Appendix for 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

1 Theoretical Analysis

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 (classification + 3D bbox regression) as $\mathcal{L}_t(\theta) = \mathbb{E}_{z\sim P_t}\big[\ell(f_\theta, z)\big]$, and the expected gradient is $g_t(\theta) = \nabla_\theta \mathcal{L}_t(\theta) = \mathbb{E}_{z\sim P_t}[\nabla_\theta \ell(f_\theta, z)]$.

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 $g_u(\theta_{u-1}) = \mathbb{E}_{z\sim P_u}\big[\nabla_\theta \ell(f_\theta, z)\big]\Big|_{\theta=\theta_{u-1}}$. 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*

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

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}.

Distributional shifts between stages (measured by W_1) together with input gradient sensitivity K and large new-stage gradients $\|g_u\|$ produce gradient misalignment that increases old-stage loss, while curvature and step size (L, η_u) further amplify this effect. Hence, this explains why I3DOD forgets in long-incremental stages.

Proof. We give a concise, self-contained proof and state the precise assumptions used. First, we assume the following for all relevant θ and $z \in \mathcal{Z}$:

A1 (Finite first moment) Each distribution P_t has finite first moment under a chosen metric $d_{\mathcal{Z}}$ on $\mathcal{Z}$, so $W_1(P, Q)$ is finite.

* Corresponding authors.

A2 (Directional Lipschitz [2]) There exists $K > 0$ such that for every unit vector $v \in \mathbb{R}^d$ the scalar map $z \mapsto v^\top \nabla_\theta \ell(f_\theta, z)$ is K-Lipschitz w.r.t. $d_{\mathcal{Z}}$. Based on the above, any fixed θ and distributions P, Q consequently satisfy

$$\|g_P(\theta) - g_Q(\theta)\| \le K\, W_1(P, Q).$$

This follows by applying Kantorovich-Rubinstein duality [5] to each directional projection and taking the supremum over unit directions.

A3 (Smoothness) For each old-stage index i, $\mathcal{L}_i$ is L-smooth on the parameter segments considered: $\|H_i(\theta)\|_{\mathrm{op}} \le L$ for θ on $[\theta_{u-1}, \theta_u]$.

Notation. For brevity write $g_u \equiv g_u(\theta_{u-1})$ and $g_i \equiv g_i(\theta_{u-1})$. Define the single-step forgetting increment

$$\Delta_i^{(u)} := \mathcal{L}_i(\theta_u) - \mathcal{L}_i(\theta_{u-1}),$$

with $\theta_u = \theta_{u-1} - \eta_u g_u$. All gradient comparisons below are taken at the same parameter point θ_{u-1}.

Directional KR lemma. By Assumption A2, for any unit v the scalar $\phi_v(z) = v^\top \nabla_\theta \ell(f_\theta, z)$ is K-Lipschitz. Kantorovich-Rubinstein duality applied to each ϕ_v and taking supremum over unit v yields

$$\|g_P(\theta) - g_Q(\theta)\| = \sup_{\|v\| \le 1} \big(\mathbb{E}_P[\phi_v] - \mathbb{E}_Q[\phi_v]\big) \le K\, W_1(P, Q).$$

Single-stage expansion and bounds. Fix $u \in \{s+1, \dots, t\}$. By Taylor's theorem there exists ξ on $[\theta_{u-1}, \theta_u]$ such that

$$\mathcal{L}_i(\theta_u) = \mathcal{L}_i(\theta_{u-1}) + \nabla \mathcal{L}_i(\theta_{u-1})^\top (\theta_u - \theta_{u-1}) + \tfrac{1}{2}(\theta_u - \theta_{u-1})^\top H_i(\xi)(\theta_u - \theta_{u-1}).$$

With $\theta_u - \theta_{u-1} = -\eta_u g_u$ we obtain the exact identity

$$\Delta_i^{(u)} = -\eta_u\, g_u^\top g_i + \tfrac{\eta_u^2}{2}\, g_u^\top H_i(\xi) g_u.$$

Decompose the first term:

$$-g_u^\top g_i = -\|g_u\|^2 + g_u^\top (g_u - g_i) \le -\|g_u\|^2 + \|g_u\|\, \|g_u - g_i\|.$$

Applying the directional KR lemma at θ_{u-1} gives $\|g_u - g_i\| \le K\, W_1(P_i, P_u)$, hence

$$-g_u^\top g_i \le -\|g_u\|^2 + K\, W_1(P_i, P_u)\, \|g_u\|.$$

By Assumption A3, L-smoothness on the segment results in $g_u^\top H_i(\xi) g_u \le L\|g_u\|^2$. Combining yields

$$\Delta_i^{(u)} \le -\eta_u \|g_u\|^2 + \eta_u K\, W_1(P_i, P_u) \|g_u\| + \tfrac{L\eta_u^2}{2} \|g_u\|^2.$$

Multi-stage summation. Summing for $u = s+1, \ldots, t$,

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

Dropping the negative alignment terms $-\eta_u \|g_u\|^2$ (for clarity, our theorem presents a conservative nonnegative bound obtained by dropping $-\eta_u \|g_u\|^2$ which should be used when one wants to capture beneficial alignment between new and old gradients.) yields the stated inequality:

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

This completes the proof, where necessary symbol explanations are as follows:

- P_t: data distribution at stage t on $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$.
- $\mathcal{L}_t(\theta) = \mathbb{E}_{z \sim P_t}[\ell(f_\theta, z)]$: population loss at stage t.
- $g_t(\theta) = \nabla_\theta \mathcal{L}_t(\theta)$: population gradient of stage t at θ.
- $g_u(\theta_{u-1})$: population gradient of stage u evaluated at θ_{u-1} (used in the update).
- θ_u, η_u: parameter after the u-th update and the step size at stage u.
- $W_1(P_i, P_u)$: 1-Wasserstein distance between P_i and P_u under $d_{\mathcal{Z}}$ (assumed finite).
- K: directional Lipschitz constant (input gradient sensitivity) giving $\|g_P(\theta) - g_Q(\theta)\| \leq K\, W_1(P, Q)$ for fixed θ.
- L: parameter-space smoothness constant (Hessian spectral bound) used to control the second-order term.

The bound is a population level (expectation) statement under the stated technical assumptions (finite first moment, directional Lipschitz of input-gradients, and local/global L-smoothness), which is expected to compactly explain why long-incremental stage I3DOD forgets.

2 Dataset and Incremental Split Details

For class lists and stage assignments, Tables 1 and 2 list the object categories used for SUN RGB-D (40 classes) and ScanNetV2 (35 classes), respectively. In both cases, classes are ranked by training-set instance count and assigned to stages in that order under the 3-, 5-, and 10-stage protocols.

3 Additional Implementation Details

As described in the main paper (Sec. 5.2), all methods share the same detector backbone, optimizer, data augmentation, and per-stage training schedule to ensure fair comparison. For TR3D [4], we train the base stage for 90 epochs

Table 1: SUN RGB-D class list and stage assignments for all 3 protocols.

#	Class	Count	3-stage	5-stage	10-stage
1	chair	9278	1	1	1
2	table	2539	1	1	1
3	pillow	1564	1	1	1
4	sofa_chair	1056	1	1	1
5	desk	933	1	1	2
6	bed	771	1	1	2
7	sofa	706	1	1	2
8	computer	682	1	1	2
9	lamp	519	1	2	3
10	box	487	1	2	3
11	garbage_bin	470	1	2	3
12	cabinet	400	1	2	3
13	shelf	342	1	2	4
14	drawer	297	1	2	4
15	night_stand	293	1	2	4
16	endtable	260	1	2	4
17	sink	222	1	3	5
18	picture	217	1	3	5
19	stool	214	1	3	5
20	coffee_table	212	1	3	5
21	bookshelf	204	2	3	6
22	painting	195	2	3	6
23	keyboard	185	2	3	6
24	dresser	182	2	3	6
25	tv	181	2	4	7
26	whiteboard	180	2	4	7
27	cpu	180	2	4	7
28	toilet	171	2	4	7
29	paper	155	2	4	8
30	ottoman	151	2	4	8
31	bench	138	3	4	8
32	recycle_bin	131	3	4	8
33	monitor	129	3	5	9
34	printer	120	3	5	9
35	plant	112	3	5	9
36	door	111	3	5	9
37	book	108	3	5	10
38	mirror	98	3	5	10
39	laptop	93	3	5	10
40	towel	91	3	5	10

Table 2: ScanNetV2 class list and stage assignments for all 3 protocols.

#	Class	Count	3-stage	5-stage	10-stage
1	chair	4357	1	1	1
2	door	2026	1	1	1
3	otherfurniture	1985	1	1	1
4	books	1554	1	1	1
5	cabinet	1427	1	1	2
6	table	1271	1	1	2
7	window	928	1	1	2
8	pillow	745	1	2	2
9	picture	661	1	2	3
10	box	657	1	2	3
11	desk	551	1	2	3
12	shelves	486	1	2	3
13	towel	481	1	2	4
14	sofa	406	1	2	4
15	sink	390	1	3	4
16	clothes	386	2	3	4
17	lamp	364	2	3	5
18	bed	307	2	3	5
19	bookshelf	300	2	3	5
20	curtain	292	2	3	5
21	mirror	279	2	3	6
22	bag	253	2	4	6
23	whiteboard	251	2	4	6
24	counter	216	2	4	7
25	toilet	201	2	4	7
26	nightstand	190	3	4	7
27	refrigerator	186	3	4	8
28	television	177	3	4	8
29	dresser	170	3	5	8
30	shower_curtain	116	3	5	9
31	bathtub	113	3	5	9
32	paper	52	3	5	9
33	person	39	3	5	10
34	floor_mat	32	3	5	10
35	blinds	22	3	5	10

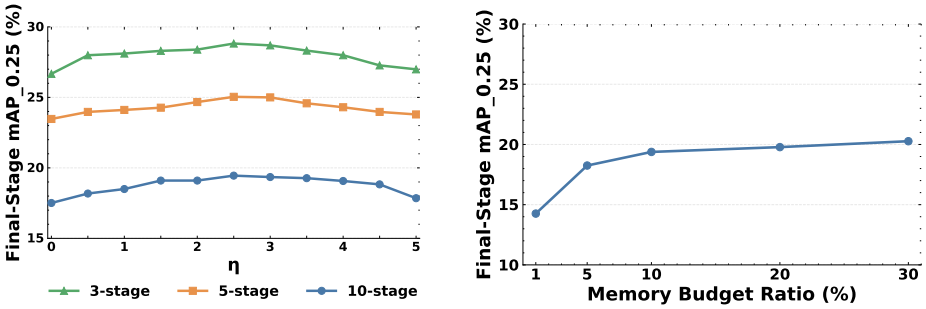
and each incremental stage for 15 epochs; for VoteNet [3], we follow the training schedule of SDCoT [6] and SDCoT++ [7]. For SDCoT [6] and SDCoT++ [7], we adopt the original configurations released with the respective papers. Both methods are model-agnostic, so substituting their default VoteNet [3] backbone with TR3D [4] requires minimal modification beyond swapping the detection model. For CPDet3D [8], which was originally designed for sparse-supervised detection, we treat each incremental stage's ground-truth annotations as the sparse labels and expand the prototype bank and classification head at every new stage; the prototype mining and cooperative refinement then provides supervision signals for old classes that lack annotations in the current stage. For AIC3DOD [1], we replace the original backbone with ours while retaining the pseudo-labeling, teacher-student knowledge distillation, and supervised training components. On ScanNetV2, we additionally apply the layout constraint loss. Because SUN RGB-D does not provide room layout annotations, AIC3DOD [1] is run on it without the layout constraint loss.

4 Computational Cost

We report the computational cost of LDMR on the SUN RGB-D 10-stage protocol with the TR3D backbone, measured on a single NVIDIA RTX A5000 GPU with an AMD EPYC 9354 CPU. During training LDMR uses approximately 6.3 GB of GPU memory, and during inference approximately 3.5 GB (batch-size dependent). Against an average baseline training time of 82 min/stage, LDMR adds only a modest 2–10 min/stage for the periodic checkpoint evaluation, learning-dynamics monitoring, and memory bank evolution, scaling with the number of scenes in the stage. At test time it introduces no inference overhead, running the unmodified TR3D detector (14.7M parameters, 78.5 GFLOPs), so its inference cost is identical to the backbone and to all compared methods.

5 Additional Ablations

Review strength. Fig. 1a examines the effect of the review strength η across all three incremental protocols. When η is too small, the reviewing emphasis is too weak to effectively retain previously learned knowledge, resulting in higher forgetting of old classes. Conversely, setting η too large places excessive emphasis on replayed samples, which interferes with the learning of new classes and degrades overall performance. A moderate value of η achieves a better trade-off across all protocols, and we adopt η=3 in all main experiments.
Memory budget ratio. Fig. 1b studies the effect of varying the memory budget as a proportion of the total training data. As expected, increasing the memory budget improves performance, since more stored exemplars provide richer supervision for knowledge consolidation. However, the gains diminish noticeably beyond a ratio of 10%, where further increases yield only marginal improvements at a higher storage cost. We therefore set the memory budget ratio to 10% in



(a) Review strength η. **(b)** Memory budget ratio, 10-stage.

Fig. 1: Additional ablation studies on SUN RGB-D (with TR3D).

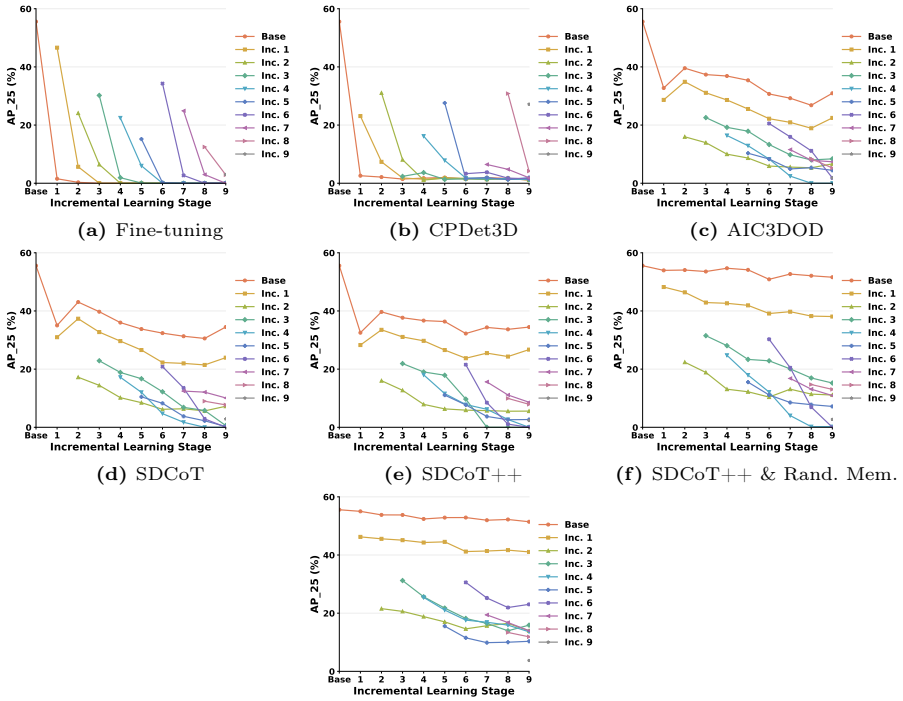
Table 3: Additional metrics (mAP_{50}, AR_{25}, AR_{50}) on SUN RGB-D and ScanNetV2 across the 3-/5-/10-stage protocols with the TR3D backbone.

		SUN RGB-D (40 classes)			**ScanNetV2** (35 classes)		
Metric	**Method**	**3-stage**	**5-stage**	**10-stage**	**3-stage**	**5-stage**	**10-stage**
mAP_{50}	SDCoT++ & Random Memory	16.65	10.74	8.55	15.79	8.88	4.56
	LDMR	**17.53**	**14.75**	**10.57**	**17.39**	**11.42**	**4.73**
AR_{25}	SDCoT++ & Random Memory	63.80	70.80	45.39	60.94	50.24	34.81
	LDMR	**76.45**	**75.32**	**68.37**	**65.18**	**50.51**	**39.25**
AR_{50}	SDCoT++ & Random Memory	35.13	29.80	21.95	27.85	18.25	10.77
	LDMR	**41.31**	**39.76**	**33.12**	**32.20**	**21.17**	**11.98**

all main experiments, as it offers a balance between performance and memory efficiency.

6 Additional Evaluation Metrics

The main paper reports mAP@0.25 as the primary metric. Here we additionally report mAP@0.50 (mAP_{50}) and average recall at both IoU thresholds (AR_{25}, AR_{50}). Table 3 compares the strongest baseline (SDCoT++ & Random Memory) against LDMR across the 3-/5-/10-stage protocols on SUN RGB-D and ScanNetV2 with the TR3D backbone. The mAP_{25} to mAP_{50} drop is large for both methods (*e.g.*, LDMR falls from 19.38 to 10.57 on SUN RGB-D and from 17.82 to 4.73 on ScanNetV2 at 10 stages), indicating that long-incremental forgetting degrades both classification and localization. LDMR improves both. At 10-stage the mAP_{50} gains of +2.02/+0.17 (SUN/Scan) reflect better localization, while the AR_{25} gains of +22.98/+4.44 and AR_{50} gains of +11.17/+1.21 reflect substantially better recall of old classes.



(a) Fine-tuning **(b)** CPDet3D **(c)** AIC3DOD

(d) SDCoT **(e)** SDCoT++ **(f)** SDCoT++ & Rand. Mem.

(g) Ours

Fig. 2: Per-stage incremental 3D object detection performance (mAP@0.25) on SUN RGB-D under the 10-stage protocol. Each curve corresponds to the set of classes introduced at an incremental stage and tracks their performance after each subsequent stage.

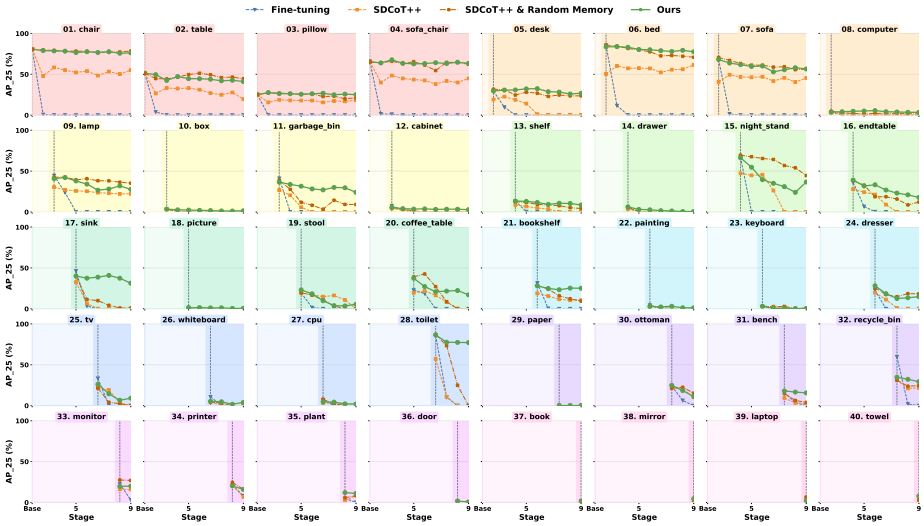


Fig. 3: Per-class incremental 3D object detection performance (AP@0.25) across 10 incremental stages on SUN RGB-D. Each subplot corresponds to one class.

7 Per-Class and Per-Stage Experimental Results

Fig. 2 shows the model performance across stages for classes grouped by the stage they are involved in training on SUN RGB-D under the 10-stage protocol. Each curve tracks the performance of one class group.

Fine-tuning and CPDet3D [8] exhibit such severe model forgetting that the performance of earlier stage groups collapses almost immediately once new classes are introduced. The pseudo-labeling based approaches (AIC3DOD [1], SDCoT [6], and SDCoT++ [7]) mitigate this issue to some extent and retain earlier groups for a few stages, but their trajectories still deteriorate significantly in later stages as the number of incremental updates grows.

SDCoT++ & Random Memory further stabilizes the trajectories, improving retention across multiple stage groups. Our method produces the most stable trajectories, particularly for class groups introduced after the second incremental stage.

Fig. 3 shows per-class test AP@0.25 trajectories across the 10 incremental stages on SUN RGB-D for four representative methods: Fine-tuning as the weakest baseline, SDCoT++ as the representative pseudo-labeling approach, SDCoT++ & Random Memory as the strongest baseline, and our LDMR. Across the majority of classes, our method exhibits more stable per-class AP curves that decline more slowly than all baselines, indicating reduced forgetting over the incremental stages. This stability holds for both frequent base-stage classes (e.g., *chair*, *table*) and classes introduced at intermediate stages (e.g., *endtable*, *bookshelf*), where baselines show steeper degradation after several stages. The effect is clearest for *sink* (class 17) and *toilet* (class 28), both of which are bathroom-

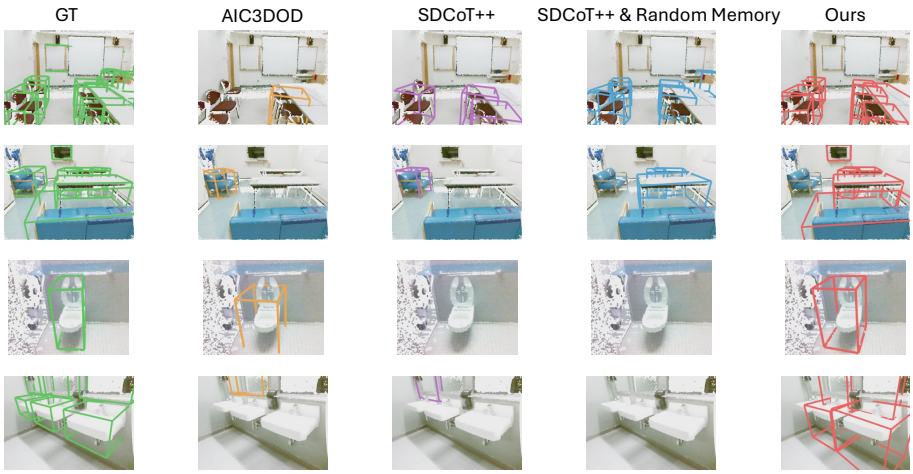


Fig. 4: Qualitative comparison of incremental 3D object detection performance over 10 incremental stages on SUN RGB-D (with TR3D).

specific objects that appear in a small subset of scenes and are therefore underrepresented in the training distribution. For these two classes, SDCoT++ and even SDCoT++ & Random Memory sharply fall to near-zero AP within a few stages of their introduction, whereas our method sustains meaningful detection performance across subsequent stages. This indicates that LDMR better preserves knowledge for under-represented classes, which are most vulnerable to the distribution shift across stages.

8 Qualitative Results

Fig. 4 presents qualitative detection results from scenes in SUN RGB-D. Each column corresponds to a different method, with the ground-truth shown on the left for reference. In the first two rows, which depict cluttered indoor scenes containing multiple objects, the baseline methods miss many instances and produce noticeably incomplete predictions. Our method recovers most ground-truth objects with tighter bounding boxes and fewer missed detections. The third and fourth rows illustrate scenes containing less frequent object categories. The baselines largely fail to detect these rare classes, producing no predictions at all or generating poorly localized boxes. Our method detects these objects with reasonable localization accuracy. This observation is consistent with the per-class results in Fig. 3, where our approach maintains the performance across stages on under-represented classes such as toilet and sink.

Fig. 5 shows qualitative results on ScanNetV2. Across all four scenes, the baseline methods fail to detect most objects, particularly in densely furnished rooms (rows 1 and 2). Our method recovers most of the ground-truth objects. In the sparser scene (row 3) and when rare objects appear (row 4), the baselines

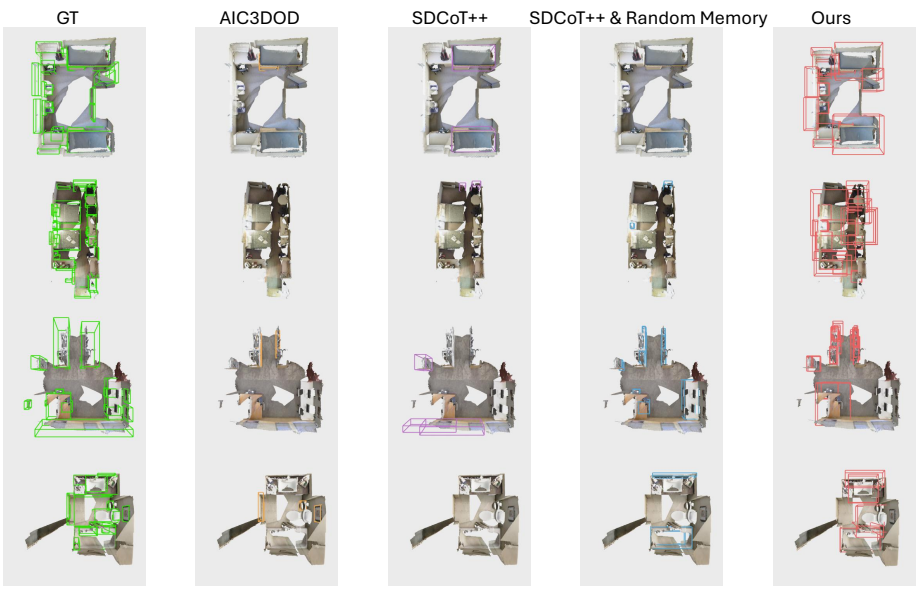


Fig. 5: Qualitative comparison of incremental 3D object detection performance over 10 incremental stages on ScanNetV2 (with TR3D).

struggle with less common categories, whereas our approach maintains reliable detection, consistent with the quantitative trends reported in our main paper.

9 Failure Cases and Limitations

(a) ScanNetV2. **(b)** SUN RGB-D.

Fig. 6: Failure cases on ScanNetV2 and SUN RGB-D.

We illustrate representative failure cases in Fig. 6. On ScanNetV2 (Fig. 6a), the model misses several ground-truth objects and occasionally produces false positives. These errors occur predominantly in scenes with sparse or incomplete point cloud reconstructions, where the lack of sufficient geometric cues

presents a fundamental challenge for point cloud-based detectors. On SUN RGB-D (Fig. 6b), the primary failure mode is missed detections in heavily cluttered scenes where severe inter-object occlusion reduces the visible surface area of individual instances, a known difficulty for 3D detection methods. In the rightmost example, the model fails to detect the three labeled garbage bins yet predicts a bounding box around a nearby unlabeled one, which reflects annotation inconsistency in the dataset. These cases suggest that the remaining performance gap stems from input data quality and annotation completeness, which are hard to resolve through algorithmic improvements alone.

References

1. 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)
2. Mishkin, A., Khaled, A., Wang, Y., Defazio, A., Gower, R.: Directional smoothness and gradient methods: Convergence and adaptivity. Advances in Neural Information Processing Systems **37**, 14810–14848 (2024)
3. 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)
4. 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)
5. Villani, C., et al.: Optimal transport: old and new, vol. 338. Springer (2009)
6. 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)
7. 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)
8. 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)