

1 A Overview

2 We begin by outlining the structure of this supplementary material. In §B, we detail the design and
 3 functionality of the Long-Term Memory module. In §C, we furnish additional visualizations that
 4 compare our approach against ESAM [1] and illustrate its behavior under online inference.

5 B More Details about Long-Term Memory

6 Analogous to standard multi-target tracking [2, 3, 4, 5, 6, 7, 8], our Long-Term Memory module
 7 employs separate workflows for training and inference. Below, we present the corresponding
 8 pseudocode for each phase.

Algorithm 1: Long-Term Memory Inference

Input: Instance embeddings \mathbf{Q}_t , boxes \mathbf{B}_t at current frame t ; Tracklet embeddings \mathbf{Q}^{Trk} , boxes \mathbf{B}^{Trk} , ages α^{Trk} ; LTM Buffer
Output: Updated tracklets \mathbf{Q}^{Trk} , boxes \mathbf{B}^{Trk} and age α^{Trk} .

▷ Affinity Computation

- 1 Applying score threshold γ to \mathbf{Q}_t and \mathbf{B}_t yields valid instance subsets $\hat{\mathbf{Q}}_t$ and $\hat{\mathbf{B}}_t$.
- 2 Compute \mathbf{E}_{ij} between $\{\hat{\mathbf{Q}}_t, \hat{\mathbf{B}}_t\}$ and all $\{\mathbf{Q}^{\text{Trk}}, \mathbf{B}^{\text{Trk}}\}$ using Eq. (1).
- 3 Compute affinity scores \mathbf{A}_{ij} from \mathbf{E}_{ij} using Eq. (2).

▷ Hungarian Assignment

- 4 Apply Hungarian algorithm to find optimal one-to-one matched pairs \mathcal{M} .

▷ Tracklet Update

- 5 Update Tracklet $\{\mathbf{Q}^{\text{Trk}}, \mathbf{B}^{\text{Trk}}, \alpha^{\text{Trk}}\}$ based on matched pairs \mathcal{M} using Eq. (3)

▷ Tracklet Initialization and Expiration

- 6 Mark unmatched tracks older than T_{life} as stale and push into LTM Buffer.

▷ LTM Reactivation

- 7 Compute $\mathbf{E}_{ij}^{\text{LTM}}$ between unmatched instances and LTM Buffer
- 8 Compute $\mathbf{A}_{ij}^{\text{LTM}}$ and apply Hungarian algorithm to get reactivation pairs \mathcal{M}_{LTM} .
- 9 Restore $\{\mathbf{Q}^{\text{LTM}}, \mathbf{B}^{\text{LTM}}, \alpha^{\text{LTM}}\}$ into tracklet embeddings based on \mathcal{M}_{LTM} using Eq. (3).

▷ New Tracklet Generation

- 10 For still-unmatched instances, initialize a new tracklet by assigning $\{\mathbf{Q}^{\text{new}}, \mathbf{B}^{\text{new}}, \alpha^{\text{new}} = 1\}$.

Algorithm 2: Long-Term Memory Training

Input: Instance embeddings \mathbf{Q}_t , boxes \mathbf{B}_t at current frame t ; Tracklet embeddings \mathbf{Q}^{Trk} , boxes \mathbf{B}^{Trk} , ages α^{Trk} ;
Output: Updated tracklets \mathbf{Q}^{Trk} , boxes \mathbf{B}^{Trk} and age α^{Trk} .

▷ Affinity Computation

- 1 Match each ground-truth annotation to its corresponding candidate in \mathbf{Q}_t and \mathbf{B}_t , yielding the valid instances $\hat{\mathbf{Q}}_t$ and $\hat{\mathbf{B}}_t$.
- 2 Compute \mathbf{E}_{ij} between $\{\hat{\mathbf{Q}}_t, \hat{\mathbf{B}}_t\}$ and corresponding $\{\mathbf{Q}^{\text{Trk}}, \mathbf{B}^{\text{Trk}}\}$ using Eq. (1).
- 3 Compute affinity scores \mathbf{A}_{ij} from \mathbf{E}_{ij} using Eq. (2).

▷ GT-based Assignment

- 4 We use ground-truth annotations to match the valid instances $\hat{\mathbf{Q}}_t$ and $\hat{\mathbf{B}}_t$ with the existing tracklets \mathbf{Q}^{Trk} and boxes \mathbf{B}^{Trk} , encoding the results in a binary matrix \mathbf{A}_{tgt} whose entries are 1 for matched pairs and 0 otherwise.

▷ Tracklet Update

- 5 Update Tracklet $\{\mathbf{Q}^{\text{Trk}}, \mathbf{B}^{\text{Trk}}, \alpha^{\text{Trk}}\}$ based on matrix \mathbf{A}_{tgt} using Eq. (3)
- 6 For unmatched instances, initialize a new tracklet by assigning $\{\mathbf{Q}^{\text{new}}, \mathbf{B}^{\text{new}}, \alpha^{\text{new}} = 1\}$.

▷ Loss Function

- 7 Compute the Cross-Entropy Loss between \mathbf{A} and the ground-truth matrix \mathbf{A}_{tgt} .

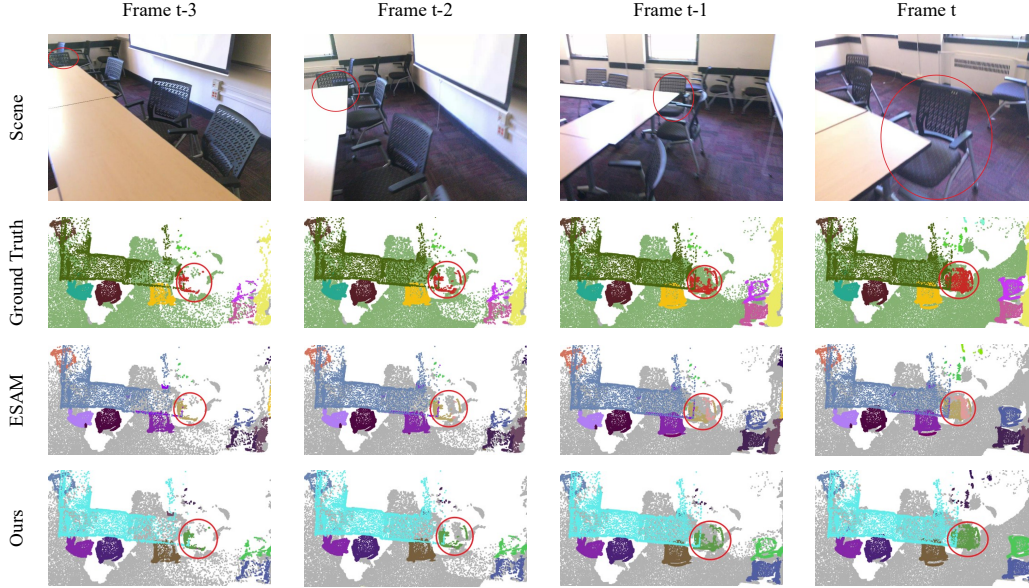


Figure 1: Visualization of segmentation results on ScanNet200 dataset in successive frames.

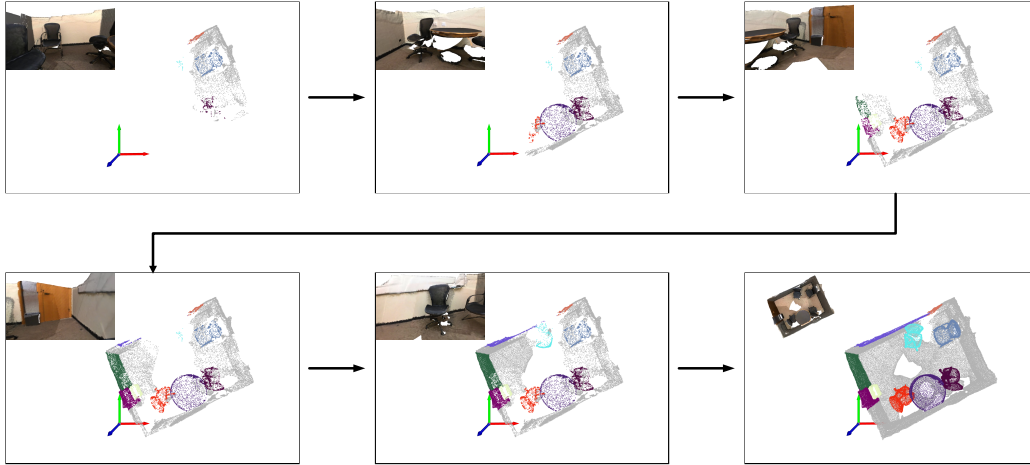


Figure 2: Online visualization of our method on ScanNet200 dataset.

9 C More Visualizations

10 In Fig. 1, we illustrate instance segmentation across successive frames on ScanNet200. Thanks to
 11 our tracking-centric online 3D segmentation framework, the method produces precise, complete
 12 object masks with robust association even when instances are occluded in preceding or following
 13 frames. Compared to ESAM, our approach exhibits substantially improved segmentation fidelity and
 14 continuity in challenging scenarios.

15 In Fig. 2, we present the online 3D segmentation data stream. The results demonstrate that our
 16 method achieves accurate, real-time instance segmentation and 3D reconstruction, thereby providing
 17 a robust foundation for downstream robotic applications.

18 References

- 19 [1] Xiuwei Xu, Huangxing Chen, Linqing Zhao, Ziwei Wang, Jie Zhou, and Jiwen Lu. Embodiedsam:
 20 Online segment any 3d thing in real time. *arXiv preprint arXiv:2408.11811*, 2024.

- 21 [2] Jiarui Cai, Mingze Xu, Wei Li, Yuanjun Xiong, Wei Xia, Zhuowen Tu, and Stefano Soatto.
22 Memot: Multi-object tracking with memory. In *Proceedings of the IEEE/CVF conference on*
23 *computer vision and pattern recognition*, pages 8090–8100, 2022.
- 24 [3] Fangao Zeng, Bin Dong, Yuang Zhang, Tiancai Wang, Xiangyu Zhang, and Yichen Wei. Motr:
25 End-to-end multiple-object tracking with transformer. In *European conference on computer*
26 *vision*, pages 659–675. Springer, 2022.
- 27 [4] Tim Meinhardt, Alexander Kirillov, Laura Leal-Taixe, and Christoph Feichtenhofer. Trackformer:
28 Multi-object tracking with transformers. In *Proceedings of the IEEE/CVF conference on computer*
29 *vision and pattern recognition*, pages 8844–8854, 2022.
- 30 [5] Yuang Zhang, Tiancai Wang, and Xiangyu Zhang. Motrv2: Bootstrapping end-to-end multi-
31 object tracking by pretrained object detectors. In *Proceedings of the IEEE/CVF conference on*
32 *computer vision and pattern recognition*, pages 22056–22065, 2023.
- 33 [6] Yanwei Li, Zhiding Yu, Jonah Philion, Anima Anandkumar, Sanja Fidler, Jiaya Jia, and Jose
34 Alvarez. End-to-end 3d tracking with decoupled queries. In *Proceedings of the IEEE/CVF*
35 *International Conference on Computer Vision*, pages 18302–18311, 2023.
- 36 [7] Shuxiao Ding, Lukas Schneider, Marius Cordts, and Juergen Gall. Ada-track: End-to-end
37 multi-camera 3d multi-object tracking with alternating detection and association. In *Proceedings*
38 *of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15184–15194,
39 2024.
- 40 [8] Ziqi Pang, Jie Li, Pavel Tokmakov, Dian Chen, Sergey Zagoruyko, and Yu-Xiong Wang. Standing
41 between past and future: Spatio-temporal modeling for multi-camera 3d multi-object tracking.
42 In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages
43 17928–17938, 2023.