

StreamVLO: Streaming Visual-LiDAR Odometry with Cumulative Drift Compensation

Mengmeng Liu¹ Jiuming Liu² Michael Ying Yang³ Chaokang Jiang⁴ Jiangtao Li⁴
Yunpeng Zhang⁴ Hesheng Wang² Francesco Nex¹ Hao Cheng^{1*}

¹University of Twente ²Shanghai Jiao Tong University ³University of Bath ⁴Phigent Robotics
{m.liu-1, h.cheng-2}@utwente.nl

1. Overview

The supplementary materials are structured as follows:

- We give more detailed illustrations about the network architecture of the MMG module in Section 2;
- More experimental results are provided in Section 3.
- We display more visualization results in Section 4.

2. Architecture of the MMG Module

Fig. 3 in the main manuscript shows the detailed network structure of MMG (MaxPooling, Mamba [3], and gMLP [7]). MMG harmonizes diverse input representations across different modalities and temporal dimensions: The gMLP encodes sequential information into a unified feature space, providing a foundation for learning from varied data sources. Mamba then establishes temporal interactions within the sequences, effectively capturing long-term dependencies across frames. Finally, MaxPooling condenses the sequence into a single, unified representation, preserving essential information in a compact form.

In this paper, we incorporate the Mamba block proposed in [3] into our method for processing image and LiDAR data due to its excellent performance and speed. Mamba [3] is designed for linear-time sequence modeling using structured state space sequence models (SSMs) [4]. These models are extended to selectively propagate or forget information along the temporal dimension based on the current input token. Specifically, let X denote the input features derived from the image and point cloud data processed by the gMLP layer. These features then serve as the input tokens for the Mamba block: $\hat{X} = \text{LN}(X)$, $\bar{X} = \sigma(\text{Conv1D}(\text{Linear}(\hat{X})))$,

$$\hat{X} = \sigma(\text{Linear}(\hat{X})), \quad (1)$$

$$Y = \text{Linear}(\text{SSM}(\bar{X})) \odot \hat{X} + X, \quad (2)$$

*Corresponding author.

Table 1. **Comparison with learning-based multi-modal odometry on KITTI 09-10 sequences.** Our StreamVLO is trained on 00-06 sequences while other models are trained on 00-08 sequences.

Method	Modalities	09		10		Mean (09-10)	
		t_{rel}	r_{rel}	t_{rel}	r_{rel}	t_{rel}	r_{rel}
Self-VLO [6]	visual+LiDAR	2.58	1.13	2.67	<u>1.28</u>	2.62	<u>1.21</u>
VIOLearner [8]	visual+inertial	<u>1.82</u>	<u>1.08</u>	<u>1.74</u>	1.38	<u>1.78</u>	1.23
StreamVLO (Ours)	visual+LiDAR	0.63	0.28	0.65	0.39	0.64	0.34

where σ denotes the SiLU activation function [5], Conv1D denotes a 1D convolution layer, LN denotes the linear normalization, and SSM is the standard selective state space model proposed by [3]. The output Y denotes the temporally encoded features for the consequent MaxPooling Layer.

3. Additional Ablation Studies

We provide more ablation and generalization studies on KITTI [2], Argoverse [1] datasets to analyze different settings of our proposed method.

3.1. Comparison with learning-based multi-modal odometry on KITTI

We assess the performance of StreamVLO against several learning-based multi-modal odometry models on sequences 09-10, as they are trained on sequences 00-08 and only report the evaluation results on sequences 09-10. As shown in Table 1, remarkably, despite being trained on fewer sequences, StreamVLO significantly outperforms both the visual-LiDAR and visual-inertial methods.

3.2. Ablation Studies

Query Numbers in the Top-k Winner-takes-all Loss. As shown in Table 2, by increasing the number of queries from 50 to 200, the performance on most of the sequences starts to decrease after $k = 100$. Hence, we opt $k = 100$ for the top-k winner-takes-all of the keypoint-aware auxiliary loss.

Table 2. The impact of varying the number of queries (top-k) for the winner-takes-all loss.

top-k	07		08		09		10		Mean	
	t_{rel}	r_{rel}	t_{rel}	r_{rel}	t_{rel}	r_{rel}	t_{rel}	r_{rel}	t_{rel}	r_{rel}
50	0.34	0.23	0.83	0.30	0.68	0.30	0.69	0.43	0.64	0.32
100	0.26	0.23	0.79	0.28	0.63	0.28	0.65	0.39	0.59	0.29
200	0.38	0.40	0.77	0.25	0.69	0.33	0.79	0.42	0.66	0.35

Table 3. The impact of varying the sub-clip length T_s , the maximum history frame length T_h , and the compensation interval T_g . The best results are **bold**.

T_s	T_h	T_g	07		08		09		10		Mean	
			t_{rel}	r_{rel}	t_{rel}	r_{rel}	t_{rel}	r_{rel}	t_{rel}	r_{rel}	t_{rel}	r_{rel}
1	30	20	0.33	0.29	0.91	0.38	0.85	0.43	0.89	0.53	0.75	0.41
3	15	20	0.29	0.31	0.76	0.29	0.67	0.31	0.70	0.35	0.61	0.32
3	30	20	0.26	0.23	0.79	0.28	0.63	0.28	0.65	0.39	0.59	0.29
3	45	20	0.30	0.33	0.83	0.29	0.67	0.32	0.70	0.41	0.63	0.34
3	30	10	0.35	0.39	0.87	0.35	0.73	0.37	0.71	0.43	0.67	0.39
3	30	30	0.31	0.32	0.87	0.32	0.61	0.31	0.75	0.45	0.64	0.35

Varying Frame Lengths. We also analyze the impact of frame lengths for the sub-clip T_s , the maximum history frame length T_h , and the compensation interval T_g . As shown in Table 3, aggregating losses (Section 3.3 in the main paper) multiple frames ($T_s = 3$) leads to a better performance than the single-frame ($T_s = 1$) losses. By varying the maximum history frame length T_h from 15 to 45, the performance first increases and then starts to decrease after increasing T_h to 30. Similarly, the compensation interval $T_g = 20$ yields the best performance among the other frame lengths. Overall, when $T_s = 3$, $T_h = 30$, and $T_g = 20$, our method achieves the best performance.

4. Visualization of the Results

4.1. 2D & 3D Trajectory Visualization

We display the 2D and 3D trajectories for all KITTI evaluation sequences (00–10) in Fig. 2 and Fig. 3, respectively. As shown, our estimated trajectories closely align with the ground truth, demonstrating the effectiveness of the proposed odometry method.

4.2. Keypoints-aware Auxiliary Loss

In Fig. 4 and Fig. 5, we provide additional visualizations of the selected top- k keypoints with minimal error relative to the ground-truth pose in our keypoints-aware auxiliary loss. Most selected keypoints lie on static objects, such as buildings, trees, and parked cars, which provide more reliable cues for consistent ego-motion estimation. In contrast, only a few keypoints fall on dynamic objects, since moving cars or pedestrians often introduce inconsistent motions. This validates our top- k winner-takes-all strategy, which explicitly favors geometrically stable regions and suppresses unreliable dynamic correspondences.

This robustness is further enhanced by the memory feature and pose banks, which implicitly filter outliers over time. Since dynamic objects usually exhibit inconsistent

motions across frames, the historical observations stored in the memory banks provide larger temporal context for identifying such inconsistencies. As a result, the model can further reduce the influence of dynamic regions beyond explicit keypoint selection. Visual comparisons are also provided in Fig. 1.

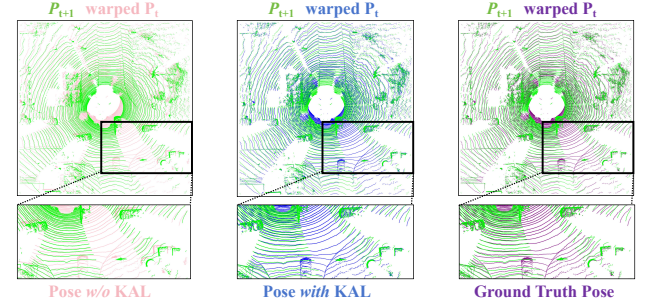


Figure 1. Comparison on Keypoint-aware Auxiliary Loss (KAL). Removing KAL leads to worse pose estimates on dynamic scenes.

References

- [1] Ming-Fang Chang, John Lambert, Patsorn Sangkloy, Jagjeet Singh, Slawomir Bak, Andrew Hartnett, De Wang, Peter Carr, Simon Lucey, Deva Ramanan, et al. Argoverse: 3d tracking and forecasting with rich maps. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8748–8757, 2019. 1
- [2] Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun. Vision meets robotics: The kitti dataset. *The International Journal of Robotics Research*, 32(11):1231–1237, 2013. 1
- [3] Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces. In *First Conference on Language Modeling*, 2024. 1
- [4] Albert Gu, Karan Goel, and Christopher Ré. Efficiently modeling long sequences with structured state spaces. In *International Conference on Learning Representations*, 2022. 1
- [5] Dan Hendrycks and Kevin Gimpel. Gaussian error linear units (gelus). *arXiv preprint arXiv:1606.08415*, 2016. 1
- [6] Bin Li, Mu Hu, Shuling Wang, Lianghao Wang, and Xiaojin Gong. Self-supervised visual-lidar odometry with flip consistency. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 3844–3852, 2021. 1
- [7] Hanxiao Liu, Zihang Dai, David So, and Quoc V Le. Pay attention to mlps. *Advances in Neural Information Processing Systems*, 34:9204–9215, 2021. 1
- [8] E Jared Shamwell, Kyle Lindgren, Sarah Leung, and William D Nothwang. Unsupervised deep visual-inertial odometry with online error correction for rgb-d imagery. *IEEE transactions on pattern analysis and machine intelligence*, 42(10):2478–2493, 2019. 1

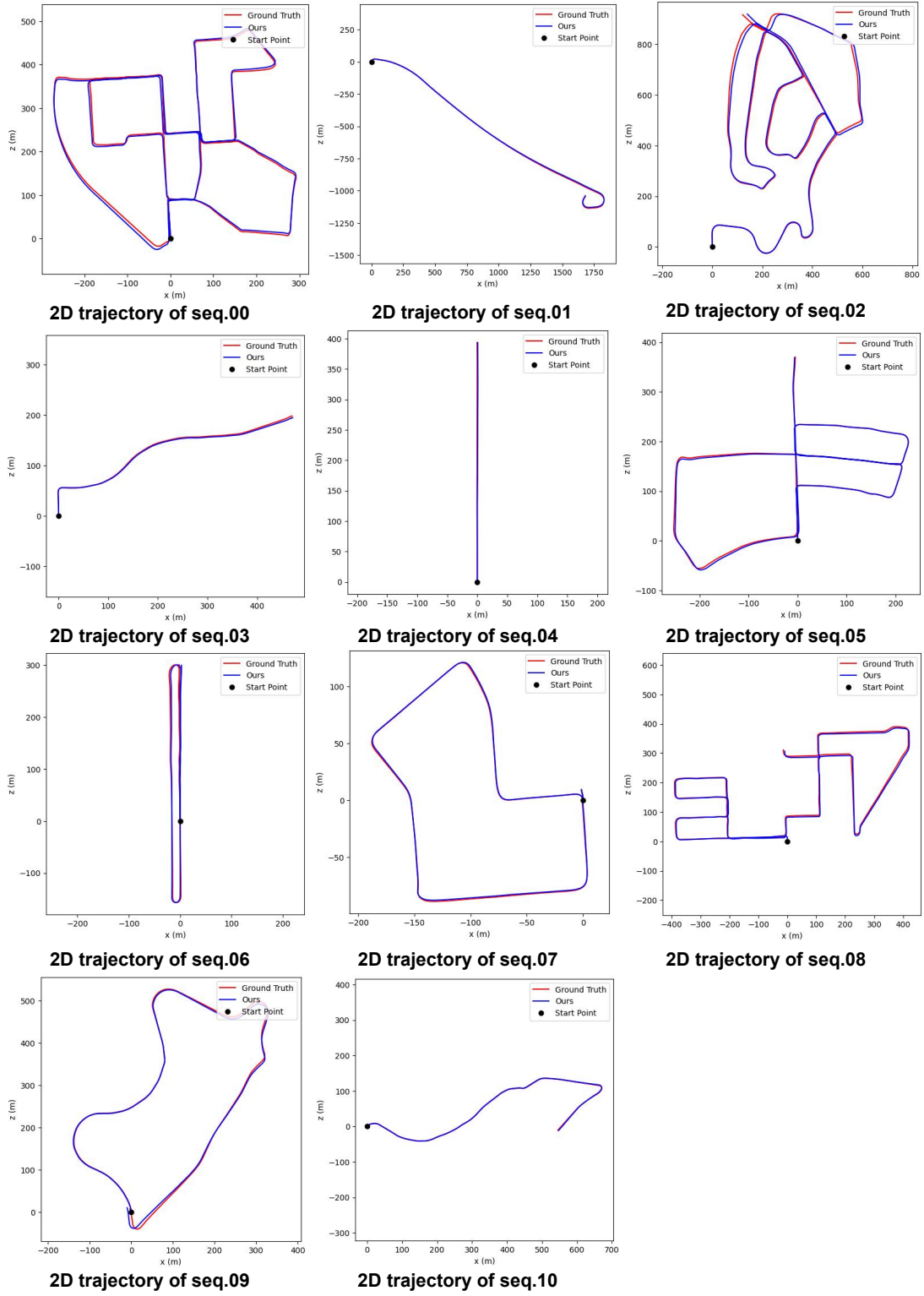
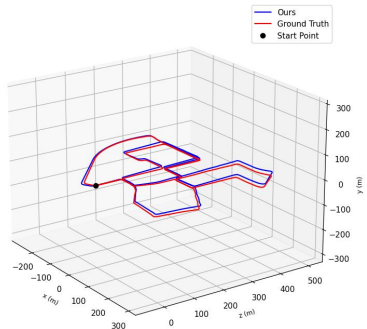
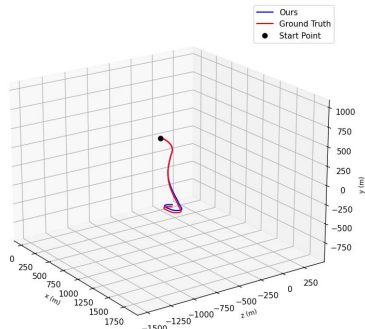


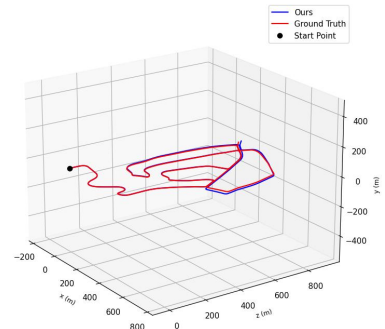
Figure 2. **The 2D trajectories of ground-truth pose and our estimated pose.** Comprehensive 2D trajectory results are shown here on 00-10 sequences of the KITTI dataset.



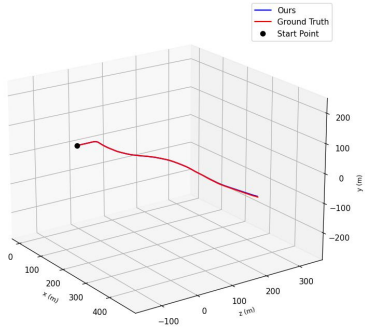
3D trajectory of seq.00



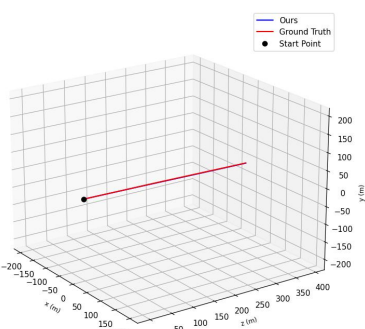
3D trajectory of seq.01



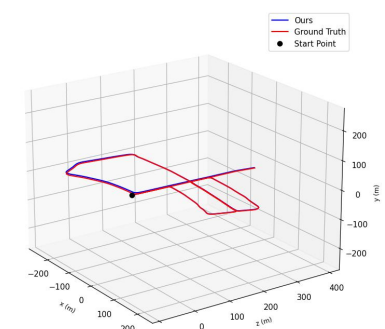
3D trajectory of seq.02



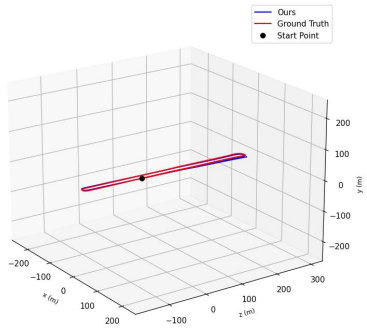
3D trajectory of seq.03



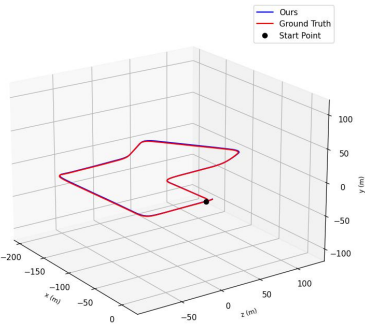
3D trajectory of seq.04



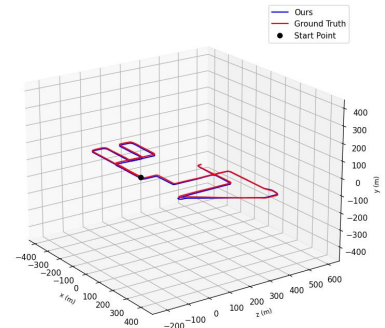
3D trajectory of seq.05



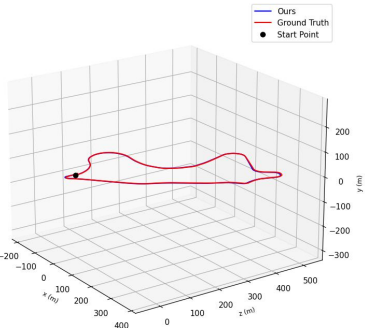
3D trajectory of seq.06



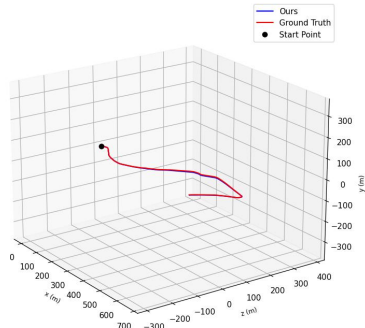
3D trajectory of seq.07



3D trajectory of seq.08



3D trajectory of seq.09



3D trajectory of seq.10

Figure 3. **The 3D trajectories of ground-truth pose and our estimated pose.** Comprehensive 3D trajectory results are shown here on 00-10 sequences of the KITTI dataset.



Figure 4. **Visualization of the Keypoint-Aware Auxiliary Loss (1).** Green points indicate the top-k queries with minimal error relative to the ground truth pose, showing that they are primarily located in regions associated with static objects, while less focus on dynamic objects (red boxes).



Figure 5. **Visualization of the Keypoint-Aware Auxiliary Loss (2).** Green points indicate the top-k queries with minimal error relative to the ground truth pose, showing that they are primarily located in regions associated with static objects, while less focus on dynamic objects (red boxes).