Scene Reconstruction as Mapping Priors for 3D Detection

Supplementary Material

Training Details

To fully utilize the proposed scalable scene reconstruction (Surfel and 3DGS) pipelines, we propose a three-stage training strategy for the proposed MPA3D approach. In the initial pre-training stage, the model is trained on 100 million internal video sequences without mapping priors (*e.g.*, Surfels and 3DGS). To ensure scalability, we leverage an off-board auto-labeler based on [1] to generate 7-DoF bounding boxes and category labels. Subsequently, we conduct a mid-training phase on a curated dataset of approximately 350K sequences. In this stage, we incorporate mapping priors (*i.e.*, Surfels and 3DGS reconstructions), and utilize high-quality manual annotations to further refine the model. Finally, we fine-tune the model on the Waymo Open Dataset (WOD) [2] training set, utilizing the full complement of mapping priors to yield the final detector.

Concatenation Fusion

In Section 4.4 of the main paper, we investigate various fusion strategies, including summation, averaging, concatenation, and gated fusion. The first two variants are straightforward, performing simple element-wise aggregation of voxel-based features across modalities. In contrast, the concatenation strategy adopts a hierarchical approach. First, the LiDAR feature $\bar{f}_{\text{lidar}}$ and the Surfel feature $\bar{f}_{\text{surfel}}$ are concatenated along the channel dimension, yielding $f_{\text{concat}} = [\bar{f}_{\text{lidar}}, \bar{f}_{\text{surfel}}]$. We then employ a PointMLP to compress the channel dimension by half, projecting the features back to the size of the original LiDAR dimension. This process is repeated to integrate the 3D Gaussian features $\bar{f}_{\text{Gaussian}}$, followed by a second PointMLP for dimensionality reduction and a residual skip connection to the initial LiDAR feature.

Influence of Different Modalities

With a smaller model MPA3D-96M, we show the effectiveness of different inputs modalities (i.e., camera, surfel, 3DGS) in Table 1. By integrating more more modalities, the 3D detection performance improves consistently.

Input Modality				L2 APH
LiDAR	Camera	Surfel	3DGS	Overall
✓				74.9
✓	✓			75.7
✓	✓	✓		76.1
✓	✓	✓	✓	77.4

Table 1. Ablation study on input modalities using MPA3D-96M model. Adding surfel and 3DGS inputs leads to improved 3D detection performance in WOD validation set.

Runtime

The latency of our baseline model is 245ms. When adding both mapping priors, it goes to 452ms. Note that 245ms is higher than 20ms reported in SWFormer paper, because we included extra modules for camera, and used a much larger transformer backbone without the optimized fused transformer kernels.

References

[1] Yingwei Li, Charles R Qi, Yin Zhou, Chenxi Liu, and Dragomir Anguelov. Modar: Using motion forecasting for 3d object detection in point cloud sequences. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9329–9339, 2023. 1

[2] Pei Sun, Henrik Kretzschmar, Xerxes Dotiwalla, Aurelien Chouard, Vijaysai Patnaik, Paul Tsui, James Guo, Yin Zhou, Yuning Chai, Benjamin Caine, Vijay Vasudevan, Wei Han, Jiquan Ngiam, Hang Zhao, Aleksei Timofeev, Scott Ettinger, Maxim Krivokon, Amy Gao, Aditya Joshi, Yu Zhang, Jonathon Shlens, Zhifeng Chen, and Dragomir Anguelov. Scalability in perception for autonomous driving: Waymo open dataset. In *CVPR*, 2020. 1