Supplementary Material for
Memory-based Adapters for Online 3D Scene Perception

Xiuwei Xu[1]*, Chong Xia[2]*, Ziwei Wang[3], Linqing Zhao[1], Yueqi Duan[2], Jie Zhou[1], Jiwen Lu[1]†
[1]Department of Automation, Tsinghua University
[2]Department of Electronic Engineering, Tsinghua University
[3]Carnegie Mellon University
{xxw21, xiac20}@mails.tsinghua.edu.cn; {ziweiwa2}@andrew.cmu.edu;
{linqingzhao}@tju.edu.cn; {duanyueqi, jzhou, lujiwen}@tsinghua.edu.cn

This supplementary material is organized as follows:

- Section A demonstrates the detailed architecture of our baseline models in three tasks and how to insert our adapters into them.
- Section B details the training hyperparameters adopted in our experiments.
- Section C details per-class experimental results.

A. Detailed Architecture

We illustrate the architectures of both image and point cloud backbones and show how to insert the memory-based adapters into them in Figure 1. For online 3D semantic segmentation, we use U-Net [7] as the image backbone and Minkowski-UNet [1] as the point cloud backbone, which is shown in Figure 1 (D) and (C) respectively. For online 3D object detection, we adopt ResNet [3] with FPN [5] as the image backbone and FCAF3D [8] as the point cloud backbone, which is shown in Figure 1 (E) and (B) respectively. For online 3D instance segmentation, we use the same image backbone as the object detection task and adopt TD3D [4] as the point cloud backbone, which is shown in Figure 1 (E) and (A) respectively. Note that for TD3D, the backbone maintains a high-resolution scene representation for ROI-wise instance prediction. We consrtuct a point cloud memory to cache this scene representation, which ensures the point clouds within each ROI are the most complete up to current time. This design helps us acquire complete instance mask by simply performing 3D NMS, which avoids complicated mask fusion strategy [6] to merge instance masks of different frames.

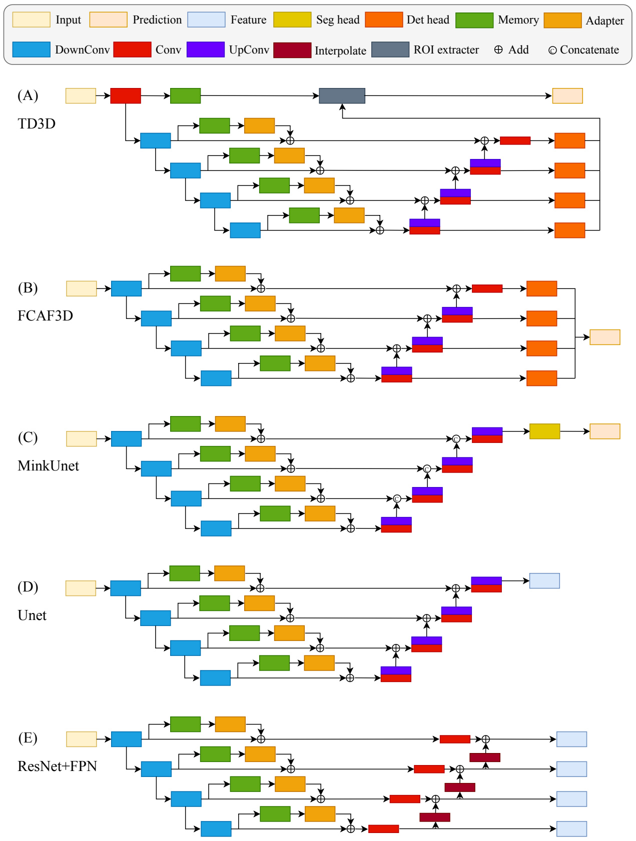

Figure 1. Details about the architectures of image and point cloud backbones and how to insert the adapters into them.

B. Training Hyperparameters

We train the online perception models in two stage. Firstly we train single-view perception model $\mathcal{M}_{SV}$ on ScanNet-25k [2]. Secondly we insert the memory-based adapters into $\mathcal{M}_{SV}$ and finetune the network on ScanNet RGB-D videos.

*Equal contribution.
†Corresponding author.

For online semantic segmentation, we set max epoch as 250, weight decay as 0.01, initial learning rate as 0.0008 and adopt AdamW optimizer with OneCycleLR scheduler for the first stage. Then we set max epoch as 36, weight decay as 0.01, initial learning rate as 0.008 and adopt AdamW optimizer with a stepwise scheduler which steps at 24 and 32 epoch for the second stage.

For online object detection, we set max epoch as 12, weight decay as 0.0001, initial learning rate as 0.001 and adopt AdamW optimizer with a stepwise scheduler which steps at 8 and 11 epoch for the first stage. Then we adopt the same hyperparameters for finetuning in the second stage.

For online instance segmentation, we set max epoch as 33, weight decay as 0.0001, initial learning rate as 0.001 and adopt AdamW optimizer with a stepwise scheduler which steps at 28 and 32 epoch for the first stage. Then we adopt the same hyperparameters for finetuning.

C. Class-specific Results

We provide class-specific experimental results of out method on three 3D scene perception tasks. Table 1 and 2 show the 3D semantic segmentation results on ScanNet and SceneNN dataset with per-class IoU. Table 3 and 4 show the 3D object detection results on ScanNet dataset with per-class AP_{25} and AP_{50}. Table 5 and 6 show the 3D object detection results on ScanNet dataset with per-class AP_{25} and AP_{50}.

References

[1] Christopher Choy, JunYoung Gwak, and Silvio Savarese. 4d spatio-temporal convnets: Minkowski convolutional neural networks. In *CVPR*, pages 3075–3084, 2019. 1

[2] Angela Dai, Angel X. Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *CVPR*, pages 5828—-5839, 2017. 1

[3] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016. 1

[4] Maksim Kolodiazhnyi, Danila Rukhovich, Anna Vorontsova, and Anton Konushin. Top-down beats bottom-up in 3d instance segmentation. *arXiv preprint arXiv:2302.02871*, 2023. 1

[5] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *CVPR*, pages 2117–2125, 2017. 1

[6] Leyao Liu, Tian Zheng, Yun-Jou Lin, Kai Ni, and Lu Fang. Ins-conv: Incremental sparse convolution for online 3d segmentation. In *CVPR*, pages 18975–18984, 2022. 1

[7] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *MICCAI*, pages 234–241. Springer, 2015. 1

[8] Danila Rukhovich, Anna Vorontsova, and Anton Konushin. Fcaf3d: fully convolutional anchor-free 3d object detection. In *ECCV*, pages 477–493. Springer, 2022. 1

Table 1. Per-class 3D semantic segmentation results (IoU) of our method on the ScanNet validation set.

	wall	floor	cabinet	bed	chair	sofa	table	door	window	bookshelf	picture	counter	desk	curtain	fridge	curtain	toilet	sink	bathtub	others	mean
Ours	85.7	97.1	63.1	80.7	89.0	76.1	73.7	66.1	63.6	77.1	41.8	65.5	61.2	58.9	61.0	72.7	95.2	77.9	94.2	53.6	72.7

Table 2. Per-class 3D semantic segmentation results (IoU) of our method on the SceneNN validation set.

	wall	floor	cabinet	bed	chair	sofa	table	door	window	bookshelf	picture	counter	desk	curtain	fridge	sink	mean
Ours	75.3	82.6	59.8	82.8	62.0	57.8	18.4	52.4	20.5	55.9	29.4	52.6	44.9	50.2	80.3	81.8	56.7

Table 3. Per-class 3D object detection results (AP_{25}) of our method on the ScanNet validation set.

	cabinet	bed	chair	sofa	table	door	window	bookshelf	picture	counter	desk	curtain	fridge	curtain	toilet	sink	bathtub	others	mean
Ours	55.2	85.4	88.7	87.2	63.3	62.5	47.3	66.2	36.0	65.2	80.1	65.0	58.1	76.3	99.7	76.7	93.3	62.0	70.5

Table 4. Per-class 3D object detection results (AP_{50}) of our method on the ScanNet validation set.

	cabinet	bed	chair	sofa	table	door	window	bookshelf	picture	counter	desk	curtain	fridge	curtain	toilet	sink	bathtub	others	mean
Ours	36.7	75.6	73.9	77.9	57.0	33.8	19.8	43.7	19.4	26.3	62.8	32.4	41.1	24.6	89.2	46.7	84.8	52.2	49.9

Table 5. Per-class 3D instance segmentation results (AP_{25}) of our method on the ScanNet validation set.

	cabinet	bed	chair	sofa	table	door	window	bookshelf	picture	counter	desk	curtain	fridge	curtain	toilet	sink	bathtub	others	mean
Ours	60.3	86.8	91.5	80.3	72.8	56.0	55.3	67.5	45.1	48.9	72.9	68.4	56.5	86.3	99.7	81.3	87.8	65.3	71.3

Table 6. Per-class 3D instance segmentation results (AP_{50}) of our method on the ScanNet validation set.

	cabinet	bed	chair	sofa	table	door	window	bookshelf	picture	counter	desk	curtain	fridge	curtain	toilet	sink	bathtub	others	mean
Ours	50.9	79.1	82.5	71.3	63.6	44.0	36.0	45.5	38.5	30.3	57.3	49.8	52.9	78.9	99.7	66.6	84.9	56.9	60.5