MemReg: Streaming Outdoor LiDAR Point Cloud Registration with Hybrid Memory Buffers

Jiuming Liu[1]*, Mengmeng Liu[2]*, Guangming Wang[3], Lihao Liu[3], Michael Ying Yang[4], Yunpeng Zhang[5], Francesco Nex[2], Hao Cheng[2], Hesheng Wang[1]†

[1] School of Automation and Intelligent Sensing, Shanghai Jiao Tong University

[2] University of Twente [3] University of Cambridge [4] University of Bath [5] Phigent Robotics

liujiuming123@gmail.com m-liu.1@utwente.nl wanghesheng@sjtu.edu.cn

Abstract

Prior outdoor LiDAR registration works are commonly limited by the pair-wise input paradigm, which neglects temporal correlations inherent within streaming sequences. In this paper, we propose a novel multi-frame outdoor point cloud registration network for streaming LiDAR scans enhanced with pose-related memory buffers. The key observation is that long-term temporal LiDAR sequences can provide rich global contextual information to complete sparse measurements, filter outliers, and address low-overlap problems, thereby boosting registration performance. Specifically, two hybrid memory buffers are designed, including an implicit memory feature buffer and an explicit memory pose buffer, to store and dynamically update pose-related temporal features. Moreover, a novel dynamic history weighting module is developed to adaptively fuse current and history pose-related features. Extensive experiments on three outdoor datasets, including KITTI, nuScenes, and Apollo-Southbay, demonstrate state-of-the-art performance of MemReg, surpassing all previous pair-wise methods and also multi-frame SLAM systems. Our method also generalizes surprisingly well to multiview indoor registration scenarios with rather competitive performance on 3DMatch, 3DLoMatch, and ScanNet. Code will be released at `https://github.com/IRMVLab/MemReg`.

1 Introduction

Point cloud registration is a fundamental task in the computer vision field, which aims to find the optimal transformation matrix between point cloud frames [1, 2]. It is widely applied in various downstream tasks, such as autonomous driving [3, 4, 5, 6, 7, 8, 9, 10, 11] and SLAM systems [12, 13, 14, 15, 16, 17].

Even though object-level and indoor point cloud registration have been widely explored [18, 19, 20, 21], the large-scale outdoor LiDAR registration is still underexplored [3, 22]. Challenges are mainly three-fold: (1) Distinct from object-level point cloud, outdoor points are typically more sparse, irregular, noisy, and have a wider spatial distribution range [23]. (2) Outliers, e.g., occlusions and highly dynamic objects, introduce inconsistent motion patterns, undermining the final ego-motion regression [22, 24]. (3) Low-overlap point pairs pose another significant challenge [18]. Due to the high speed of ego vehicles in autonomous driving scenarios, consecutive LiDAR scans often have large translational displacements, making it difficult to search for neighboring points across frames [25]. To tackle these issues, HRegNet [3] addresses mismatches and ensures reliable correspondences caused by outliers with bilateral and neighborhood consensus in point-matching layers. RegFormer [22] uses a projection-aware Transformer with enlarged receptive fields to handle the low-overlap

*Equal Contribution.

†Corresponding Author.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

issue. *However, these prior outdoor registration methods only take pair-wise point clouds as inputs, ignoring intrinsically rich temporal information in streamlined LiDAR sequences.*

Inspired by recent advances in developing temporal cues for tasks like semantic segmentation [26, 27, 28, 29], object detection [30, 31], and multi-object tracking [32, 33], we aim to improve registration accuracy by extending pair-wise frames to streaming multi-frame inputs. The key motivations are: (1) Long-term temporal information helps complete scenes by supplementing sparse and noisy observations across consecutive frames. (2) Multi-frame inputs enhance temporal consistency, improving the recognition and rejection of outliers for effective point correlation. (3) For the low-overlap issue, motion priors from past timestamps can provide strong pose initialization and reduce the distance between initially distant point pairs. These advantages are demonstrated in Sec. 5.

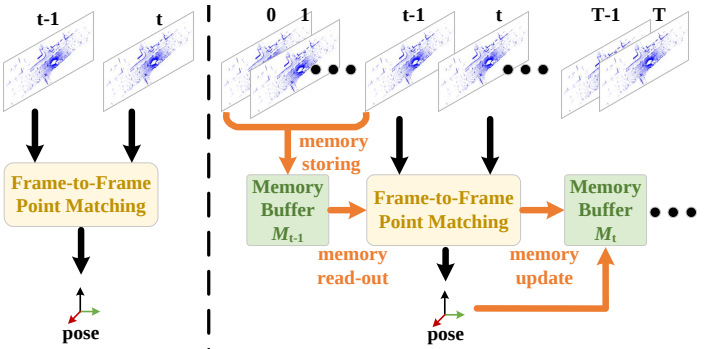


Figure 1: **Comparison with prior pair-wise registration methods.** Prior methods are commonly constrained by pair-wise registration (a), lacking temporal exploration in LiDAR sequences. In contrast, we propose a streaming multi-frame registration network using long-term temporal priors from hybrid memory buffers (b).

In this paper, we propose a novel streaming multi-frame LiDAR registration network, termed MemReg, designing temporal memory buffers and dynamic history weighting to explore inherent temporal consistency as in Fig. 1. Temporal cues from multiple history frames are progressively stored and retrieved in an autoregressive frame-by-frame manner to improve each pair-wise registration, rather than directly outputting multi-frame poses. Hybrid memory buffers are designed to store implicit temporal features (*Memory Feature Buffer*) and explicit historical transformations (*Memory Pose Buffer*) with a fixed memory length. Specifically, when the current pair of frames is delivered into the network, the *Memory Feature Buffer* is first updated by concatenating the current motion features and pre-stored multi-frame history motion features. Then, a Mamba-based [34, 35] temporal encoding layer is developed to establish long-term temporal dependencies across the updated memory features with high efficiency. Similarly, current pose transformations and temporally stored ones are processed by another Mamba module as a strong pose initialization in the *Memory Pose Buffer*, generating transformation features. To effectively merge history and current pose-related features, a dynamic history weighting module inspired by [36, 37] is designed to rescale history and current features and adaptively weight current features with multi-frame historical ones. Finally, the current pose is regressed and refined, which is used to update memory buffers for the next timestamp within sequential LiDAR pairs autoregressively.

Overall, the key contributions of this paper are:

- We propose a novel outdoor LiDAR point cloud registration paradigm enhanced by multi-frame streaming motions, which improves the robustness to challenging cases like outliers and low-overlap inputs.
- Explicit and implicit memory buffers store temporal information from historical frames. Memory Feature Buffer holds implicit pose-related features, while Memory Pose Buffer maintains explicit history poses for enhanced pose initialization.
- To effectively integrate history and current features, a dynamic history weighting module is designed to adaptively merge features on different scales.
- Our MemReg achieves state-of-the-art performance on outdoor registration datasets, including KITTI [38], nuScenes [39], and Apollo-Southbay [40]. Furthermore, our method generalizes well to the indoor multiview and pairwise registration tasks on 3DMatch, 3DLoMatch [41], and ScanNet [42] benchmarks.

2 Related Work

Outdoor Point Cloud Registration. Point cloud registration [43, 41, 44, 45, 46, 47, 19, 48, 49, 50, 51, 52, 53, 54, 55, 56, 57] is a crucial task that targets at finding the optimal transformation matrix between point cloud frames. Recently, large-scale outdoor registration has raised remarkable research attention, but the original sparsity, irregularity, and high-dynamic characteristics of outdoor LiDAR points pose significant challenges [58, 59, 60]. As a pioneering work, HRegNet [3] proposes bilateral consensus and neighborhood consensus to reduce mismatches. RegFormer [22, 61] designs a projection-aware Transformer, incorporating the cross-attention and all-to-all point-gathering strategy for reliable cross-frame correlation. To address density disparities, GCLNet [62] introduces a group-wise contrastive learning to extract density-invariant geometric features for registration. SDMNet [63] develops a sparse-to-dense registration network with a sparse matching stage and a local-dense matching stage. Darls [64] designs a density adaptive registration method for large-scale point clouds. To improve generalization, UGP [65] proposes inter-frame cross attention and progressive self attention. However, these previous works typically take pair-wise-only point pairs and neglect the inherent temporal consistency within LiDAR sequences. In this paper, we propose a novel outdoor registration model enhanced by streamlined temporal consistency.

Temporal Modeling with Memory. Sequential 4D point cloud streams have great potential in boosting various dense prediction tasks [66, 67, 68, 69]. Different from a static point cloud with limited observations, the additional temporal dimension can supplement incomplete measurements [26], display consistent motion priors [32], and further boost the performance of original single-frame tasks [66, 70]. Recently, increasing research has focused on storing temporal information with a memory mechanism. STMM [71] develops a recurrent unit with memory storage for object detection, modeling the long-term temporal appearance and motion dynamics. MeMOT [72] designs a memory-based multi-object tracking framework that utilizes a long-term spatio-temporal memory storing identity features of tracked objects. MemorySeg [26] proposes a 3D latent memory representation to improve current predictions with past frames. MEMFOF [73] and ARFlow [74] present memory-enhanced optical flow pipeline based on high-resolution predictions. Inspired by these works, we design both explicit and implicit memory buffers for temporal information storage.

3 Method

3.1 Overall Architecture

Given a stream of LiDAR point cloud sequences $\{PC_0, PC_1, ..., PC_{\mathrm{V}-1}\}^{\mathrm{V}\times\mathrm{N}\times 3}$, we propose a multi-frame registration framework as described in Fig. 2, where buffers storing multiple history memories would enhance the performance of each pair-wise registration. Specifically, to correlate current point pairs PC_{t} and $PC_{\mathrm{t}+1}$, hierarchical point features are first extracted, and then matched in the coarsest layer following HRegNet [23]. Afterwards, history motion information is read-out from memory buffers on both feature level and pose level in Sec. 3.2. To dynamically merge history and current features, we also design a dynamic history weighting module in Sec. 3.3. Finally, estimated coarse poses are iteratively refined, generating the final rotation matrix $\hat{R} \in \mathbb{R}^{3\times 3}$ and translation vector $\hat{t} \in \mathbb{R}^{3\times 1}$. Memory buffers are also dynamically updated with current features and current poses in Sec. 3.4. The network is supervised by the loss function defined in Sec. 3.5.

3.2 Memory Initialization and Read-out

Unlike pair-wise registration, streaming point cloud videos are the inputs of our network for long-term temporal modeling. To take advantage of temporal motion priors sufficiently, we design two memory buffers storing both cross-frame implicit feature-level memory and explicit pose-level memory. Our memory buffers $M_0^{\mathrm{F}} \in \mathbb{R}^{\mathrm{T}\times \mathrm{D_f}}$ of features and $M_0^{\mathrm{P}} \in \mathbb{R}^{\mathrm{T}\times \mathrm{D_p}}$ of poses are initialized by empty sets, where the temporal length of both memory buffers is maintained as T.

Implicit Feature Memory Read-out. With motion features F_{t} and poses R_{t}, t_{t} extracted from the coarse matching layer in the current timestamp, we first read-out the temporally-stored features using a sliding window mechanism. When current motion features are delivered into the Memory Feature Buffer $M_{\mathrm{t}-1}^{\mathrm{F}}$, the temporally-farthest feature in the buffer $\hat{F}_{\mathrm{t}-\mathrm{T}-1}$ is discarded. Other $T-1$

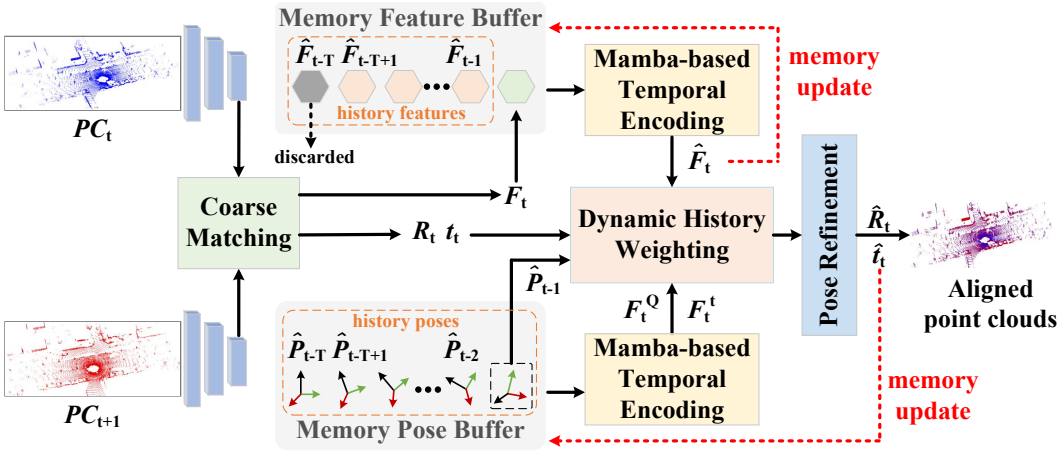


Figure 2: **The overview architecture of MemReg.** We leverage the long-term temporal features from history frames to enhance the current registration. During the current registration, memory features and memory poses are first read-out and interacted with current motion features through the Mamba-based Temporal Encoding module and Dynamic History Weighting module. Then, the current refined poses will, in turn, update memory buffers for next-timestamp estimation.

temporally-nearest features $\{\hat{F}_{t-T}, ..., \hat{F}_{t-1}\}$ are concatenated with the current motion feature F_t as:

$$M^{F}_{t-1} : \{\hat{F}_{t\text{-}T-1}, \hat{F}_{t\text{-}T+1}, ..., \hat{F}_{t\text{-}1}\} \Rightarrow \{\hat{F}_{t\text{-}T}, ..., \hat{F}_{t\text{-}1}, F_t\}, \tag{1}$$

where $\{\}$ indicates feature concatenation along the temporal dimension. The concatenated features formulate the input tokens of the following Mamba-based Temporal Encoding [34, 75]:

$$\hat{M}^{F}_{t-1} = \text{LN}(M^{F}_{t-1}), \tag{2}$$

$$\overline{M}^{F}_{t} = \delta(\text{DW}(\text{Linear}(\hat{M}^{F}_{t-1}))), \tag{3}$$

$$\hat{M}^{F}_{t} = \delta(\text{Linear}(\hat{M}^{F}_{t-1})), \tag{4}$$

$$\hat{F}_t = \text{Linear}(\text{SSM}(\overline{M}^{F}_{t})) \odot \hat{M}^{F}_{t} + M^{F}_{t-1}, \tag{5}$$

where δ denotes the SiLU activation [76]. DW is the depth-wise convolution [77]. LN indicates the Layer Normalization. SSM indicates the standard selective state space model [34]. The output $\hat{F}_t$ is the temporally encoded feature.

Explicit Pose Memory Read-out. T-length nearest Memory Pose Buffer M^{P}_{t-1} consists of historically estimated rotation matrices $M^{R}_{t-1} = \{\hat{Q}_{t\text{-}T-1}, \hat{Q}_{t\text{-}T}, ..., \hat{Q}_{t\text{-}1}\}$ and translation matrices $M^{t}_{t-1} = \{\hat{t}_{t\text{-}T-1}, \hat{t}_{t\text{-}T}, ..., \hat{t}_{t\text{-}1}\}$, which are respectively embedded by two separate MLP layers first to lift their dimensions. To facilitate regression, the rotation matrix here is represented by quaternion vectors $\hat{Q}_t \in \mathbb{R}^{4\times1}$. Finally, the embedded rotation features and translation features are fed into two Mamba blocks for the temporal modeling, respectively:

$$F^{Q}_{t} = \text{Mamba}(\text{MLP}(\{\hat{Q}_{t\text{-}T-1}, \hat{Q}_{t\text{-}T}, ..., \hat{Q}_{t\text{-}1}\})), \tag{6}$$

$$F^{t}_{t} = \text{Mamba}(\text{MLP}(\{\hat{t}_{t\text{-}T-1}, \hat{t}_{t\text{-}T}, ..., \hat{t}_{t\text{-}1}\})), \tag{7}$$

where $\hat{Q}_t$ and $\hat{t}_t$ respectively denote the t-th temporally-nearest quaternion and translation vectors stored in the Memory Pose Buffer. Here, Mamba indicates the same operation as in Memory Feature Buffer, about which we avoid detailed descriptions for simplicity.

3.3 Dynamic History Weighting

After retrieving the history-interacted features, the key problem is how to merge history-interacted features with the current motion features for enhanced registration performance. Since there may be scale gaps, the history and current poses should be re-scaled for effective information merging. Inspired by the great success of the dynamic weighting mechanism in [36, 37], a dynamic history

weighting layer is introduced to adaptively merge the history pose-related features, history poses, current pose-related features, and current poses.

To be specific, the inputs of our dynamic history weighting layer are composed of the history pose $\hat{Q}_{t-1}, \hat{t}_{t-1}$ from the latest previous timestamp, history pose-encoded features F_t^Q, F_t^t, current pose R_t, t_t, and current pose-encoded features $\hat{F}_t$. The current rotation matrix R_t is also converted to the quaternion vector Q_t for consistency. History and current pose-related features are first concatenated:

$$F_d^Q = \hat{Q}_{t-1} \oplus F_t^Q \oplus Q_t \oplus \hat{F}_t, \tag{8}$$

where $\oplus$ indicates concatenation along the channel dimension. Then, the dynamic weighting head is designed as:

$$W_1 = W_{d_1} F_d^Q, \quad W_2 = W_{d_2} F_d^Q, \tag{9}$$

$$F_{d_1}^Q = \mathrm{ReLU}(\mathrm{Norm}(W_1 F_d^Q)), \quad Q_t^d = W_2 F_{d_1}^Q, \tag{10}$$

where W_{d_1} and W_{d_2} are trainable parameters, and the output dimension is set to the same as the quaternion vector. Here, we only give descriptions of the quaternion vector weighting, and the same applies to the translation matrix weighting, generating the weighted translation vector t_t^d. By providing history pose-related features to the dynamic prediction head as inputs, the learnable weights enable the network to adjust to different scales dynamically and re-weight current pose estimations with history motion priors.

3.4 Memory Update and Pose Refinement

Our fixed-length memory buffers follow the sliding window mechanism, which are progressively updated as streaming point frames are delivered sequentially into the network. For the Memory Feature Buffer update, we directly concatenate the encoded current features $\hat{F}_t$ with T-1 previous features:

$$M_{\mathrm{t}}^{\mathrm{F}} = \{\hat{F}_{\mathrm{t\text{-}T+1}}, ..., \hat{F}_{\mathrm{t\text{-}1}}, \hat{F}_{\mathrm{t}}\}. \tag{11}$$

For the Memory Pose Buffer update, the pose refinement in [23] is first adopted to refine the poses Q_t^d and t_t^d to get the final pose estimation $\hat{R}_t$ and $\hat{t}_t$. The Memory Pose Buffer is then updated by concatenating previous T-1 poses and the current pose estimate as:

$$M_{\mathrm{t}}^{\mathrm{R}} = \{\hat{R}_{\mathrm{t\text{-}T+1}}, ..., \hat{R}_{\mathrm{t\text{-}1}}, \hat{R}_t\}, \quad M_{\mathrm{t}}^{\mathrm{t}} = \{\hat{t}_{\mathrm{t\text{-}T+1}}, ..., \hat{t}_{\mathrm{t\text{-}1}}, \hat{t}_t\}. \tag{12}$$

3.5 Loss Function

Our network is supervised by reducing the discrepancy for each pair-wise registration sample [23, 22]. The rotation and translation loss functions are:

$$\mathcal{L}_{\text{rot}} = ||\hat{R}^T R - \mathrm{I}||_2, \quad \mathcal{L}_{\text{trans}} = ||\hat{t} - t||_2, \tag{13}$$

where $\hat{R}, \hat{t}$ indicate the estimated rotation and translation matrices. R, t are the ground truth rotation and translation matrices. I denotes the identity matrix.

The overall loss function is: $\mathcal{L} = \mathcal{L}_{\text{trans}} + \mathcal{L}_{\text{rot}}$.

4 Experiment

4.1 Experimental Settings

Datasets and Evaluation Metrics. We conduct outdoor evaluations on three commonly used LiDAR datasets: KITTI odometry [38], nuScenes [39], and Apollo-Southbay [40]. To further demonstrate the generalization ability of our method, we also utilize indoor datasets including ScanNet [42], 3DMatch, and 3DLoMatch [41] for evaluation. Following the protocols in previous works [23, 63, 22], we use three main metrics to assess our method: Relative Translation Error (RTE), Relative Rotation Error (RRE), and Registration Recall. A successful registration is decided only when the RTE and RRE are both within certain thresholds θ_T, θ_R. To avoid the unreliable error metric report by minor failure samples, we follow previous works [3, 25, 22] to calculate the average RTE and RRE only for successful registration samples.

Table 1: **Outdoor point cloud registration on the KITTI dataset [38].** '*' indicates the baseline on which we introduce multi-frame memory.

Methods	KITTI Odometry			
	RTE (m)	RRE (deg)	Recall	Time (ms)
ICP (P2P) [43]	0.045 ± 0.054	0.112 ± 0.093	14.25%	472.2
ICP (P2Pl) [43]	0.044 ± 0.041	0.145 ± 0.153	33.56%	461.7
RANSAC [78]	0.126 ± 0.067	0.538 ± 0.396	91.94%	459.4
FGR [79]	0.929 ± 0.592	0.963 ± 0.806	39.42%	506.1
IDAM [80]	0.659 ± 0.483	1.057 ± 0.939	70.92%	33.4
FMR [81]	0.657 ± 0.422	1.493 ± 0.847	90.57%	85.5
DGR [82]	0.322 ± 0.319	0.374 ± 0.302	98.71%	1496.6
HRegNet [3]	0.056 ± 0.075	0.178 ± 0.196	99.77%	106.2
RegFormer [22]	0.082 ± 0.112	0.228 ± 0.209	99.80%	98.3
OKR-Net [83]	0.070 ± 0.130	0.230 ± 0.220	99.22%	100.4
HRegNet* [23]	0.047 ± 0.037	0.147 ± 0.120	**100.0**%	136.0
HDMNet [25]	0.050 ± 0.057	0.159 ± 0.152	99.85%	120.2
PTT [84]	0.063	0.230	99.80%	-
UPG [65]	0.071	0.245	99.80%	-
LSReg-Net [85]	0.060 ± 0.070	0.200 ± 0.170	99.92%	92.5
Ours	$\mathbf{0.032 \pm 0.036}$	$\mathbf{0.122 \pm 0.096}$	**100.0**%	133.0

Table 2: **Outdoor point cloud registration on the nuScenes dataset [39].** '*' indicates the baseline on which we introduce multi-frame memory.

Methods	nuScenes			
	RTE (m)	RRE (deg)	Recall	Time (ms)
ICP (P2P) [43]	0.252 ± 0.510	0.253 ± 0.502	18.78%	82.0
ICP (P2Pl) [43]	0.153 ± 0.296	0.212 ± 0.306	36.8%	44.5
RANSAC [78]	0.206 ± 0.186	0.738 ± 0.704	60.91%	187.4
FGR [79]	0.708 ± 0.622	1.007 ± 0.924	32.2%	284.6
DCP [86]	1.087 ± 0.491	2.065 ± 1.141	58.58%	45.5
IDAM [80]	0.467 ± 0.410	0.793 ± 0.783	87.98%	32.6
FMR [81]	0.603 ± 0.391	1.610 ± 0.974	92.06%	61.1
DGR [82]	0.211 ± 0.183	0.476 ± 0.430	98.41%	523.0
HRegNet [3]	0.122 ± 0.112	$0.273 \pm \mathbf{0.197}$	**100.0**%	87.3
RegFormer [22]	0.198	0.223	99.90%	85.6
OKR-Net [83]	0.140 ± 0.130	0.280 ± 0.210	99.96%	60.0
HRegNet* [23]	0.110 ± 0.096	0.285 ± 0.209	**100.0**%	120.1
HDMNet [25]	0.114 ± 0.102	0.274 ± 0.206	**100.0**%	102.9
UPG [65]	0.179	0.270	**100.0%**	-
LSReg-Net [85]	0.150 ± 0.130	0.320 ± 0.220	99.98%	81.2
Ours	$\mathbf{0.069 \pm 0.055}$	$\mathbf{0.261 \pm 0.197}$	**100.0**%	119.3

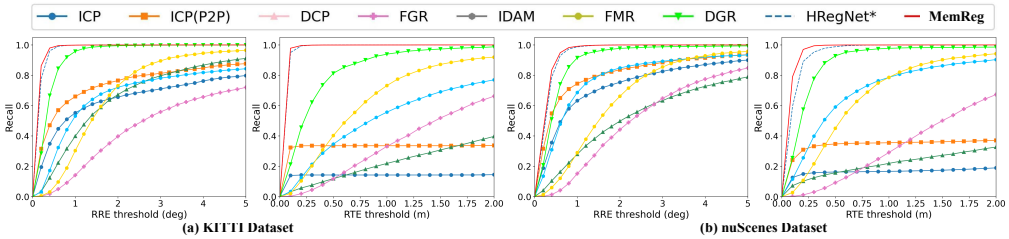


Figure 3: **The registration recall under different RRE or RTE thresholds on the KITTI and nuScenes datasets.**

Implementation Details. We first voxelize the input points with a voxel size of 0.3 m. Following [3, 63, 25], down-sampling is adopted for the input points, where 16,384 points are randomly sampled for KITTI and Apollo-Southbay, and 8,192 points for nuScenes. Experiments are conducted on a single NVIDIA RTX3090 GPU. The Adam optimizer is adopted with $\beta_1 = 0.9$, $\beta_2 = 0.999$. The initial learning rate is 0.001 and exponentially decays every 200,000 steps until reaching 0.00001. The batch size is set to 8. We choose HRegNet* [23] here as the baseline, on which we introduce multi-frame temporal information. Registration is successful when $\theta_T = 5$m, $\theta_R = 2$ degree (deg) by default. The temporal length in memory buffers T is set to 20 for KITTI and Apollo-Southbay, and 10 for nuScenes.

4.2 Quantitative Results

Comparison Results on KITTI. As in Table 1, our model not only outperforms previous optimization methods, e.g., ICP [43], RANSAC [78], by a large margin, but also learning-based methods. Compared to the previous SOTA method HRegNet* [23], our model achieves a 31.9% lower average RTE and a 17.0% lower average RRE enhanced by adding streaming temporal information. Also, compared to the baseline [23], our method requires a slightly shorter runtime for one inference. The experiment results demonstrate both the accuracy and efficiency of our proposed method.

Comparison Results on nuScenes. As in Table 2, our method achieves a 100% registration recall and outperforms previous optimization-based and learning-based methods with smaller registration errors, pushing the limit of performance further. Compared to the recent SOTA method HRegNet* [23], our model results in only half of its translation errors (0.069 m) and also a much lower rotation error (0.261 deg).

Comparison Results on Apollo-Southbay. As in Table 3, our method achieves 100% registration recall and significantly surpasses all previous methods, especially in the RTE metric, establishing a new SOTA in registration performance.

Comparison with Odometry or SLAM systems. We also compare with SLAM methods in Table 4. Our method not only surpasses standard odometry methods with sequential inputs like PWCLO

Table 3: **Outdoor point cloud registration performance on the Apollo-Southbay dataset [40].** '*' indicates the baseline method on which we introduce multi-frame temporal information.

Methods	Apollo-Southbay			
	RTE (m)	RRE (deg)	Recall	Time (ms)
ICP (P2P) [43]	0.100 ± 0.039	0.063 ± 0.305	39.93%	482.1
ICP (P2Pl) [43]	0.039 ± 0.179	0.046 ± 0.257	46.45%	470.2
RANSAC [78]	0.125 ± 0.114	0.361 ± 0.368	83.72%	552.1
FGR [79]	0.470 ± 0.498	0.664 ± 0.762	50.50%	496.0
DCP [86]	1.174 ± 0.499	2.155 ± 1.254	28.49%	41.0
IDAM [80]	0.456 ± 0.416	0.361 ± 0.429	74.03%	32.0
FMR [81]	0.653 ± 0.488	0.727 ± 0.658	83.78%	83.4
DGR [82]	0.132 ± 0.151	0.127 ± 0.146	99.64%	2507.0
HRegNet* [23]	0.034 ± 0.037	0.079 ± 0.079	99.88%	129.5
Ours	$\mathbf{0.024 \pm 0.031}$	$\mathbf{0.072 \pm 0.074}$	**100.0%**	127.6

Table 4: **Comparison with SLAM systems on KITTI odometry [38].** We report the average sequence translational RMSE (%). 'Multi.' indicates methods modeling multi-frame correlation.

Method	Multi.	07	08	09	10	Mean(07-10)
LONet [87]	✗	1.70	2.12	1.37	1.80	1.75
PWCLO [88]	✗	1.16	1.68	0.88	2.14	1.47
TransLO [12]	✗	0.55	1.29	0.95	1.18	0.99
DVLO [89]	✗	0.46	1.09	0.85	0.88	0.82
LAGLO [90]	✗	1.02	1.48	2.29	2.00	2.15
LOAM [91]	✓	0.69	1.18	1.20	1.51	1.15
DELO [92]	✓	0.58	1.36	1.23	1.53	1.18
DSLO [93]	✓	0.58	1.16	**0.72**	1.29	0.94
DVLO4D [16]	✓	0.43	0.95	0.77	0.76	0.73
Ours	✓	**0.38**	**0.81**	0.75	**0.72**	**0.67**

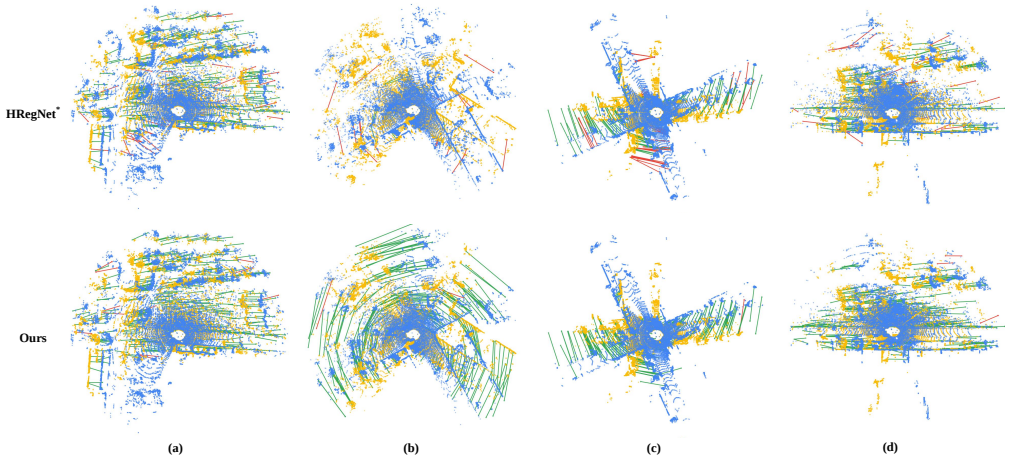


Figure 4: **Visualization of point cloud registration on KITTI.** Yellow and blue points indicate input point pair frames, with green lines showing accurate correspondences (inliers) and red lines showing inaccurate ones (outliers). This figure is best viewed in zoomed-in color mode. Compared to HRegNet* [23], our method MemReg estimates more accurate correspondences.

[88], DVLO [89], and LAGLO [90], but also outperforms methods that explicitly model multi-frame correlation like DELO [92], DSLO [93], and DVLO4D [16]. Our registration-only method also ha better registration results than LOAM [91], a standard LiDAR SLAM system with loop closure.

Results Under Stricter Thresholds. Since the threshold setting in [23] is rather loose, registration recalls are relatively saturated in these three datasets and the performance comparison with previous methods is unclear. Therefore, we vary a series of stricter thresholds in Fig. 3 for KITTI and nuScenes datasets, where our method consistently outperforms previous optimization- and learning-based methods.

4.3 Qualitative Results

Registration Samples. As shown in Fig. 4, compared to HRegNet* [23], introducing temporal information with memory buffers can significantly improve the ratio of successfully correlated keypoints. This demonstrates the effectiveness of our designed streaming temporal information on filtering outliers.

Weights in Dynamic History Weighting. We display the weights W_2 and corresponding 2D trajectories in Fig. 5, where the W_2 matrix is projected to a normalized 3D vector

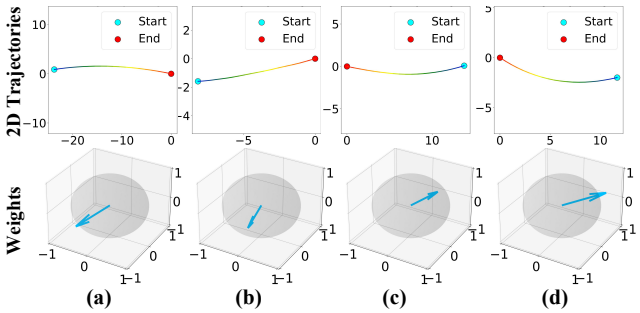


Figure 5: **Visualization of dynamic history weighting.**

Table 5: **Generalization to pair-wise indoor registration on 3DMatch and 3DLoMatch.** We introduce our memory buffers into a recent SOTA indoor baseline HybridReg [95] for evaluation.

Methods	3DMatch			3DLoMatch		
	RRE↓	RTE↓	RR↑	RRE↓	RTE↓	RR↑
FCGF [45]	2.0	6.6	85.1	3.2	10.0	40.1
D3Feat [101]	2.2	6.7	81.6	3.4	10.3	37.3
PREDATOR [18]	2.0	6.4	89.0	3.1	9.3	59.8
CoFiNet [102]	2.0	6.2	89.3	3.3	9.4	67.5
RegTR [20]	**1.6**	4.9	92.0	2.8	7.7	64.8
GeoTransformer [19]	**1.6**	5.3	91.5	2.5	**7.4**	**74.0**
HybridReg [95]	**1.6**	5.2	92.7	2.5	7.5	73.1
HybridReg + Ours	**1.6**	**4.8**	**92.9**	**2.4**	**7.4**	73.5

Table 6: **Generalization to multiview indoor registration on 3D(Lo)Match and ScanNet.** 'Rot.' and 'Trans.' are rotation and translation mean errors.

Methods	3DMatch	3DLoMatch	ScanNet	
	RR(%)↑	RR(%)↑	Rot. (°)↓	Trans. (m)↓
EIGSE3 [96]	40.1	26.5	40.6/37.1	0.88/0.84
L1-IRLS [97]	68.6	49.0	41.8/34.0	1.05/1.01
RotAvg [97]	77.2	63.0	38.5/31.6	0.96/0.83
LITS [98]	80.8	65.2	24.9/19.9	0.65/0.56
HARA [99]	83.8	79.1	34.7/31.3	0.86/0.17
SGHR [94]	96.2	81.6	21.7/19.0	0.56/0.49
SMVR [100]	96.2	82.0	19.8/17.5	0.55/0.45
SGHR + Ours	**97.1**	**82.8**	**19.2/16.3**	**0.52/0.44**

via PCA. Similar weights (lower figures) are observed in similar temporal trajectories (upper figures) such as (a)&(b), (c)&(d), demonstrating that our Dynamic History Weighting module can effectively learn weights from history trajectories.

4.4 Generalization to Indoor Registration

To demonstrate the generalization ability of our method, we also evaluate our method MemReg on indoor-level registration datasets: 3DMatch, 3DLoMatch [41], and ScanNet [42]. Current indoor methods can be divided into two categories in terms of different input frames: pairwise registration and multiview registration. As in Table 5, we first adapt our memory buffer module to general pairwise settings based on a recent SOTA method, HybridReg [95]. Compared to HybridReg, there is an obvious increase in registration recall and lower RRE, RTE on both 3DMatch and 3DLoMatch benchmarks. We also evaluate our method's compatibility in Table 6 on indoor multiview registration, where the network inputs are a series of partial indoor segments with overlap. We choose SGHR [94] as the baseline enhanced by our memory. Our method outperforms recent state-of-the-art methods [96, 97, 98, 99, 100], which are specially designed for the multiview indoor registration task. We also visualize the comparison in Fig. 6. Our method registers input scans better compared to SGHR [94], demonstrating our method's strong generalization ability for indoor-level scenes.

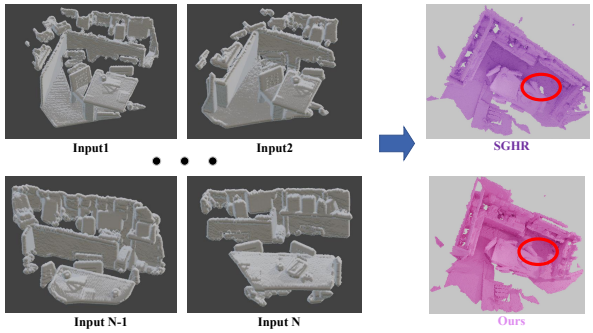


Figure 6: **Multiview indoor registration on 3DMatch [41].** We adapt our model to indoor scenes [94], where our MemReg has more accurate pose estimation for areas with the red circle than the baseline method SGHR.

4.5 Ablation Studies

In this section, we conduct comprehensive ablation studies to demonstrate the effectiveness of each proposed novelty.

Dynamic History Weighting (DHW). As in Table 7, removing DHW results in an 18.8% higher RTE and 11.5% higher RRE, indicating that DHW can effectively unify the history and current pose-related features. We also compare with another pose estimator in [89], which has much higher RTE and RRE.

Table 7: **Ablation studies of the Dynamic History Weighting (DHW) on KITTI.**

Model	RTE (m)	RRE (deg)	Time (ms)
w/o DHW	0.038 ± 0.038	0.136 ± 0.113	138
with estimator in [89]	0.036 ± 0.040	0.132 ± 0.104	135
with DHW	$\mathbf{0.032 \pm 0.036}$	$\mathbf{0.122 \pm 0.096}$	**133**

Significance of Temporal Memory Buffers. Table 8 shows ablation studies for two memory buffers. Without the Memory Pose Buffer (MPB), the average RTE increases by 31.3% and the average RRE

Table 8: **Ablation studies of the two memory buffers on KITTI.**

MFB	MPB	RTE (m)	RRE (deg)	Time (ms)
✗	✗	0.047 ± 0.046	0.164 ± 0.135	**126**
✓	✗	0.042 ± 0.042	0.156 ± 0.124	130
✗	✓	0.035 ± 0.039	0.126 ± 0.106	132
✓	✓	$\mathbf{0.032 \pm 0.036}$	$\mathbf{0.122 \pm 0.096}$	133

Table 9: **Ablation studies of different temporal lengths on KITTI.**

T	RTE (m)	RRE (deg)	Time (ms)
5	0.037 ± 0.041	0.136 ± 0.116	**128**
10	0.033 ± 0.036	0.127 ± 0.101	129
20	$\mathbf{0.032 \pm 0.036}$	$\mathbf{0.122 \pm 0.096}$	133
30	$\mathbf{0.032} \pm 0.038$	0.129 ± 0.105	140

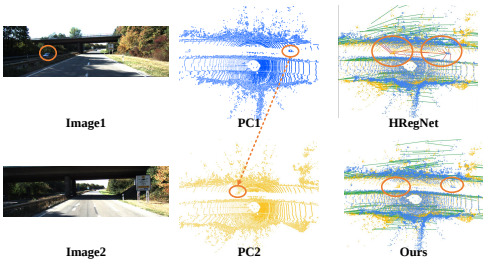


Figure 7: **Comparison on dynamics.** We evaluate on highway scenes with high dynamic cars (orange circled). HRegNet* [23] wrongly estimates correspondences on dynamics, while ours filters influence with no estimated correspondences on dynamics.

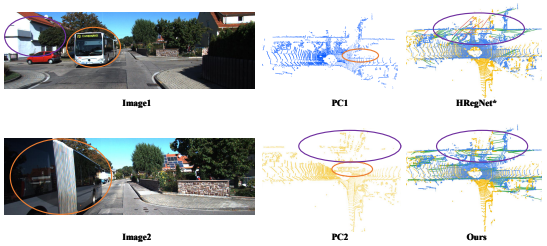


Figure 8: **Comparison on scenes with significant occlusions.** For scenes with large occluded areas by the bus (orange circled). HRegNet* [23] wrongly estimates correspondences for occluded objects (purple circled), but ours can filter their influence with no correspondences estimated on occluded objects.

by 27.9%. Removing the Memory Feature Buffer (MFB) results in a 9.4% higher average RTE and a 3.3% higher average RRE.

Different Temporal Lengths. Table 9 shows the comparison of different settings on temporal lengths in memory buffers. Accuracy increases largely from T=5 to 20 but is saturated when T>20. This is because 20 frames have enough historical observations for memory read-out. Points farther than 20 frames share less geometric and motion similarity, which may even degrade accuracy. More ablation studies about temporal interaction methods, rotation and translation thresholds are in the Appendix.

5 Discussions

Temporal information introduced by our designed memory buffers can address challenging cases where pair-wise registration works struggle, including low-overlap, dynamics, and occlusions.

Low-Overlap Inputs. Low-overlap registration is still an open issue because real-world LiDAR scans may have large distances or rotations, especially for low-frequency scanning LiDAR sensors, fast-moving cars, and street corners with large rotations. In these cases, the classical nearest neighboring strategy, e.g., KNN, may fail to capture precise point correspondences. In Table 10, we evaluate the low-overlap registration performance with larger frame intervals, which demonstrates the effectiveness of our designed memory buffers compared to the pair-wise baseline method HRegNet* [23].

Table 10: **Evaluation on low-overlap input pairs with 20-frame and 30-frame intervals on KITTI.**

Method	Intervals	RTE (m)	RRE (deg)	Recall
HRegNet* [23]	20-frame	0.085 ± 0.152	0.235 ± 0.213	96.2%
Ours	20-frame	$\mathbf{0.041 \pm 0.039}$	$\mathbf{0.148 \pm 0.127}$	**100.0%**
HRegNet* [23]	30-frame	0.123 ± 0.195	0.315 ± 0.302	76.3%
Ours	30-frame	$\mathbf{0.046 \pm 0.042}$	$\mathbf{0.169 \pm 0.141}$	**100.0%**

Dynamics and Occlusions. Dynamic or occluded objects introducing inconsistent motion patterns are commonly viewed as outliers in the outdoor registration task [22], which undermines the accuracy of consistent pose regression. We visualize the estimated correspondences for scenes with high dynamics in Fig. 7 and significant occlusions in Fig. 8. Our method can mitigate their influence with fewer or even no estimated correspondences on these outlier points. Therefore, the subsequent pose regression would be less influenced by these outliers.

6 Conclusion

In this paper, we introduced a memory buffer using streaming history features to enhance outdoor point cloud registration. Two temporal buffers are designed to store feature-level and pose-level memories from previous timestamps, exploiting long-term history motion priors to address challenging outdoor scenes. Additionally, a Mamba-based temporal interaction module and a dynamic history weighting module are developed to merge historical and current features. Extensive experiments demonstrate the effectiveness of our proposed method on KITTI, nuScenes, and Apollo-Southbay datasets. Our MemReg also has excellent generalization ability on the indoor registration benchmarks like 3DMatch, 3DLoMatch, and ScanNet.

7 Acknowledgments

This work was supported in part by the Natural Science Foundation of China under Grant 62225309,U24A20278, 62361166632, the Explorers Program of Shanghai (Basic Research Funding) under Grant 25TS1415100, in part by the Shanghai Municipal Special Program for Basic Research on General AI Foundation Models under Grant 2025SHZDZX025G01.

References

[1] Zixu Li, Yupeng Hu, Zhiwei Chen, Haokun Wen, Xuemeng Song, and Liqiang Nie. Combiner: Composed image retrieval guided by attribute-based neighbor relations. *IEEE Transactions on Image Processing*, 2026.

[2] Zixu Li, Yupeng Hu, Zhiwei Chen, Mingyu Zhang, Zhiheng Fu, and Liqiang Nie. Conesep: Cone-based robust noise-unlearning compositional network for composed image retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16897–16909, June 2026.

[3] Fan Lu, Guang Chen, Yinlong Liu, Lijun Zhang, Sanqing Qu, Shu Liu, and Rongqi Gu. Hregnet: A hierarchical network for large-scale outdoor lidar point cloud registration. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 16014–16023, 2021.

[4] Hengrui Zhang, Yuhu Cheng, CL Chen, and Xuesong Wang. Diffusion subgoal planning for long-horizon offline goal-conditioned reinforcement learning. *arXiv preprint arXiv:2609.34575*, 2026.

[5] Mengmeng Liu, Diankun Zhang, Jiuming Liu, Jianfeng Cui, Hongwei Xie, Guang Chen, Hangjun Ye, Michael Ying Yang, Francesco Nex, and Hao Cheng. Driveva: Video action models are zero-shot drivers. In *European Conference on Computer Vision*, pages 315–335. Springer, 2026.

[6] Chaokang Jiang, Desen Zhou, Jiuming Liu, and Li Sun. Vectorworld: Efficient streaming world model via diffusion flow on vector graphs. In *Forty-third International Conference on Machine Learning*, 2026.

[7] Mengmeng Liu, Diankun Zhang, Jiuming Liu, Jianfeng Cui, Hongwei Xie, Guang Chen, Hangjun Ye, Francesco Nex, Hao Cheng, and Michael Ying Yang. Universe: Unified video action models for autonomous driving with flexible mask-modulated modality generation. *arXiv preprint arXiv:2607.05133*, 2026.

[8] Mengmeng Liu, Hao Cheng, and Michael Ying Yang. Tracing the influence of predecessors on trajectory prediction. In *2023 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, pages 3245–3255. IEEE, 2023.

[9] Yifei Dong, Fengyi Wu, Yilong Dai, Lingdong Kong, Guangyu Chen, Yetong Sha, Qiyu Hu, Feng Liu, Siyu Huang, Qi Dai, and Zhi-Qi Cheng. Language-conditioned world modeling for visual navigation, 2026.

[10] Yifei Dong, Fengyi Wu, Guangyu Chen, Lingdong Kong, Qiyu Hu, Yuxuan Zhou, Xu Zhu, Jingdong Sun, Jun-Yan He, Qi Dai, Alexander G. Hauptmann, and Zhi-Qi Cheng. Towards unified world models for visual navigation via memory-augmented planning and foresight. In *European Conference on Computer Vision*, pages 276–295. Springer, 2026.

[11] Yifei Dong, Fengyi Wu, Qi He, Lingdong Kong, Heng Li, Minghan Li, Zebang Cheng, Yuxuan Zhou, Jingdong Sun, Qi Dai, Alexander G Hauptmann, and Zhi-Qi Cheng. Ha-vln 2.0: An open benchmark and leaderboard for human-aware navigation in discrete and continuous environments with dynamic multi-human interactions, 2026.

[12] Jiuming Liu, Guangming Wang, Chaokang Jiang, Zhe Liu, and Hesheng Wang. Translo: A window-based masked point transformer framework for large-scale lidar odometry. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 1683–1691, 2023.

[13] Tianchen Deng, Guole Shen, Tong Qin, Jianyu Wang, Wentao Zhao, Jingchuan Wang, Danwei Wang, and Weidong Chen. Plgslam: Progressive neural scene represenation with local to global bundle adjustment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19657–19666, 2024.

[14] Mengmeng Liu, Jiuming Liu, Michael Ying Yang, Chaokang Jiang, Jiangtao Li, Yunpeng Zhang, Hesheng Wang, Francesco Nex, and Hao Cheng. Streamvlo: Streaming visual-lidar odometry with cumulative drift compensation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 39086–39097, 2026.

[15] Siting Zhu, Guangming Wang, Hermann Blum, Jiuming Liu, Liang Song, Marc Pollefeys, and Hesheng Wang. Sni-slam: Semantic neural implicit slam. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 21167–21177. IEEE, 2024.

[16] Mengmeng Liu, Michael Ying Yang, Jiuming Liu, Yunpeng Zhang, Jiangtao Li, Sander Oude Elberink, George Vosselman, and Hao Cheng. Dvlo4d: Deep visual-lidar odometry with sparse spatial-temporal fusion. In *2025 IEEE International Conference on Robotics and Automation (ICRA)*, pages 9740–9747. IEEE, 2025.

[17] Tianchen Deng, Chang Nie, Shuhong Liu, Wenhua Wu, Jianfei Yang, Shenghai Yuan, Jiuming Liu, Zhe Liu, Danwei Wang, and Hesheng Wang. Voxelized 3D Gaussian representation for dense visual SLAM on embedded vision system. *International Journal of Computer Vision*, 134:445, 2026.

[18] Shengyu Huang, Zan Gojcic, Mikhail Usvyatsov, Andreas Wieser, and Konrad Schindler. Predator: Registration of 3d point clouds with low overlap. In *Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition*, pages 4267–4276, 2021.

[19] Zheng Qin, Hao Yu, Changjian Wang, Yulan Guo, Yuxing Peng, and Kai Xu. Geometric transformer for fast and robust point cloud registration. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11143–11152, 2022.

[20] Zi Jian Yew and Gim Hee Lee. Regtr: End-to-end point cloud correspondences with transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6677–6686, 2022.

[21] Sheng Ao, Qingyong Hu, Hanyun Wang, Kai Xu, and Yulan Guo. Buffer: Balancing accuracy, efficiency, and generalizability in point cloud registration. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1255–1264, 2023.

[22] Jiuming Liu, Guangming Wang, Zhe Liu, Chaokang Jiang, Marc Pollefeys, and Hesheng Wang. Regformer: An efficient projection-aware transformer network for large-scale point cloud registration. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8451–8460, 2023.

[23] Fan Lu, Guang Chen, Yinlong Liu, Lijun Zhang, Sanqing Qu, Shu Liu, Rongqi Gu, and Changjun Jiang. Hregnet: A hierarchical network for efficient and accurate outdoor lidar point cloud registration. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(10):11884–11897, 2023.

[24] Jiuming Liu, Guangming Wang, Weicai Ye, Chaokang Jiang, Jinru Han, Zhe Liu, Guofeng Zhang, Dalong Du, and Hesheng Wang. Difflow3d: Toward robust uncertainty-aware scene flow estimation with iterative diffusion-based refinement. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15109–15119. IEEE, 2024.

[25] Weiyi Xue, Fan Lu, and Guang Chen. Hdmnet: A hierarchical matching network with double attention for large-scale outdoor lidar point cloud registration. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 3393–3403, 2024.

[26] Enxu Li, Sergio Casas, and Raquel Urtasun. Memoryseg: Online lidar semantic segmentation with a latent memory. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 745–754, 2023.

[27] Jiangwei Lao, Weixiang Hong, Xin Guo, Yingying Zhang, Jian Wang, Jingdong Chen, and Wei Chu. Simultaneously short-and long-term temporal modeling for semi-supervised video semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14763–14772, 2023.

[28] Xuechao Chen, Shuangjie Xu, Xiaoyi Zou, Tongyi Cao, Dit-Yan Yeung, and Lu Fang. Svqnet: Sparse voxel-adjacent query network for 4d spatio-temporal lidar semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8569–8578, 2023.

[29] Yu Zheng, Guangming Wang, Jiuming Liu, Marc Pollefeys, and Hesheng Wang. Spherical frustum sparse convolution network for lidar point cloud semantic segmentation. *Advances in Neural Information Processing Systems*, 37:121827–121858, 2024.

[30] Shihao Wang, Yingfei Liu, Tiancai Wang, Ying Li, and Xiangyu Zhang. Exploring object-centric temporal modeling for efficient multi-view 3d object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3621–3631, 2023.

[31] Jinghua Hou, Zhe Liu, dingkang liang, Zhikang Zou, Xiaoqing Ye, and Xiang Bai. Query-based temporal fusion with explicit motion for 3d object detection. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

[32] Ruopeng Gao and Limin Wang. Memotr: Long-term memory-augmented transformer for multi-object tracking. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9901–9910, 2023.

[33] Zhiwei Chen, Yupeng Hu, Zixu Li, Zhiheng Fu, Guozhi Qiu, Weili Guan, and Liqiang Nie. Egoadapt: A multi-scene egocentric adaptation method for cvpr 2026 hd-epic vqa challenge. *arXiv preprint arXiv:2605.24500*, 2026.

[34] Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces. *arXiv preprint arXiv:2312.00752*, 2023.

[35] Mengmeng Liu, Jiuming Liu, Yunpeng Zhang, Jiangtao Li, Michael Ying Yang, Francesco Nex, and Hao Cheng. 4dstr: Advancing generative 4d gaussians with spatial-temporal rectification for high-quality and consistent 4d generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 7224–7232, 2026.

[36] Xu Jia, Bert De Brabandere, Tinne Tuytelaars, and Luc V Gool. Dynamic filter networks. *Advances in neural information processing systems*, 29, 2016.

[37] Görkay Aydemir, Adil Kaan Akan, and Fatma Güney. Adapt: Efficient multi-agent trajectory prediction with adaptation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8295–8305, 2023.

[38] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *2012 IEEE conference on computer vision and pattern recognition*, pages 3354–3361. IEEE, 2012.

[39] Holger Caesar, Varun Bankiti, Alex H Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuscenes: A multimodal dataset for autonomous driving. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11621–11631, 2020.

[40] Weixin Lu, Yao Zhou, Guowei Wan, Shenhua Hou, and Shiyu Song. L3-net: Towards learning based lidar localization for autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6389–6398, 2019.

[41] Andy Zeng, Shuran Song, Matthias Nießner, Matthew Fisher, Jianxiong Xiao, and Thomas Funkhouser. 3dmatch: Learning local geometric descriptors from rgb-d reconstructions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1802–1811, 2017.

[42] Angela Dai, Angel X Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5828–5839, 2017.

[43] P.J. Besl and Neil D. McKay. A method for registration of 3-d shapes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 14(2):239–256, 1992.

[44] Haowen Deng, Tolga Birdal, and Slobodan Ilic. Ppfnet: Global context aware local features for robust 3d point matching. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 195–205, 2018.

[45] Christopher Choy, Jaesik Park, and Vladlen Koltun. Fully convolutional geometric features. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 8958–8966, 2019.

[46] Yasuhiro Aoki, Hunter Goforth, Rangaprasad Arun Srivatsan, and Simon Lucey. Pointnetlk: Robust & efficient point cloud registration using pointnet. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7163–7172, 2019.

[47] Zhen Dong, Fuxun Liang, Bisheng Yang, Yusheng Xu, Yufu Zang, Jianping Li, Yuan Wang, Wenxia Dai, Hongchao Fan, Juha Hyyppä, et al. Registration of large-scale terrestrial laser scanner point clouds: A review and benchmark. *ISPRS Journal of Photogrammetry and Remote Sensing*, 163:327–342, 2020.

[48] Juncheng Mu, Lin Bie, Shaoyi Du, and Yue Gao. Colorpcr: Color point cloud registration with multi-stage geometric-color fusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21061–21070, 2024.

[49] Shoujun Jia, Chun Liu, Hangbin Wu, Weihua Huan, and Shufan Wang. Incremental registration towards large-scale heterogeneous point clouds by hierarchical graph matching. *ISPRS Journal of Photogrammetry and Remote Sensing*, 213:87–106, 2024.

[50] Zixu Li, Yupeng Hu, Zhiwei Chen, Qinlei Huang, Guozhi Qiu, Zhiheng Fu, and Meng Liu. Retrack: Evidence-driven dual-stream directional anchor calibration network for composed video retrieval. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 23373–23381, 2026.

[51] Zixu Li, Yupeng Hu, Zhiheng Fu, Zhiwei Chen, Yongqi Li, and Liqiang Nie. Tema: Anchor the image, follow the text for multi-modification composed image retrieval. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 24421–24442, 2026.

[52] Hong Chen, Pei Yan, Sihe Xiang, and Yihua Tan. Dynamic cues-assisted transformer for robust point cloud registration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21698–21707, 2024.

[53] Yongzhe Yuan, Yue Wu, Xiaolong Fan, Maoguo Gong, Qiguang Miao, and Wenping Ma. Inlier confidence calibration for point cloud registration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5312–5321, 2024.

[54] Haobo Jiang, Mathieu Salzmann, Zheng Dang, Jin Xie, and Jian Yang. Se (3) diffusion model-based point cloud registration for robust 6d object pose estimation. *Advances in Neural Information Processing Systems*, 36:21285–21297, 2023.

[55] Kexue Fu, Mingzhi Yuan, Changwei Wang, Weiguang Pang, Jing Chi, Manning Wang, and Longxiang Gao. Dual focus-attention transformer for robust point cloud registration. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 11769–11778, 2025.

[56] Shuo Lu, Yinuo Xu, Kecheng Yu, Siru Jiang, Yongcan Yu, Yubin Wang, Haitao Yang, Yuxiang Zhang, Bin Wang, Ran He, et al. Worldcoder-bench: Benchmarking physically grounded 3d world synthesis. *arXiv preprint arXiv:2606.01869*, 2026.

[57] Linrui Gong, Jiuming Liu, Junyi Ma, Lihao Liu, Yaonan Wang, and Hesheng Wang. Eadreg: Probabilistic correspondence generation with efficient autoregressive diffusion model for outdoor point cloud registration. *arXiv preprint arXiv:2411.15271*, 2024.

[58] Jiuming Liu, Weicai Ye, Guangming Wang, Chaokang Jiang, Lei Pan, Jinru Han, Zhe Liu, Guofeng Zhang, and Hesheng Wang. Difflow3d: Hierarchical diffusion models for uncertainty-aware 3d scene flow estimation. *IEEE transactions on pattern analysis and machine intelligence*, 2025.

[59] Chaokang Jiang, Guangming Wang, Jiuming Liu, Hesheng Wang, Zhuang Ma, Zhenqiang Liu, Zhujin Liang, Yi Shan, and Dalong Du. 3dsflabelling: Boosting 3d scene flow estimation by pseudo auto-labelling. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15173–15183. IEEE, 2024.

[60] Jiuming Liu, Zheng Huang, Mengmeng Liu, Tianchen Deng, Francesco Nex, Hao Cheng, and Hesheng Wang. Topolidm: Topology-aware lidar diffusion models for interpretable and realistic lidar point cloud generation. In *2025 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 8180–8186. IEEE, 2025.

[61] Jiuming Liu, Guangming Wang, Zhe Liu, Chaokang Jiang, Haoang Li, Mengmeng Liu, Tianchen Deng, Marc Pollefeys, Michael Ying Yang, and Hesheng Wang. Regformer++: An efficient large-scale 3d lidar point registration network with projection-aware 2d transformer. *arXiv preprint arXiv:2603.14290*, 2026.

[62] Quan Liu, Hongzi Zhu, Yunsong Zhou, Hongyang Li, Shan Chang, and Minyi Guo. Density-invariant features for distant point cloud registration. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 18215–18225, 2023.

[63] Fan Lu, Guang Chen, Yinlong Liu, Yibing Zhan, Zhijun Li, Dacheng Tao, and Changjun Jiang. Sparse-to-dense matching network for large-scale lidar point cloud registration. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(9):11270–11282, 2023.

[64] Jiawei Wang, Yan Zhuang, Fei Yan, and Hong Zhang. Density adaptive registration of large-scale point clouds in diverse outdoor environments. *IEEE Robotics and Automation Letters*, 2025.

[65] Zhenxuan Zeng, Qiao Wu, Xiyu Zhang, Lin Yuanbo Wu, Pei An, Jiaqi Yang, Ji Wang, and Peng Wang. Unlocking generalization power in lidar point cloud registration. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 22244–22253, 2025.

[66] Haiyan Wang and Yingli Tian. Sequential point clouds: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.

[67] Jiuming Liu, Chaojun Ni, Mengmeng Liu, Chensheng Peng, Fangjinhua Wang, Sitian Shen, Marc Pollefeys, Masayoshi Tomizuka, Ayush Tewari, and Per Ola Kristensson. Towards interactive video world modeling: Frontiers, challenges, benchmarks, and future trends. *arXiv preprint arXiv:2606.01164*, 2026.

[68] Mengmeng Liu, Hao Cheng, Lin Chen, Hellward Broszio, Jiangtao Li, Runjiang Zhao, Monika Sester, and Michael Ying Yang. Laformer: Trajectory prediction for autonomous driving with lane-aware scene constraints. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 2039–2049. IEEE, 2024.

[69] Hao Cheng, Mengmeng Liu, Lin Chen, Hellward Broszio, Monika Sester, and Michael Ying Yang. Gatraj: A graph-and attention-based multi-agent trajectory prediction model. *ISPRS Journal of Photogrammetry and Remote Sensing*, 205:163–175, 2023.

[70] Zhiheng Fu, Zixu Li, Zhiwei Chen, Fangxu Liu, Yupeng Hu, Weili Guan, and Liqiang Nie. Egoaction: Egocentric action composition with reliability-aware temporal fusion for the epic-kitchens action detection challenge at cvpr 2026. *arXiv preprint arXiv:2605.24496*, 2026.

[71] Fanyi Xiao and Yong Jae Lee. Video object detection with an aligned spatial-temporal memory. In *Proceedings of the European conference on computer vision (ECCV)*, pages 485–501, 2018.

[72] Jiarui Cai, Mingze Xu, Wei Li, Yuanjun Xiong, Wei Xia, Zhuowen Tu, and Stefano Soatto. Memot: Multi-object tracking with memory. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8090–8100, 2022.

[73] Vladislav Bargatin, Egor Chistov, Alexander Yakovenko, and Dmitriy Vatolin. Memfof: High-resolution training for memory-efficient multi-frame optical flow estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 8187–8196, October 2025.

[74] Jiuming Liu, Mengmeng Liu, Siting Zhu, Yunpeng Zhang, Jiangtao Li, Michael Yang, Francesco Nex, Hao Cheng, and Hesheng Wang. Arflow: Auto-regressive optical flow estimation for arbitrary-length videos via progressive next-frame forecasting. In *International Conference on Learning Representations*, volume 2026, pages 130125–130154, 2026.

[75] Jiuming Liu, Jinru Han, Lihao Liu, Angelica I Aviles-Rivero, Chaokang Jiang, Zhe Liu, and Hesheng Wang. Mamba4d: Efficient 4d point cloud video understanding with disentangled spatial-temporal state space models. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 17626–17636. IEEE, 2025.

[76] Dan Hendrycks and Kevin Gimpel. Gaussian error linear units (gelus). *arXiv preprint arXiv:1606.08415*, 2016.

[77] François Chollet. Xception: Deep learning with depthwise separable convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1251–1258, 2017.

[78] Martin A Fischler and Robert C Bolles. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, 24(6):381–395, 1981.

[79] Qian-Yi Zhou, Jaesik Park, and Vladlen Koltun. Fast global registration. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II 14*, pages 766–782. Springer, 2016.

[80] Jiahao Li, Changhao Zhang, Ziyao Xu, Hangning Zhou, and Chi Zhang. Iterative distance-aware similarity matrix convolution with mutual-supervised point elimination for efficient point cloud registration. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXIV 16*, pages 378–394. Springer, 2020.

[81] Xiaoshui Huang, Guofeng Mei, and Jian Zhang. Feature-metric registration: A fast semi-supervised approach for robust point cloud registration without correspondences. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11366–11374, 2020.

[82] Christopher Choy, Wei Dong, and Vladlen Koltun. Deep global registration. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2514–2523, 2020.

[83] Zijian Wang, Xiaosu Xu, Yiqing Yao, Nuo Li, and Yehao Liu. Okr-net: Overlapping keypoints registration network for large-scale lidar point clouds. *IEEE Robotics and Automation Letters*, 2023.

[84] Meiling Wang, Guangyan Chen, Yi Yang, Li Yuan, and Yufeng Yue. Point tree transformer for point cloud registration. *IEEE Transactions on Circuits and Systems for Video Technology*, 2025.

[85] Yucheng Tao, Xiuqing Yang, Hanqi Wang, Jian Wang, Zhiyuan Li, and Huawei Liang. Lsreg-net: An end-to-end registration network for large-scale lidar point cloud in autonomous driving. *IEEE Sensors Journal*, 2025.

[86] Yue Wang and Justin M Solomon. Deep closest point: Learning representations for point cloud registration. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 3523–3532, 2019.

[87] Qing Li, Shaoyang Chen, Cheng Wang, Xin Li, Chenglu Wen, Ming Cheng, and Jonathan Li. Lo-net: Deep real-time lidar odometry. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8473–8482, 2019.

[88] Guangming Wang, Xinrui Wu, Zhe Liu, and Hesheng Wang. Pwclo-net: Deep lidar odometry in 3d point clouds using hierarchical embedding mask optimization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 15910–15919, 2021.

[89] Jiuming Liu, Dong Zhuo, Zhiheng Feng, Siting Zhu, Chensheng Peng, Zhe Liu, and Hesheng Wang. Dvlo: Deep visual-lidar odometry with local-to-global feature fusion and bi-directional structure alignment. In *European Conference on Computer Vision*, pages 475–493. Springer, 2024.

[90] Fangzhou Tang, Shuting Zhang, Bocheng Zhu, and Junren Sun. Enhanced lidar odometry for autonomous vehicular positioning system using local feature enhancement and global motion constraint. *IEEE Transactions on Vehicular Technology*, 2025.

[91] Ji Zhang, Sanjiv Singh, et al. Loam: Lidar odometry and mapping in real-time. In *Robotics: Science and systems*, volume 2, pages 1–9. Berkeley, CA, 2014.

[92] Sk Aziz Ali, Djamila Aouada, Gerd Reis, and Didier Stricker. Delo: Deep evidential lidar odometry using partial optimal transport. In *2023 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, pages 4519–4528. IEEE, 2023.

[93] Huixin Zhang, Guangming Wang, Xinrui Wu, Chenfeng Xu, Mingyu Ding, Masayoshi Tomizuka, Wei Zhan, and Hesheng Wang. Dslo: Deep sequence lidar odometry based on inconsistent spatio-temporal propagation. In *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 10672–10677. IEEE, 2024.

[94] Haiping Wang, Yuan Liu, Zhen Dong, Yulan Guo, Yu-Shen Liu, Wenping Wang, and Bisheng Yang. Robust multiview point cloud registration with reliable pose graph initialization and history reweighting. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9506–9515, 2023.

[95] Keyu Du, Hao Xu, Haipeng Li, Hong Qu, Chi-Wing Fu, and Shuaicheng Liu. Hybridreg: Robust 3d point cloud registration with hybrid motions. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 2789–2797, 2025.

[96] Federica Arrigoni, Beatrice Rossi, and Andrea Fusiello. Spectral synchronization of multiple views in se (3). *SIAM Journal on Imaging Sciences*, 9(4):1963–1990, 2016.

[97] Avishek Chatterjee and Venu Madhav Govindu. Robust relative rotation averaging. *IEEE transactions on pattern analysis and machine intelligence*, 40(4):958–972, 2017.

[98] Zi Jian Yew and Gim Hee Lee. Learning iterative robust transformation synchronization. In *2021 International Conference on 3D Vision (3DV)*, pages 1206–1215. IEEE, 2021.

[99] Seong Hun Lee and Javier Civera. Hara: A hierarchical approach for robust rotation averaging. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15777–15786, 2022.

[100] Li Fang, Tianyu Li, Shudong Zhou, and Yanghong Lin. Robust multiview point cloud registration using algebraic connectivity and spatial compatibility. *IEEE Transactions on Geoscience and Remote Sensing*, 2024.

[101] Xuyang Bai, Zixin Luo, Lei Zhou, Hongbo Fu, Long Quan, and Chiew-Lan Tai. D3feat: Joint learning of dense detection and description of 3d local features. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6359–6367, 2020.

[102] Hao Yu, Fu Li, Mahdi Saleh, Benjamin Busam, and Slobodan Ilic. Cofinet: Reliable coarse-to-fine correspondences for robust pointcloud registration. *Advances in Neural Information Processing Systems*, 34:23872–23884, 2021.

[103] Charles Ruizhongtai Qi, Li Yi, Hao Su, and Leonidas J Guibas. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. *Advances in neural information processing systems*, 30, 2017.

[104] Yue Liu, Yunjie Tian, Yuzhong Zhao, Hongtian Yu, Lingxi Xie, Yaowei Wang, Qixiang Ye, Jianbin Jiao, and Yunfan Liu. Vmamba: Visual state space model. *Advances in neural information processing systems*, 37:103031–103063, 2024.

[105] Kunchang Li, Xinhao Li, Yi Wang, Yinan He, Yali Wang, Limin Wang, and Yu Qiao. Videomamba: State space model for efficient video understanding. In *European conference on computer vision*, pages 237–255. Springer, 2024.

[106] Dingkang Liang, Xin Zhou, Wei Xu, Xingkui Zhu, Zhikang Zou, Xiaoqing Ye, Xiao Tan, and Xiang Bai. Pointmamba: A simple state space model for point cloud analysis. *Advances in neural information processing systems*, 37:32653–32677, 2024.

[107] Jiuming Liu, Ruiji Yu, Yian Wang, Yu Zheng, Tianchen Deng, Weicai Ye, and Hesheng Wang. Point mamba: A novel point cloud backbone based on state space model with octree-based ordering strategy. *arXiv preprint arXiv:2403.06467*, 2024.

[108] Rui She, Sijie Wang, Qiyu Kang, Kai Zhao, Yang Song, Wee Peng Tay, Tianyu Geng, and Xingchao Jian. Posdiffnet: Positional neural diffusion for point cloud registration in a large field of view with perturbations. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 231–239, 2024.

[109] Guangyan Chen, Meiling Wang, Yi Yang, Li Yuan, and Yufeng Yue. Fast and robust point cloud registration with tree-based transformer. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 773–780. IEEE, 2024.

[110] Shaocheng Yan, Pengcheng Shi, and Jiayuan Li. Ml-semreg: Boosting point cloud registration with multi-level semantic consistency. In *European Conference on Computer Vision*, pages 19–37. Springer, 2024.

[111] Yue Pan, Tao Sun, Liyuan Zhu, Lucas Nunes, Iro Armeni, Jens Behley, and Cyrill Stachniss. Register any point: Scaling 3d point cloud registration by flow matching. In *European Conference on Computer Vision*, pages 21–43. Springer, 2026.

A Overview

The supplementary materials are structured as follows:

- We first provide more details about the network architecture in Sec. B.
- The evaluation metrics are described in Sec. C.
- Additional experiment results are presented in the Sec. D to demonstrate the superiority of our method and the effectiveness of the proposed contributions.
- More visualizations on the KITTI, nuScenes, and Apollo-Southbay datasets are provided in Sec. E.
- Additionally, we append the video demos of the outdoor registration performance on four sequences of KITTI (Seq.01, Seq.08, Seq.09, Seq.10) to the supplementary materials.

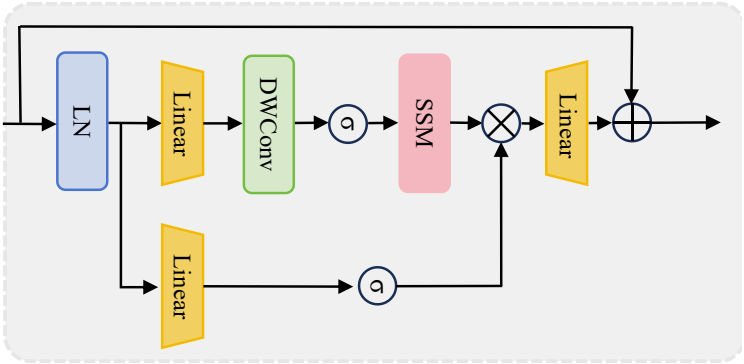


Figure 9: **The architecture of the Mamba module.** We utilize the standard Mamba [34] module without any task-specific design.

B Network Architecture

B.1 Feature Extraction and Coarse Matching

We follow HRegNet [23] to construct the hierarchical point feature extraction and coarse point matching layers.

Feature Extraction. Hierarchical point features are extracted as in [23]. For each layer, Weighted Farthest Point Sampling (WFPS) [103] is first utilized to sample a set of center keypoints. Around each keypoint, K Nearest Neighbor (KNN) and Shared Multi-layer Perceptron (Shared MLP) are then used to search and aggregate the features of its k nearest neighboring points:

$$f_i = \max_{k=1,2,\cdots,\mathrm{K}} \mathrm{MLP}((x_i^k - x_i) \oplus f_i^k), \tag{14}$$

where x_i^k denotes the k-th nearest point and f_i^k is the corresponding point feature. Max is the Max-pooling operation.

Coarse Point Matching. The core issue in point cloud registration is to find precise and reliable correspondences between two frames. At the coarsest layer, where points commonly possess the largest receptive field, finding accurate correspondences can guarantee the generally reliable registration performance. To achieve this, we consider geometric features, descriptor features, and similarity features from current point pairs as in HRegNet [23]. Specifically, for each keypoint in PC_{t} as the clustering center, we perform KNN to select a set of neighboring candidate points in $PC_{\mathrm{t+1}}$. The center keypoint and its K neighboring points form a cluster for feature embedding. Geometric features consist of both the absolute coordinates and the relative distance of the keypoint and its neighboring points. Descriptor features include the saliency uncertainties of keypoints and also the local descriptors for each keypoint and its neighboring points. Similarity features are considered in both bilateral consensus and neighborhood consensus. Bilateral consensus ensures the bi-directional optimal matching scores in both the forward (keypoints in PC_{t} searching KNN in $PC_{\mathrm{t+1}}$) and the reverse (keypoints in $PC_{\mathrm{t+1}}$ searching KNN in PC_{t}) processes. Neighborhood consensus indicates that the neighboring keypoints around the correspondence keypoints should possess similar features. The outputs of the coarse matching layer are cross-frame features F_t and initially estimated pose R_t, t_t. However, these features and poses contain no temporal information, which will interact with and update temporal memory buffers (Secs. 3.3-3.5 in the main paper).

B.2 Mamba Architecture

Recently, a novel architecture named Mamba [34] has shown promising effectiveness, especially for long-sequence modeling ability, outperforming Transformer-based counterparts in image processing [104], video understanding [105], and point cloud [106, 107] fields. In this paper, we adopt the vanilla Mamba architecture [34] in our temporal encoding layer as illustrated in Fig. 9. Without any task-specific designs, our Mamba-based temporal encoding turns out to be effective enough to embed long-term dependencies for history and current motion features.

Table 11: **Implementation details of coarse matching and pose refinement layers.** In the coarse matching and pose refinement layers, 'N' denotes the number of the selected neighboring points in KNN. 'MLP1' and 'MLP2' represent the output dimension of two MLP layers, respectively, for correspondence prediction.

Module	Layer	N	MLP1	MLP2
Coarse Matching	Layer 1	8	[128, 128, 128]	[128, 128, 1]
Module	Layer	N	MLP1	MLP2
Pose Refinement	Layer 2	8	[256, 256, 256]	[256, 256, 1]
	Layer 3	8	[512, 512, 512]	[512, 512, 1]

Table 12: **More comparison results on the KITTI odometry [38].**

Methods	KITTI Odometry		
	RTE (m)	RRE (deg)	Recall
PosDiffNet [108]	0.066	0.240	99.80%
TrT-Net [109]	0.063	0.230	99.80%
ML-SemReg [110]	0.052	0.200	97.91%
RAP [111]	0.043	0.139	**100.0%**
Ours	**0.032**	**0.122**	**100.0**%

B.3 Pose Refinement Layer

We adopt the same pose refinement process as in HRegNet [23], where finer registration layers only consider the geometric features and descriptor features. Similarity features are abandoned due to high computational costs. With the correlated features, a soft correspondence layer is introduced to predict keypoint correspondences and their confidence scores [3]. Finally, the Weighted Kabsch algorithm is applied to calculate the transformation matrices ΔR and Δt. Without loss of generality, taking layer l as an example, the refinement layer can be represented by:

$$R^l = \Delta R \times R^{l+1}, \tag{15}$$

$$t^l = \Delta R \times t^{l+1} + \Delta t, \tag{16}$$

where R^l, t^l indicate the refined poses at a finer layer. R^{l+1}, t^{l+1} indicate the un-refined poses at a coarser layer. More implementation details are provided in Table 11.

C Evaluation Metrics

Here, we provide the equations to calculate RTE (Relative Translation Error) and RRE (Relative Rotation Error). RTE is calculated by:

$$\text{RTE} = ||\hat{t} - t||_2. \tag{17}$$

RRE is defined as:

$$\text{RRE} = \arccos(\text{Tr}(\hat{R}^{\text{T}} R - 1)/2), \tag{18}$$

where $\hat{R}, \hat{t}$ indicate the respective estimated rotation matrix and translation vector, and R, t are the ground truth rotation matrix and translation vector.

D Additional Experimental Results

More Comparison Results. We also supplement more comparisons with recent state-of-the-art methods including PosDiffNet [108], TrT-Net [109], ML-SemReg [110], and RAP [111]. Since they only release the average rotation and translation metrics, we also follow them for a fair comparison. As shown in Table 12, our method has substantial performance advantages compared to these methods, which demonstrates the superiority of our proposed temporal buffers.

Table 13: **Comparison results with HRegNet* [23] under different translation thresholds θ_T on the KITTI dataset.** θ_R is set to 1deg.

θ_T (m)	Method	RTE (m)	RRE (deg)	Recall
0.05	HRegNet*	0.0302 ± 0.0110	0.1583 ± 0.1155	72.08%
	Ours	$\mathbf{0.0243 \pm 0.0109}$	$\mathbf{0.1175 \pm 0.0837}$	**87.50%**
0.1	HRegNet*	0.0406 ± 0.0201	0.1606 ± 0.1195	95.60%
	Ours	$\mathbf{0.0294 \pm 0.0172}$	$\mathbf{0.1184 \pm 0.0889}$	**97.93%**
0.2	HRegNet*	0.0444 ± 0.0271	0.1613 ± 0.1192	97.62%
	Ours	$\mathbf{0.0310 \pm 0.0225}$	$\mathbf{0.1198 \pm 0.0881}$	**98.13%**
0.3	HRegNet*	0.0455 ± 0.0308	0.1616 ± 0.1194	99.52%
	Ours	$\mathbf{0.0318 \pm 0.0259}$	$\mathbf{0.1197 \pm 0.0879}$	**99.88%**

Table 14: **Comparison results with HRegNet* [23] under different rotation thresholds θ_R on the KITTI dataset.** θ_T is set to 0.05m.

θ_R (deg)	Method	RTE (m)	RRE (deg)	Recall
0.25	HRegNet*	0.0302 ± 0.0109	0.1210 ± 0.0627	57.20%
	Ours	$\mathbf{0.0241 \pm 0.0109}$	$\mathbf{0.1006 \pm 0.0538}$	**81.10%**
0.5	HRegNet*	0.0304 ± 0.0109	0.1524 ± 0.0973	67.38%
	Ours	$\mathbf{0.0242 \pm 0.0109}$	$\mathbf{0.1158 \pm 0.0781}$	**87.12%**
1	HRegNet*	0.0304 ± 0.0109	0.1593 ± 0.1121	68.41%
	Ours	$\mathbf{0.0243 \pm 0.0109}$	$\mathbf{0.1192 \pm 0.0882}$	**87.71%**
2	HRegNet*	0.0304 ± 0.0109	0.599 ± 0.1147	68.45%
	Ours	$\mathbf{0.0243 \pm 0.0109}$	$\mathbf{0.1193 \pm 0.0890}$	**87.72%**

Performance Under More Challenging Thresholds. We also compare with baseline HRegNet* [23] in terms of different settings θ_T and θ_R in Table 13 and Table 14. From these tables, much obviously higher recalls, lower RRE, and lower RTE can be observed compared to HRegNet*. This demonstrates the great potential of our method in more challenging outdoor scenes with larger rotations and translations.

Ablation Studies for Various Temporal Interaction Methods. For effective memory read-out, we explore different temporal interaction methods, including temporal averaging, GRU-based, Attention-based, and Mamba-based approaches. As shown in Table 15, the Mamba-based method outperforms the others, particularly in reducing rotation errors. This improvement is due to the effectiveness of the Mamba model in capturing long-term temporal dependencies.

Ablation Studies for Different Mamba Layers. We also vary the number of Mamba layers from 1 to 3 in Table 16. Performance remains stable across all settings, demonstrating our model is insensitive to the specific design choice in Mamba. Notably, we choose 1 as the layer number of the Mamba module because increasing the number cannot bring a significant performance increase.

Ablation Studies for Learnable Parameters in DHW. Learnable weights are of great significance to the dynamic fusion of history and current pose-related features in the Dynamic History Weighting (DHW). To prove this, we replace the learnable weighting matrices W_1, W_2 with fixed ones. As shown in Table 17, using fixed W_1 or W_2 increases both RTE and RRE compared to the learnable DHW, showing that dynamic weighting truly works.

E Visualization Results

Outliers: Dynamic and Occlusion Cases. Dynamics and occlusions are two main challenging outlier cases for outdoor point cloud registration [22]. Here, in Fig. 10, we first show one case of

Table 15: **Ablation studies of temporal encoding methods on KITTI.**

Model	RTE (m)	RRE (deg)	Recall	Time (ms)
Temporal Averaging	0.045 ± 0.041	0.143 ± 0.117	**100%**	131
GRU	0.034 ± 0.039	0.133 ± 0.115	**100%**	135
Attention	0.033 ± 0.039	0.138 ± 0.121	**100%**	136
Mamba (Ours)	$\mathbf{0.032 \pm 0.036}$	$\mathbf{0.122 \pm 0.096}$	**100%**	**133**

Table 16: **Ablation studies for different Mamba layers on KITTI.**

Mamba layers	RTE (m)	RRE (deg)	Recall
1	$\mathbf{0.032 \pm 0.036}$	0.122 ± 0.096	**100%**
2	$\mathbf{0.032} \pm 0.037$	$\mathbf{0.120} \pm 0.103$	**100%**
3	$\mathbf{0.032} \pm 0.037$	$0.121 \pm \mathbf{0.092}$	**100%**

Table 17: **Ablation studies of the learnable parameters (W_1, W_2) in Dynamic History Weighting.**

Method	RTE (m)	RRE (deg)	Recall
Ours w/ fixed W_1	0.036 ± 0.039	0.131 ± 0.103	**100%**
Ours w/ fixed W_2	0.037 ± 0.041	0.134 ± 0.107	**100%**
Ours	$\mathbf{0.032 \pm 0.036}$	$\mathbf{0.122 \pm 0.096}$	**100%**

dynamics. As illustrated in the figure, when a dynamic car runs toward the ego-car at a high speed, parts of the dynamic car will be unobserved. For example, given only Frame 2 and Frame 3, the registration is difficult as the head of the car is not scanned in Frame 3. However, introducing more temporal inputs (Frame 1) will provide more geometric cues of the dynamic car and help recognize this outlier. Another common outlier is occlusion. For example, the occluded car in Frame 2 of Fig. 11 is often seen as an outlier with pairwise inputs (e.g., Frames 1&2), reducing registration accuracy. Sequential observations help identify the occlusion, as the car reappears in Frame 3 and can be correctly matched with Frame 1.

Registration Errors. We display the registration errors of two scenes in Fig. 12. Compared to HRegNet* [23], our method generates much smaller errors thanks to the sufficient exploration of the temporal information.

Dynamic Weighting Visualization. We visualize the weighting matrix using PCA and its corresponding 2D trajectories in Fig. 13. From the figure, it is clear that similar historical trajectories often share similar weights, e.g., (a, b, c), (d, e), (f, g, h), and (i, j). This demonstrates the effectiveness of our Dynamic History Weighting method in learning weights from historical poses to influence the current motion features.

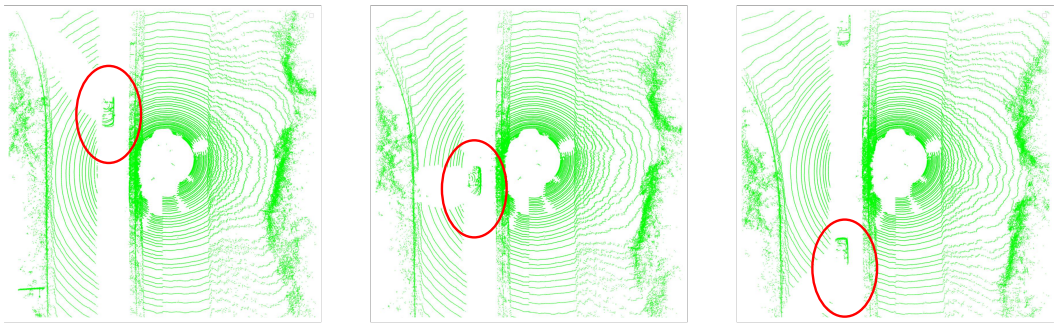
Registration Samples. We display more visualization about the point cloud registration performance on three large-scale outdoor datasets: KITTI in Fig. 14, nuScenes in Fig. 15, and Apollo-Southbay in Fig. 16. From these figures, it is clear that our proposed method accurately registers input point pairs even though there are large displacements (low-overlap problem).

F Limitations

This paper introduces a temporally enhanced outdoor registration methods, which store memory buffers to establish long-term temporal correlations. Although this method works well in both outdoor and multiview indoor methods, its wider generalization ability on non-temporal data needs further validation.

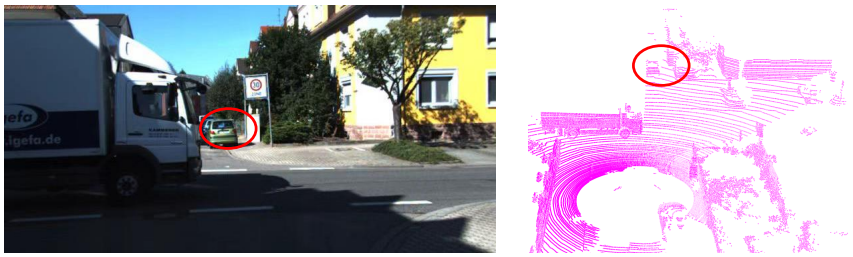
G Broader Impacts

This paper proposes a novel streaming registration method for outdoor autonomous driving scenarios. Its application scenes can be extended to outdoor and indoor robotics for embodied AI tasks. Also, the real-world deployment of our method with further efficiency improvement is highly promising.



Frame 1 **Frame 2** **Frame 3**

Figure 10: **A dynamic case in KITTI.** In a highway scene, partial observations hinder the recognition of the highly dynamic car.



(a) Frame 1



(b) Frame 2



(c) Frame 3

Figure 11: **An occlusion case in KITTI.** A pairwise pose regression can fail in occluded scenarios, e.g., the static car in the red circle in Frame 2, breaking the correlation in Frames 1&2 and 2&3.

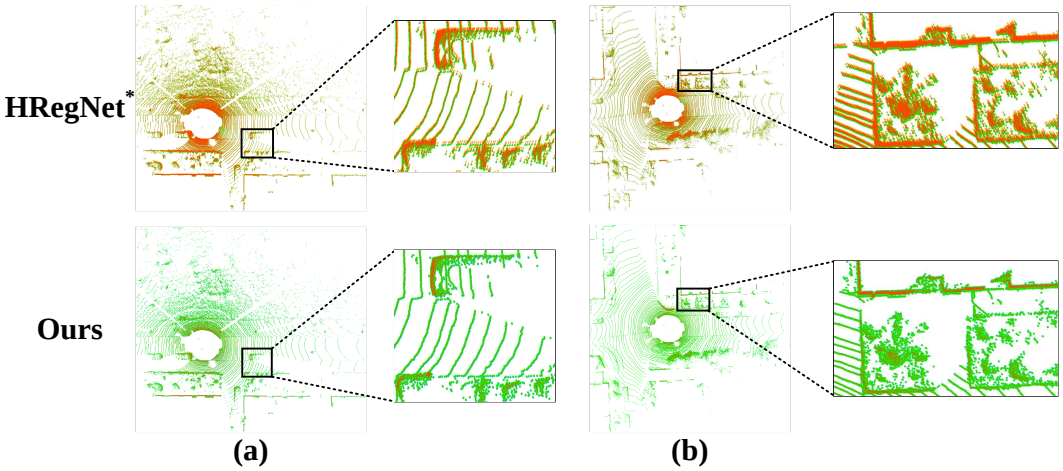


Figure 12: **Registration error visualization.** Our estimation (blue points) shows much smaller errors (red line) compared to HRegNet* [23] (yellow points) relative to GT (green points).

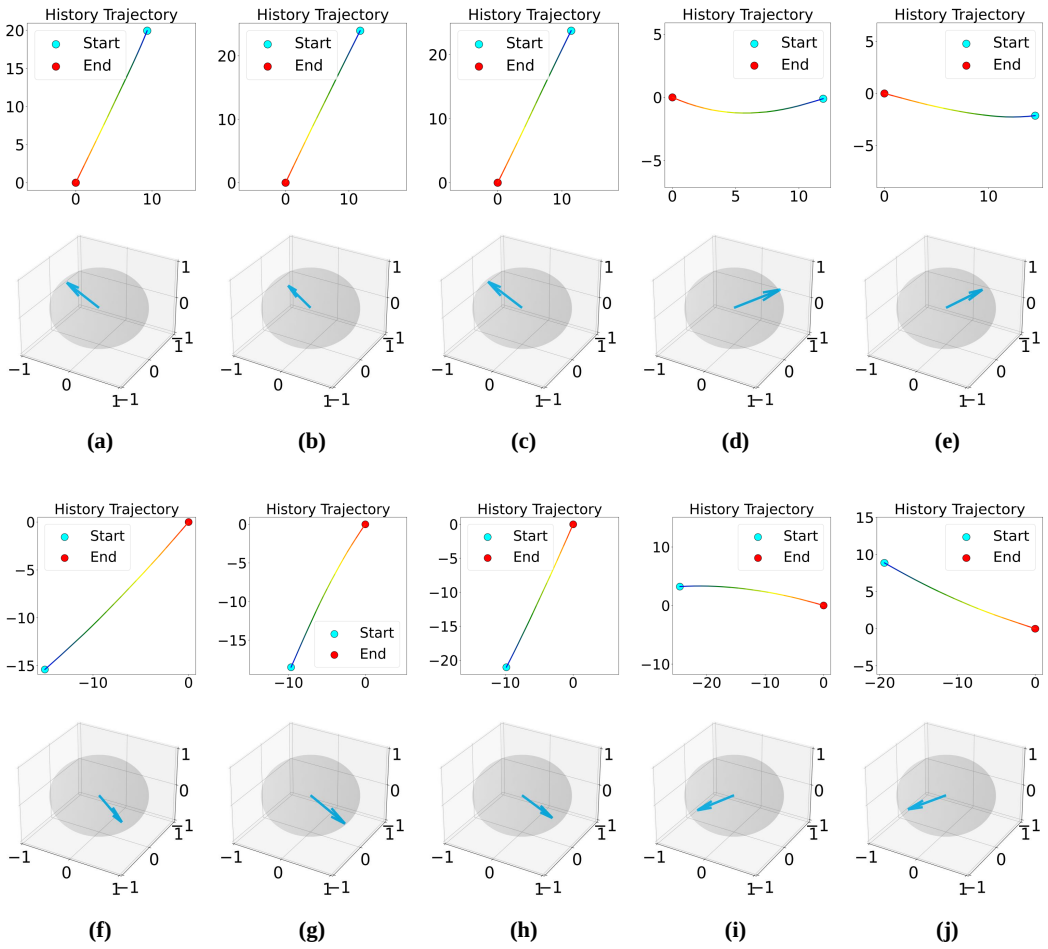


Figure 13: **The visualization of Dynamic History Weighting.** We show some samples about 2D trajectories (upper figures) and their corresponding weights (lower figures), which are generated from the PCA analysis of the weight matrix W_2.

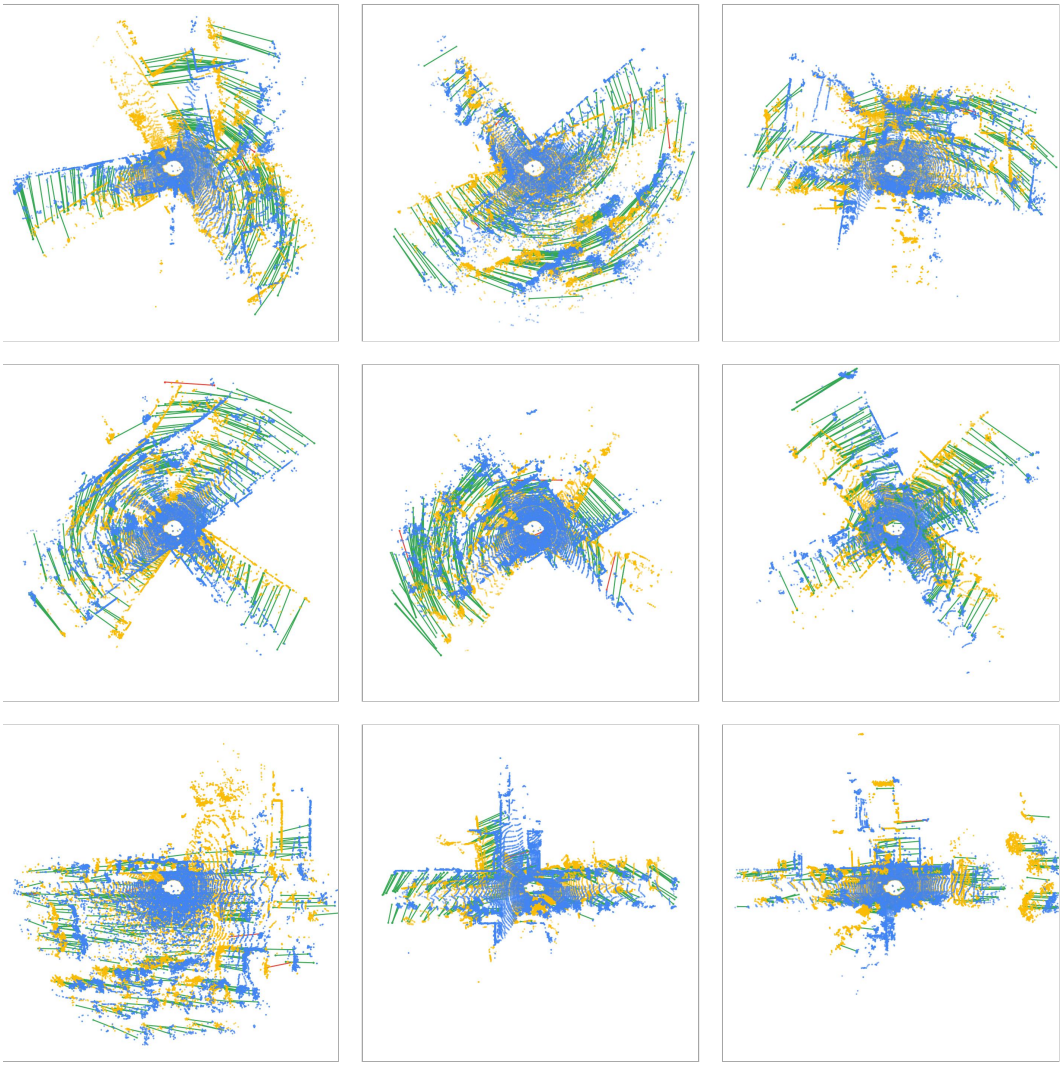


Figure 14: **The visualization of point cloud registration on KITTI.** Yellow and blue points indicate input point pair frames, with green lines showing accurate correspondences (inliers) and red lines showing inaccurate ones (outliers). This figure is best viewed in zoomed-in color mode.

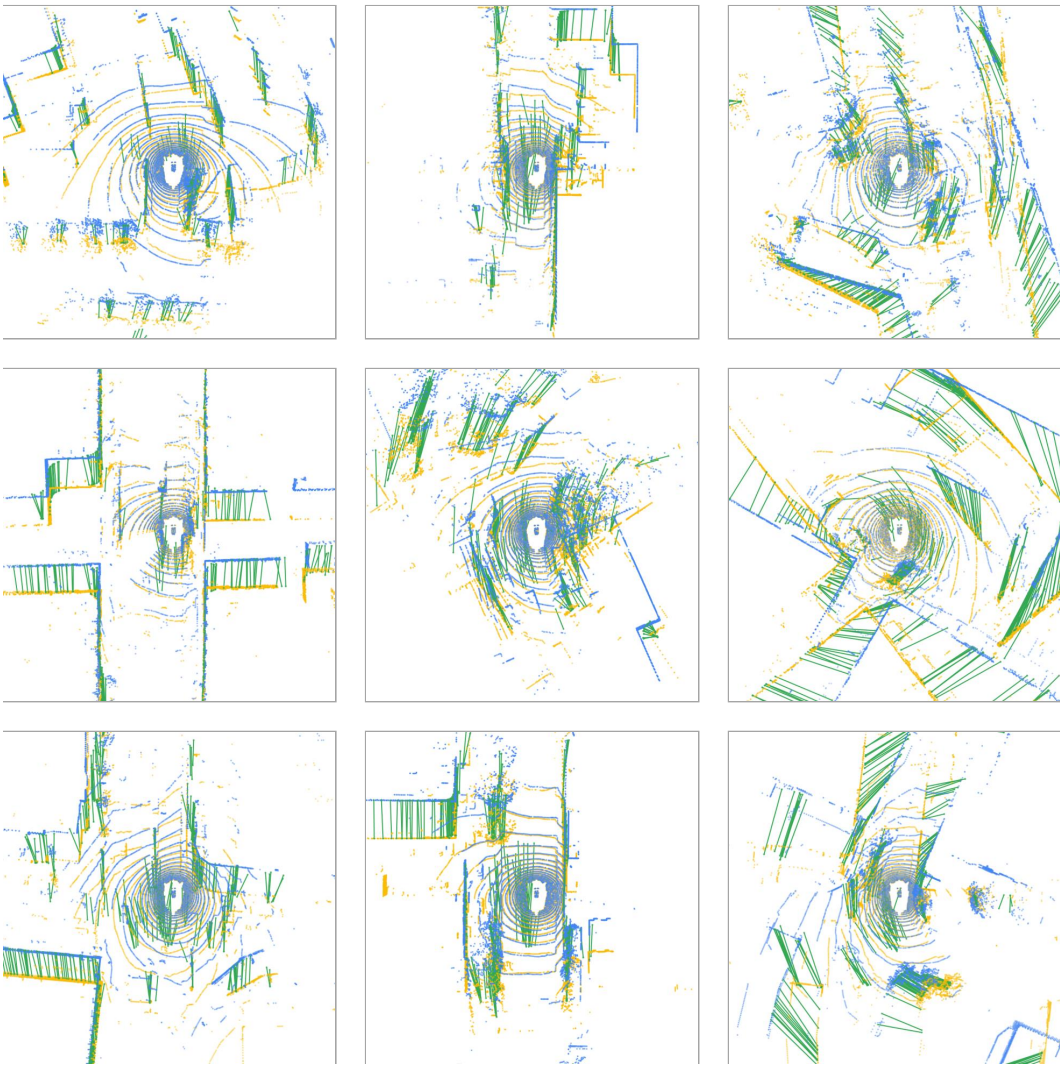


Figure 15: **The visualization of point cloud registration on nuScenes.** Yellow and blue points indicate input point pair frames, with green lines showing accurate correspondences (inliers) and red lines showing inaccurate ones (outliers). This figure is best viewed in zoomed-in color mode.

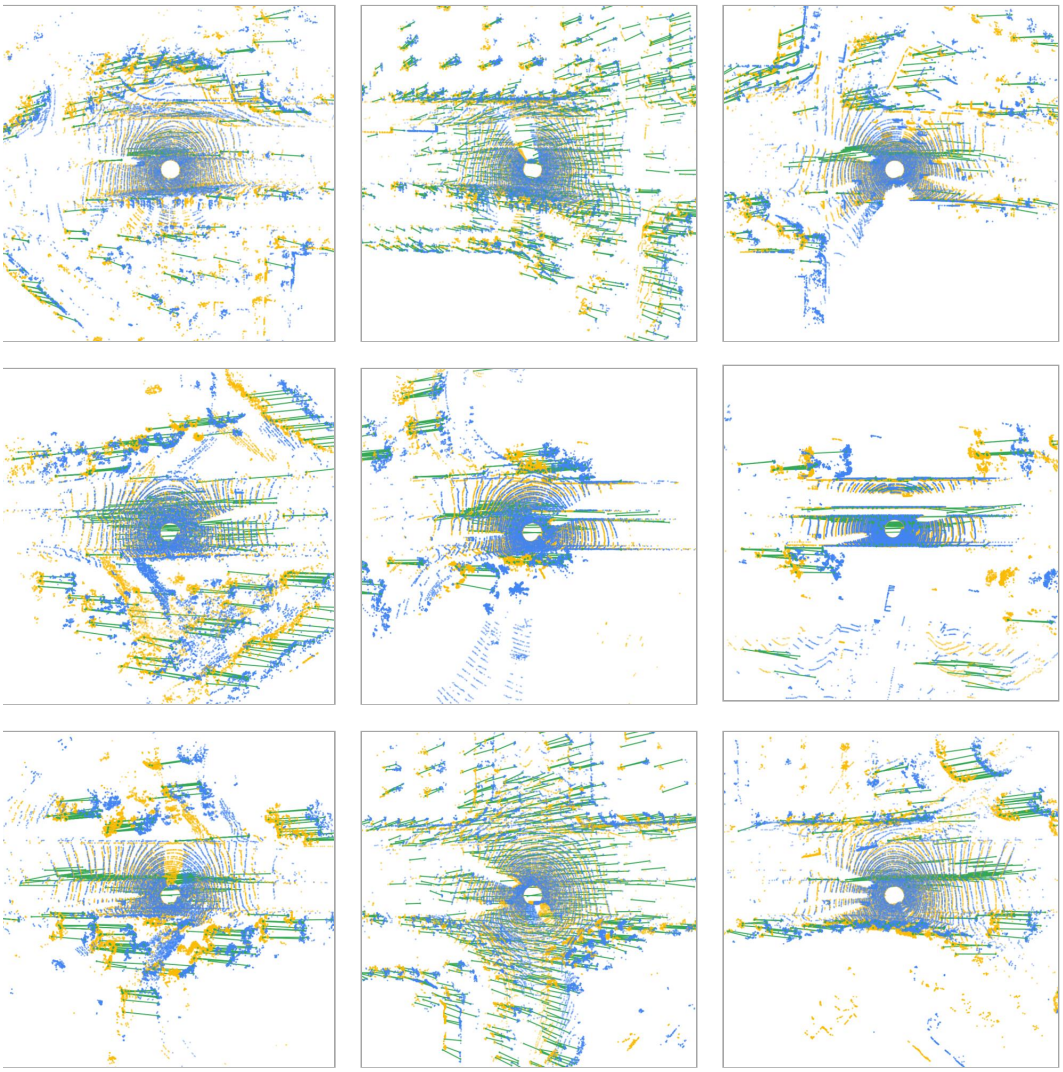


Figure 16: **The visualization of point cloud registration on Apollo-Southbay.** Yellow and blue points indicate input point pair frames, with green lines showing accurate correspondences (inliers) and red lines showing inaccurate ones (outliers). This figure is best viewed in zoomed-in color mode.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The main claims in the abstract and introduction faithfully reflect the paper's contributions and scope and match theoretical and experimental results.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We have included a discussion about the paper's limitations in Section F of the Appendix.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [N/A]

   Justification: The paper does not present new theoretical results, theorems, or formal guarantees. The equations in the method section define the model architectures and loss functions.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: The paper fully disclose all the information needed to reproduce the main experimental results. We also supplement the implementation code in the supplementary materials.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: The experiments use publicly available datasets and benchmarks, including KITTI, nuScenes, Apollo, 3D(Lo)Match, and ScanNet, following their official access procedures and licenses. The code will be open-sourced after publication.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: The paper specifies the datasets, validation splits, input modality, evaluation metrics, optimizer, learning rate, weight decay, batch size, training steps, feature resolution, and architecture details. Additional architecture and verification details are provided in the supplementary material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

   Justification: The paper hasn't included the error bars due to highly fixed results under our default settings.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: The paper has provided sufficient information on the computer resources such as latency on all datasets in the main paper.

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics.

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: We have included a broader impact clarification in Section G of the Appendix, including wider application scenarios like embodied AI tasks, and also real-world deployment with high efficiency.

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The paper poses no such risks.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: The creators or original owners of assets (e.g., code, data, models), used in the paper, are properly credited and the license and terms of use are explicitly mentioned and properly respected.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: The paper has communicated the details of the dataset/code/model as part of submissions via structured templates.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [N/A]

Justification: The paper does not involve crowdsourcing nor research with human subjects.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [N/A]

Justification: The core method development in this research does not involve LLMs as any important, original, or non-standard components.