3-stage	5-stage	10-stage
LDMR (full)	**29.12**	**25.10**	**19.38**
w/o HR	26.67	23.47	17.51
w/o ME	27.44	23.17	18.70
w/o HR & ME	25.18	20.02	15.54

Table 3: Comparison of alternative memory-bank schemes in SDCoT++ & Random Memory (SUN RGB-D, TR3D, 10-stage's mAP@0.25).

Memory-bank scheme	mAP@0.25
Random	15.54
Reservoir	16.12
Max #objects	15.31
Lowest recall	14.10
LDMR	**19.38**

Memory-bank selection schemes. To verify that the improvement does not merely come from using a memory bank, we equip the SDCoT++ & Random Memory baseline with alternative memory-bank selection schemes (Tab. 3). Reservoir sampling, max-object-count, and lowest-recall heuristics get 14.10–16.12 mAP, all below LDMR's 19.38, confirming that the gain stems from our design rather than from the bank itself.

Scoring-function design choices. Tab. 4 ablates scoring functions. For the object-count weight in Eqs. (4) and (9), the log form $\log(1+n_{s,c})$ (19.38) outperforms the linear $n_{s,c}$ (18.55) and sublinear $\sqrt{n_{s,c}}$ (19.01) forms, as strong regulation best prevents object-dense scenes from dominating the score. Moreover, each term in the learnability score (Eq. 9) is necessary: removing the under-learning weight $\omega_t(c)$, the count weight $\log(1+n_{s,c})$, or the learning gain $G_t(s,c)$ reduces performance to 19.13, 17.86, and 18.88, respectively. Besides, simplifying the under-learning weight (Eq. 10) to a low-recall only form drops performance.

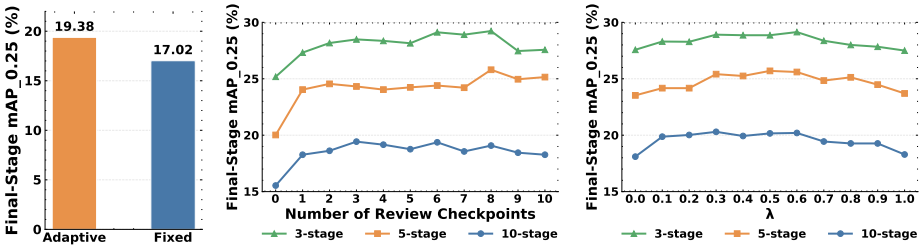
Adaptive *vs.* fixed reviewing. Fig. 5a compares adaptive-weight reviewing against fixed-weight reviewing on the 10-stage protocol. Our approach outperforms the fixed-weight variant, showing that the learning-dynamics-driven adaptation effectively allocates review effort as forgetting patterns vary across stages.

Reviewing granularity. Fig. 5b examines the effect of the reviewing granularity, i.e., the number of sub-stages I into which each incremental stage is divided. Each stage is split by $I-1$ review checkpoints into I sub-stages (e.g., a single checkpoint for two sub-stages). Across all incremental protocols, performance clearly improves as the number of review checkpoints increases from 0 (I increases from 1), indicating that finer-grained monitoring of learning dynamics enables more targeted review. Meanwhile, performance plateaus for too many sub-stages, suggesting that the reviewing segments should not be too short.

Trade-off between learnability and diversity. Fig. 5c studies the effect of the trade-off parameter λ in the memory evolution objective (Eq. 14). Setting λ=0 (learnability only, no diversity) or λ=1 (diversity only, no learnability) generally results in lower performance, while the middle value of λ improves results across all protocols, with the best performance reached around λ=0.3 $\sim$ 0.6. Beyond this range, performance gradually declines, indicating that insufficient learnability or diversity hurts performance. The result suggests that both terms are necessary and that a moderately balanced setting works best.

Table 4: Ablation on the scoring-function design choices (SUN RGB-D, TR3D, 10-stage mAP@0.25). Each variant alters one component of LDMR.

Component	Design choice	mAP@0.25
Count weight (Eqs. 4,9)	linear: $n_{s,c}$	18.55
	sublinear: $\sqrt{n_{s,c}}$	19.01
	log: $\log(1+n_{s,c})$ (final design)	**19.38**
Learnability (Eq. 9)	w/o $\omega_t(c)$ (class under-learning weight)	19.13
	w/o $\log(1+n_{s,c})$ (object-count weight)	17.86
	w/o $G_t(s,c)$ (scene-level learning gain)	18.88
Under-learning weight (Eq. 10)	low-recall only	18.45



(a) Reviewing weight. **(b)** Number of review checkpoints. **(c)** Trade-off λ.

Fig. 5: (a) Ablation study on adaptive and fixed sampling weights in the human-like intra-stage review on SUN RGB-D (with TR3D). (b) Ablation study on the reviewing granularity. (c) Effect of the trade-off parameter λ between learnability and diversity.

6 Conclusion

In this work, we studied long-incremental 3D object detection and showed that existing pipelines can enter a self-reinforcing cycle, where stage-wise supervision shift degrades old-class performance and progressively corrupts pseudo labels. To break this cycle, we proposed Learning-Dynamics-driven Memory and Review (LDMR), which converts learning dynamics into training decisions through human-like intra-stage review and scene-aware cross-stage memory evolution. Across SUN RGB-D and ScanNetV2 under multiple incremental protocols, LDMR consistently improves the performance by mitigating model forgetting, with more pronounced gains in longer incremental stages, highlighting the importance of our learning-dynamics-driven framework for long-horizon 3D continual learning.

7 Acknowledgements

This research is supported by the Agency for Science, Technology and Research (A*STAR) under its MTC Programmatic Funds (Grant No. M23L7b0021), and the Ministry of Education, Singapore, under its MOE Academic Research Fund Tier 2 (MOE-T2EP20124-0013).

References

1. Cao, Y., Wu, F., Chen, D.Z., Zhong, Y., Hong, L., Xu, D.: Vggt-det: Mining vggt internal priors for sensor-geometry-free multi-view indoor 3d object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4708–4717 (2026)
2. Cheng, Q., Wan, Y., Wu, L., Hou, C., Zhang, L.: Continuous subspace optimization for continual learning. In: Belgrave, D., Zhang, C., Lin, H., Pascanu, R., Koniusz, P., Ghassemi, M., Chen, N. (eds.) Advances in Neural Information Processing Systems. vol. 38, pp. 15376–15398 (2025)
3. Cheng, Z., Chen, C., Zhao, Z., Qian, P., Li, X., Yang, X.: Coco-teach: A contrastive co-teaching network for incremental 3d object detection. In: 2023 IEEE International Conference on Image Processing (ICIP). pp. 1990–1994 (2023)
4. Cheng, Z., Wu, F., Qian, P., Zhao, Z., Yang, X.: Aic3dod: Advancing indoor class-incremental 3d object detection with point transformer architecture and room layout constraints. In: 2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). pp. 7512–7521 (2025)
5. Dai, A., Chang, A.X., Savva, M., Halber, M., Funkhouser, T., Nießner, M.: Scannet: Richly-annotated 3d reconstructions of indoor scenes. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 5828–5839 (2017)
6. Feng, T., Wang, M., Yuan, H.: Overcoming catastrophic forgetting in incremental object detection via elastic response distillation. In: Proceedings of the IEEE conference on Computer Vision and Pattern Recognition. pp. 9417–9426 (2022)
7. Han, Y., Zhao, N., Chen, W., Ma, K.T., Zhang, H.: Dual-perspective knowledge enrichment for semi-supervised 3d object detection. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 38, pp. 2049–2057 (2024)
8. Jiao, P., Zhao, N., Chen, J., Jiang, Y.G.: Unlocking textual and visual wisdom: Open-vocabulary 3d object detection enhanced by comprehensive guidance from text and image. In: European Conference on Computer Vision. pp. 376–392 (2024)
9. Kirkpatrick, J., Pascanu, R., Rabinowitz, N., Veness, J., Desjardins, G., Rusu, A.A., Milan, K., Quan, J., Ramalho, T., Grabska-Barwinska, A., et al.: Overcoming catastrophic forgetting in neural networks. Proceedings of the National Academy of Sciences **114**(13), 3521–3526 (2017)
10. Kolodiazhnyi, M., Vorontsova, A., Skripkin, M., Rukhovich, D., Konushin, A.: Unidet3d: Multi-dataset indoor 3d object detection. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 39, pp. 4365–4373 (2025)
11. Lazarow, J., Griffiths, D., Kohavi, G., Crespo, F., Dehghan, A.: Cubify anything: Scaling indoor 3d object detection. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 22225–22233 (2025)
12. Lemeshko, A., Gabdullin, B., Drozdov, N., Konushin, A., Rukhovich, D., Kolodiazhnyi, M.: Zoo3d: Zero-shot 3d object detection at scene level. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 25820–25829 (2026)
13. Li, Z., Hoiem, D.: Learning without forgetting. IEEE Transactions on Pattern Analysis and Machine Intelligence **40**(12), 2935–2947 (2017)
14. Liu, Y., Schiele, B., Vedaldi, A., Rupprecht, C.: Continual detection transformer for incremental object detection. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 23799–23808 (2023)
15. Liu, Y., Cong, Y., Goswami, D., Liu, X., Van De Weijer, J.: Augmented box replay: Overcoming foreground shift for incremental object detection. In: Proceedings of the IEEE international conference on computer vision. pp. 11367–11377 (2023)

16. Liu, Z., Zhang, Z., Cao, Y., Hu, H., Tong, X.: Group-free 3d object detection via transformers. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 2929–2938 (2021)
17. Misra, I., Girdhar, R., Joulin, A.: An end-to-end transformer model for 3d object detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 2906–2917 (2021)
18. Peng, C., Zhao, K., Lovell, B.C.: Faster ilod: Incremental learning for object detectors based on faster rcnn. Pattern Recognition Letters **140**, 109–115 (2020)
19. Qi, C.R., Litany, O., He, K., Guibas, L.J.: Deep hough voting for 3d object detection in point clouds. In: proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 9277–9286 (2019)
20. Rebuffi, S.A., Kolesnikov, A., Sperl, G., Lampert, C.H.: icarl: Incremental classifier and representation learning. In: Proceedings of the IEEE conference on Computer Vision and Pattern Recognition. pp. 2001–2010 (2017)
21. Rukhovich, D., Vorontsova, A., Konushin, A.: Fcaf3d: Fully convolutional anchor-free 3d object detection. In: European Conference on Computer Vision. pp. 477–493 (2022)
22. Rukhovich, D., Vorontsova, A., Konushin, A.: Tr3d: Towards real-time indoor 3d object detection. In: 2023 IEEE International Conference on Image Processing (ICIP). pp. 281–285. IEEE (2023)
23. Savadikar, C., Dai, M., Wu, T.: Cheem: Continual learning by reuse, new, adapt and skip-a hierarchical exploration-exploitation approach. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 25066–25076 (2026)
24. Shen, Y., Geng, Z., Yuan, Y., Lin, Y., Liu, Z., Wang, C., Hu, H., Zheng, N., Guo, B.: V-detr: Detr with vertex relative position encoding for 3d object detection. In: International Conference on Learning Representations (2024)
25. Sheng, H., Cai, S., Zhao, N., Deng, B., Liang, Q., Zhao, M.J., Ye, J.: Ct3d++: Improving 3d object detection with keypoint-induced channel-wise transformer. International Journal of Computer Vision **133**(7), 4817–4836 (2025)
26. Shmelkov, K., Schmid, C., Alahari, K.: Incremental learning of object detectors without catastrophic forgetting. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 3420–3429 (2017)
27. Smith, J.S., Karlinsky, L., Gutta, V., Cascante-Bonilla, P., Kim, D., Arbelle, A., Panda, R., Feris, R., Kira, Z.: Coda-prompt: Continual decomposed attention-based prompting for rehearsal-free continual learning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 11909–11919 (2023)
28. Song, S., Lichtenberg, S.P., Xiao, J.: Sun rgb-d: A rgb-d scene understanding benchmark suite. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 567–576 (2015)
29. Wang, H., Ding, L., Dong, S., Shi, S., Li, A., Li, J., Li, Z., Wang, L.: Cagroup3d: Class-aware grouping for 3d object detection on point clouds. Advances in Neural Information Processing Systems pp. 29975–29988 (2022)
30. Wang, J., Zhao, N.: Uncertainty meets diversity: A comprehensive active learning framework for indoor 3d object detection. In: Proceedings of the IEEE conference on Computer Vision and Pattern Recognition. pp. 20329–20339 (2025)
31. Wang, Z., Zhang, Z., Lee, C.Y., Zhang, H., Sun, R., Ren, X., Su, G., Perot, V., Dy, J., Pfister, T.: Learning to prompt for continual learning. In: Proceedings of the IEEE conference on Computer Vision and Pattern Recognition. pp. 139–149 (2022)

32. Wu, Y., Piao, H., Huang, L.K., Wang, R., Li, W., Pfister, H., Meng, D., Ma, K., Wei, Y.: SD-loRA: Scalable decoupled low-rank adaptation for class incremental learning. In: The Thirteenth International Conference on Learning Representations (2025)
33. Wu, Y., Wang, K., Pan, Y., Zhao, N.: Ccf: Complementary collaborative fusion for domain generalized multi-modal 3d object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 18745–18754 (2026)
34. Xu, J., Zhao, N.: Stream3d: Streaming zero-shot 3d instance segmentation with multi-view noise mask filtering and manifold refining. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Findings. pp. 327–337 (June 2026)
35. Yuan, S., Xu, J., Hu, P., Zhu, X., Zhao, N.: Graph smoothing for enhanced local geometry learning in point cloud analysis. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 40, pp. 12250—12258 (2026)
36. Yun, P., Cen, J., Liu, M.: Conflicts between likelihood and knowledge distillation in task incremental learning for 3d object detection. In: 2021 International Conference on 3D Vision (3DV). pp. 575–585 (2021)
37. Yun, P., Liu, Y., Liu, M.: In defense of knowledge distillation for task incremental learning and its application in 3d object detection. IEEE Robotics and Automation Letters **6**(2), 2012–2019 (2021)
38. Zhang, X., Ye, H., Li, J., Tang, Q., Li, Y., Guo, Y., Guo, J.: Prompt3d: Random prompt assisted weakly-supervised 3d object detection. In: Proceedings of the IEEE conference on Computer Vision and Pattern Recognition. pp. 28046–28055 (2024)
39. Zhao, N., Chua, T.S., Lee, G.H.: Sess: Self-ensembling semi-supervised 3d object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11079–11087 (2020)
40. Zhao, N., Lee, G.H.: Static-dynamic co-teaching for class-incremental 3d object detection. In: Proceedings of the AAAI Conference on Artificial Intelligence. pp. 3436–3445 (2022)
41. Zhao, N., Qian, P., Wu, F., Xu, X., Yang, X., Lee, G.H.: Sdcot++: Improved static-dynamic co-teaching for class-incremental 3d object detection. IEEE Transactions on Image Processing **34**, 4188–4202 (2025)
42. Zhou, D.W., Wang, Q.W., Qi, Z.H., Ye, H.J., Zhan, D.C., Liu, Z.: Class-incremental learning: A survey. IEEE Transactions on Pattern Analysis and Machine Intelligence **46**(12), 9851–9873 (2024)
43. Zhu, Y., Hui, L., Yang, H., Qian, J., Xie, J., Yang, J.: Learning class prototypes for unified sparse-supervised 3d object detection. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 9911–9920 (2025)
44. Zhu, Y., Qian, J., Yang, J., Xie, J., Zhao, N.: Few-shot incremental 3d object detection in dynamic indoor environments. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 18786–18795 (2026)