TRAINING-FREE GLOBAL GEOMETRIC ASSOCIATION FOR 4D LIDAR PANOPTIC SEGMENTATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Dominant paradigms for 4D LiDAR panoptic segmentation are usually required to train deep neural networks with large superimposed point clouds or design dedicated modules for instance association. However, these approaches perform redundant point processing and consequently become computationally expensive, yet still overlook the rich geometric priors inherently provided by raw point clouds. To this end, we introduce GEO-4D, a simple yet effective training-free framework that unifies spatial and temporal reasoning, enabling holistic LiDAR perception over long time horizons. Specifically, we propose a global geometric association strategy that establishes consistent instance correspondences by estimating an optimal transformation between instance-level point sets. To mitigate instability caused by structural inconsistencies in point cloud observations, we propose a global geometry-aware soft matching mechanism that enforces spatially coherent point-wise correspondences grounded in the spatial distribution of instance point sets. Furthermore, our carefully designed pipeline, which considers three instance types—static, dynamic, and missing—offers computational efficiency and occlusion-aware matching. Our extensive experiments across both SemanticKITTI and nuScenes demonstrate that our method consistently outperforms state-of-the-art approaches, even without additional training or extra point cloud inputs. These results underscore the potential of inherent geometric priors in point clouds as an efficient solution for maintaining consistent instance identities over time.

1 INTRODUCTION

Autonomous driving systems heavily depend on precise perception of dynamic 3D environments with geometrically accurate measurements captured by LiDAR sensors. In particular, LiDAR segmentation plays a crucial role in enabling comprehensive point-wise understanding. In 3D environments, LiDAR panoptic segmentation (Kirillov et al., 2019; Milioto et al., 2020) originates from semantic and instance segmentation, which jointly identify class categories and instance IDs. Along with the recent significant advances in frame-wise perception, 4D LiDAR panoptic segmentation (Aygun et al., 2021) has emerged to capture temporal relationships in a holistic manner. In this context, ensuring temporal continuity is a principal purpose in 4D LiDAR panoptic segmentation, even under instance motions.

With its potential benefits, as shown in Table 1, the mainstream of prior works can be categorized into three discrete paradigms: 1) IoU-based, 2) Query-propagated, and 3) Detect & Track association. One solution (Aygun et al., 2021; Kreuzberg et al., 2022; Zhu et al., 2023; Yilmaz et al., 2024; Hong et al., 2024), superimposing point clouds, is widely adopted for establishing 4D spatio-temporal volume. It enables consistent instance tracking over time by measuring overlapping regions. However, despite its simplicity, stacking point clouds (i.e., increasing the number of points per batch) leads to a quadratic growth in computational overhead, which is an inherent challenge in point cloud processing (Liu et al., 2023; Lai et al., 2022; Wu et al., 2022; 2024).

Table 1: Conceptual comparison of 4D LiDAR panoptic segmentation paradigms.

Strategy	Frame-wise Perception	No Temporal Training	Off-the-shelf 3D Perception Model
IoU-based	✗	✗	✗
Query Prop.	✓	✗	✗
Detect & Track	✓	✗	✓
Geo-4D (Ours)	✓	✓	✓

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

An alternative direction (Marcuzzi et al., 2023; Athar et al., 2023; Marcuzzi et al., 2022) seeks to mitigate this issue by treating each scan independently, as demonstrated in the 2nd and 3rd rows of Table 1. They are primarily derived from the two major paradigms of multi-object tracking, tracking-by-detection (Bewley et al., 2016; Wojke et al., 2017; Zhang et al., 2022b; Qin et al., 2024) and query propagation (Zeng et al., 2022; Zhang et al., 2023b; Meinhardt et al., 2022), to assign consistent IDs for instances. Both approaches differ substantially in training strategies, yet commonly suffer from error accumulation (Zhang et al., 2023b; Segu et al., 2024) and reliance on costly and dataset-dependent training. Consequently, we argue that this suboptimal design fails to fully exploit the strong perceptual capabilities of 3D panoptic networks, which naturally raises the following question: *Can instance association be achieved efficiently using only a 3D panoptic network?*

To answer this, we propose a global geometric association strategy for consistent and reliable instance correspondence, even without any training process. Iterative Closest Point (ICP) is widely used for point cloud registration (Wang & Solomon, 2019; Zhang et al., 2023a; Choy et al., 2020), but its applicability beyond that has been largely underexplored. Here, we integrate ICP into the 4D LiDAR panoptic segmentation framework, thereby enabling explicit instance correspondence through geometric alignment under estimated inter-scan transformations. Additionally, we introduce a novel paradigm for point-wise correspondence, Global Geometry-aware Soft Matching (GGSM), enhancing the robustness to structural noise by enforcing instance-aware geometric correspondences. Traditional ICP performs well on point cloud pairs under ideal conditions, but it relies on nearest-neighbor point correspondences. In real-world noisy measurements acquired from on-vehicle sensors, such point-wise matching often yields incorrect transformations and ultimately leads to erroneous instance associations, as it ignores the global instance structure. In contrast, our GGSM allows reliable correspondences based on the overall geometric context rather than local proximity. Specifically, it identifies point pairs that minimize the transport cost by solving the earth mover's problem (Cuturi, 2013) between two instance point set distributions. As a result, instance association shifts from greedy local pairing to globally consistent alignment, yielding more stable and accurate matching even under challenging environments.

In addition, in real-world observations, instances may remain stationary, move across frames, or temporarily disappear due to occlusion or sensing limitations. Motivated by these instance states, we classify instances into three types: static, dynamic, and missing. For static instances, we apply statistics-based geometric filtering using the mean and covariance of their point distributions to preserve spatial consistency across frames. This filtering step also significantly reduces the number of matching candidates before applying global geometric association to dynamic instances, thereby ensuring computational efficiency. Moreover, we incorporate a short-term memory bank to recover temporarily unobservable instances. We retain only unmatched instances within a short-term window, preserving temporal consistency with minimal memory overhead.

To wrap it up, we named our overall pipeline as **GEO-4D** and demonstrate its effectiveness through extensive experiments on the SemanticKITTI (Behley et al., 2019) and nuScenes (Fong et al., 2022) benchmarks. Note that GEO-4D achieves superior performance over existing state-of-the-art methods and ranks 1st on the SemanticKITTI test leaderboard, even under the minimal setting of a single scan and without any additional training. The contributions of this work are summarized as follows: (1) we introduce GEO-4D, a training-free 4D LiDAR panoptic segmentation pipeline that can be flexibly integrated with various 3D panoptic segmentation networks; (2) we propose a global geometric association strategy with instance state conditioned solutions to improve association robustness while efficiently handling computational complexity and occlusion; and (3) we demonstrate the superiority of GEO-4D through extensive experiments and in-depth analyses on challenging benchmarks.

2 RELATED WORK

LiDAR Panoptic Segmentation. LiDAR panoptic segmentation handles two different basis tasks, semantic and instance segmentation. Proposal-based methods Liu et al. (2019); Mohan & Valada (2021); Xiong et al. (2019) first identify *thing* points and then group them into instances, whereas proposal-free methods (Hong et al., 2021; Razani et al., 2021; Zhou et al., 2021; Li et al., 2022) directly predict semantic labels and class agnostic instance segments. 4D LiDAR panoptic segmentation further requires instance identities to remain consistent across sequential scans. Most existing methods (Zhang et al., 2025; Aygun et al., 2021; Kreuzberg et al., 2022; Zhu et al., 2023; Yilmaz

et al., 2024; Hong et al., 2024) aggregate multiple consecutive scans and associate instances based on overlapping regions. Alternatively, CA-Net (Marcuzzi et al., 2022) uses feature similarity, while Mask4D (Marcuzzi et al., 2023) maintains identities by iteratively updating track queries. These approaches often incur additional computational cost or depend on specific network designs. In contrast, we propose GEO-4D that is a flexible, efficient, and well-balanced instance association through direct geometric registration.

Multi-Object Tracking (MOT). Tracking-by-detection (TBD) (Bewley et al., 2016; Wojke et al., 2017; Zhang et al., 2022b; Qin et al., 2024) has become the *de facto* standard due to its simplicity and the rapid progress of detectors. Here, the motion model (e.g., Kalman Filter (Kalman, 1960)) plays a key role by estimating future object states to maintain tracking continuity. Recent DETR-based methods (Zeng et al., 2022; Zhang et al., 2023b; Meinhardt et al., 2022) instead perform detection and tracking jointly with query-based representations, and similar paradigms have been extended to 3D tracking (Weng et al., 2020; Zhou et al., 2020; Pang et al., 2022; Zhang et al., 2022a). Unlike conventional MOT, however, 4D LiDAR panoptic segmentation requires point-level scene understanding rather than bounding box tracking. GEO-4D follows the TBD paradigm while grounding instance association in the geometric structure of the underlying point clouds.

Iterative Closest Point (ICP). ICP (Besl & McKay, 1992) is a classical point set registration method with numerous extensions (Bustos & Chin, 2017; Li, 2021; Rusu et al., 2009; Bouaziz et al., 2013; Rusinkiewicz & Levoy, 2001; Segal et al., 2009; Yang et al., 2015). It has also been incorporated into deep learning–based applications (Wang & Solomon, 2019; Zhang et al., 2023a; Choy et al., 2020). ICP-Flow (Lin & Caesar, 2024), which estimates point-wise scene flow between adjacent scans using instance-level clustering, is closely related to our work. However, they only consider point-wise motion estimation between two temporally adjacent scans by leveraging well-separated instance clusters. In contrast, we use ICP for persistent instance association across sequential scans. Remarkably, our Global Geometry-aware Soft Matching enables instance centric association, even when instance point sets are noisy, imprecise, or fragmented.

3 METHOD

3.1 PRELIMINARY

Iterative Closest Point (ICP). Conventionally, for point cloud registration, ICP estimates the optimal transformation by iteratively minimizing the difference between source and target point sets. Given two point sets $\mathcal{P}_{\text{src}} = \{p_1, ..., p_I\}$ and $\mathcal{P}_{\text{dst}} = \{q_1, ..., q_J\}$ in $\mathbb{R}^3$, it starts with the initial guess of rigid transformation $(\mathbf{R}^{(0)}, \mathbf{t}^{(0)})$, where the rotation $\mathbf{R} \in \text{SO}(3)$ and translation $\mathbf{t} \in \mathbb{R}^3$, which satisfies the following formula:

$$q \approx \mathbf{R}p + \mathbf{t} \ . \tag{1}$$

Here, in order to compute the alignment error, pairwise correspondences are established using a nearest-neighbor search algorithm (e.g., KD-tree (Bentley, 1975)) as:

$$\mathcal{C}^{(k)} = \{(p, \hat{q}) \mid \hat{q} = \underset{q \in \mathcal{P}_{\text{dst}}}{\arg\min} ||\mathbf{R}^{(k)}p + \mathbf{t}^{(k)} - q||, \ \forall p \in \mathcal{P}_{\text{src}}\}, \tag{2}$$

where k denotes the iteration. Classical ICP adopts l_2 distance as the error metric between the matched point pairs. Ultimately, the optimal rigid transformation is obtained by iteratively solving a series of least-squares problems as:

$$\mathbf{R}^{(k+1)}, \mathbf{t}^{(k+1)} = \underset{\mathbf{R}, \mathbf{t}}{\arg\min} \sum_{(p, \hat{q}) \in \mathcal{C}^{(k)}} ||\mathbf{R}p + \mathbf{t} - \hat{q}||^2 \ , \tag{3}$$

to update the rotation $\mathbf{R}$ and translation $\mathbf{t}$ until convergence. With this theoretical background, we investigate how ICP contributes to consistent instance ID assignment across frames while enhancing robustness against outliers.

3.2 GEO-4D

Our proposed GEO-4D is developed on top of a pre-trained 3D LiDAR panoptic segmentation model (Hong et al., 2021; Yilmaz et al., 2024), which is specialized in frame-wise perception.

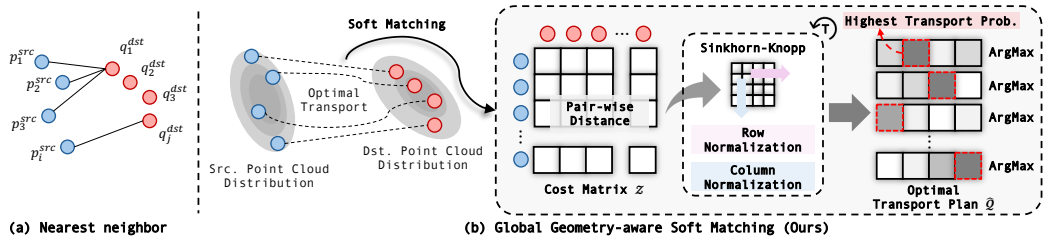

Figure 1: (a) Nearest neighbor aligns each source point with its closest counterpart, focusing only on local point proximity. (b) In contrast, our Global-Geometry-aware Soft Matching computes correspondences via an optimal transport plan $\hat{\mathcal{Q}}$, enabling instance-aware matching that respects the global geometry of each instance point set.

Following the standard assumption in LiDAR panoptic segmentation, we transform all points into the world coordinate system using the ego pose. Through this model, we obtain the per-point semantic and instance prediction $(s_i,\ \ell_i)$ for a given point cloud, where $s_i \in \{1, ..., S\}$, $\ell_i \in \mathbb{N}_0$, and S is the number of classes. Prior to the association stage, we construct a set of matching candidates between t and $t-1$ times by pairing source and destination instances as follows:

$$\mathcal{M} = \mathcal{G}_{\text{src}} \times \mathcal{G}_{\text{dst}} = \{(a, b) \mid a \in \mathcal{G}_{\text{src}},\ b \in \mathcal{G}_{\text{dst}}\}, \tag{4}$$

where $\mathcal{G}_{\text{src}}$ and $\mathcal{G}_{\text{dst}}$ are respectively denoted as the sets of unique instance IDs in the source and destination. Then, we derive the filtered set $\mathcal{M}'$ by keeping only the pairs in $\mathcal{M}$ that share the same semantic class. In this study, we categorize instances into dynamic, static, and missing types to facilitate effective instance association across frames.

Static Instance. Under the assumption that both the center and the shape of a non-moving rigid object remain invariant, we filter static instances based on their centers and further refine them using covariance cues to improve selection fidelity. Given the matching candidates $\mathcal{M}'$, we first represent each instance by the set of 3D points belonging to it. For each instance pair $(a'_{\text{src}},\ b'_{\text{dst}}) \in \mathcal{M}'$, we denote the corresponding point sets as $\mathcal{P}(a'_{\text{src}})$ and $\mathcal{P}(b'_{\text{dst}})$, respectively. Then, we compute the centroid of each point set:

$$\mu_s = \frac{1}{|\mathcal{P}(a'_{\text{src}})|} \sum_{p \in \mathcal{P}(a'_{\text{src}})} p, \qquad \mu_d = \frac{1}{|\mathcal{P}(b'_{\text{dst}})|} \sum_{q \in \mathcal{P}(b'_{\text{dst}})} q, \tag{5}$$

where $p, q \in \mathbb{R}^3$ represent the world coordinates of points. We keep only the pairs satisfying $||\mu_d - \mu_s||_2 < \tau_{center}$, where τ_{center} is the distance threshold for identifying non-moving objects based on center consistency. Although this naive approach performs well under ideal conditions, it suffers when applied to noisy predicted point sets, resulting in unsuitable pair associations. Thus, we introduce an additional covariance-based constraint that measures the similarity of the spatial spread and geometric structure of the instance point sets. Specifically, the covariance of each point set is computed as follows:

$$\sigma_s = \text{Cov}(\mathcal{P}(a'_{\text{src}})), \quad \sigma_d = \text{Cov}(\mathcal{P}(b'_{\text{dst}})). \tag{6}$$

To assess shape consistency, we measure the similarity between the two covariance matrices and use the predefined threshold τ_{cov} to control the discrepancy between source and destination point set:

$$\frac{||\sigma_d - \sigma_s||_F}{\text{tr}(\sigma_d) + \text{tr}(\sigma_s)} < \tau_{cov}, \tag{7}$$

where $\text{tr}(\cdot)$ and $\|\cdot\|_F$ denote the matrix trace and the Frobenius norm. Then, for all candidate pairs satisfying the filtering conditions, we choose the smallest discrepancy in Eq. 7 for each source instance to determine the best-matching pairs while avoiding duplicate associations.

Dynamic Instance. After filtering static instances, we construct the remaining candidate pair set $\mathcal{M}''$. These candidate pairs are then fed into the global geometric association, which serves as the core component of our framework for training-free instance association. In the vanilla ICP loop, the first step is finding point pairs to compute the distance by utilizing the nearest-neighbor algorithm. While ICP has a powerful ability to register point cloud sets, determining the corresponding points

under noisy instance predictions potentially causes erroneous matches. Therefore, we propose a Global Geometry-aware Soft Matching that facilitates stable correspondence estimation within ICP.

Global Geometry-aware Soft Matching starts by regarding the point cloud as a probability distribution rather than a discrete set of individual points. For a candidate pair $(a''_{\text{src}},\ b''_{\text{dst}}) \in \mathcal{M}''$, we denote their corresponding point sets as $\mathcal{P}(a''_{\text{src}}) = \{p_i\}_0^{I-1}$ and $\mathcal{P}(b''_{\text{dst}}) = \{q_j\}_0^{J-1}$, where I and J denote the number of points in each instance. We then formulate the correspondence estimation as an Optimal Transport (OT) problem (Cuturi, 2013). Specifically, we seek a transport plan $\mathcal{Q}$ that minimizes the cost of transforming one distribution into the other as follows:

$$\text{OT}(\mathcal{P}(a''_{\text{src}}),\ \mathcal{P}(b''_{\text{dst}})) := \min_{\mathcal{Q}} \langle \mathcal{Z},\ \mathcal{Q} \rangle_F - \epsilon H(\mathcal{Q}), \tag{8}$$

where $\langle \cdot \rangle_F$ denotes Frobenius dot-product. Here, the cost matrix $\mathcal{Z}$ represents the pairwise transport cost, defined as $\mathcal{Z}_{ij} = ||p_i - q_j||_2^2$. The entropy H and ϵ control the smoothness of the transport plan as a regularization term. Then, we utilize the Sinkhorn-Knopp algorithm (Sinkhorn, 1967) that provides an efficient solution to the entropy-regularized optimal transport problem by iteratively rescaling the rows and columns of $\mathcal{Q}$ until the marginal constraints are satisfied (details in supplementary material). Afterwards, we derive the final correspondences from the source to the destination point by taking the argmax operation over each row of the optimal transport plan $\hat{\mathcal{Q}}$ as:

$$j^*(i) = \arg\max_j \hat{\mathcal{Q}}_{ij}, \quad \mathcal{C}^{(k)}_{\text{GGSM}} = \{(p_i, q_{j^*(i)}) \mid i = 0, \ldots, I-1\}. \tag{9}$$

That is, we approximate point-to-point correspondences by selecting the highest transport probability from the minimized transport plan. As a result, this strategy is based on considering the instance instead of local points (see Figure 1), thereby alleviating the vulnerability to outliers and incorrect predictions.

Returning to the ICP procedure, we follow the standard process, initializing the transformation with a histogram-based method (Lin & Caesar, 2024) to reduce sensitivity to initial estimates and subsequently running iterative refinement to get the final rigid transformation. After obtaining the transformation $(\mathbf{R}, \mathbf{t})$, we determine whether two instances correspond by evaluating the IoU between their transformed point sets. Specifically, points with errors smaller than the predefined threshold τ_{dist} are treated as inliers. Then, in order to filter valid matching candidates, we evaluate alignment accuracy between source and destination instances using IoU scores as follows:

$$\delta(a''_{\text{src}}, b''_{\text{dst}}) = \mathbb{1}[\text{IoU}(\mathcal{F}(\mathcal{P}(a''_{\text{src}})), \mathcal{P}(b''_{\text{dst}})) \geq \tau_{\text{iou}}], \tag{10}$$

where $\text{IoU}(\cdot)$ denotes the IoU calculation, and $\mathcal{F}$ applies the estimated transformation obtained by ICP. Using Eq. 10, we assign the same instance ID to matched candidates. In our experiments, this simple matching strategy consistently outperforms more complex assignment schemes. Additional comparisons are provided in the supplementary material.

Missing Instance. Occlusions stemming from the inherent limitations of LiDAR sensor often break temporal continuity beyond two consecutive scans. To mitigate this, we introduce a memory bank $\mathcal{B}$ that caches the world coordinates, semantic labels, and instance IDs of unmatched objects in both static and dynamic association stages. Owing to bounded memory capacity, $\mathcal{B}$ keeps only entries observed within a short window of the previous w_{mem} scans. If $\mathcal{B}$ is not empty, we retrieve the instance information and apply the global geometric association strategy to restore temporal consistency.

4 Experiments

In this section, we present experimental results to answer the following research questions: **(Q1)** How well does GEO-4D perform compared with different types of association methods? **(Q2)** Why are the geometric properties of point clouds essential for reliable instance association? **(Q3)** Does GEO-4D generalize consistently across different segmentation outputs and datasets? **(Q4)** How does global geometric consistency help GGSM avoid erroneous local correspondences? **(Q5)** Does GEO-4D maintain robust performance regardless of object type and point density?

4.1 Experiment Setup

Datasets. We choose two representative benchmarks. SemanticKITTI (Behley et al., 2019) captures point clouds using a 64-beam LiDAR at 10 Hz and provides point-wise annotations for 19 semantic

Table 2: **Quantitative results on SemanticKITTI.** We compare GEO-4D with prior methods in terms of the standard 4D panoptic metric (LSTQ), complementary tracking metrics, and computational efficiency. Memory usage is reported in GB, and inference time is reported in seconds per scan. Best and second best results are shown in **bold** and underlined, respectively.

		Validation						Test		Efficiency	
Method	#Sc.	LSTQ↑	S_{assoc} ↑	MOTSA↑	sMOTSA↑	PTQ↑	sPTQ↑	LSTQ↑	S_{assoc} ↑	Mem.↓	Inf. Time↓
4D-StOP (Kreuzberg et al., 2022)	2	66.4	71.8	31.0	27.1	53.4	53.7	62.9	67.3	15.5	3.67
Eq-4D-StOP (Zhu et al., 2023)	2	68.9	74.8	43.6	39.6	56.4	56.7	65.4	69.7	11.0	2.89
Mask4Former (Yilmaz et al., 2024)	2	72.0	76.3	47.3	42.1	48.3	48.5	68.4	67.3	10.4	3.51
Mask4D (Marcuzzi et al., 2023)	1	71.4	75.4	52.1	48.1	58.4	58.5	64.3	66.4	11.9	1.12
Mask4Former + CA-Net (Marcuzzi et al., 2022)	1	70.4	73.4	48.4	43.1	48.9	49.1	66.5	62.8	5.71	1.56
Mask4Former + GEO-4D	**1**	**72.7**	**78.3**	**73.3**	**65.7**	**59.6**	**59.7**	**70.3**	**70.0**	**3.15**	**1.10**

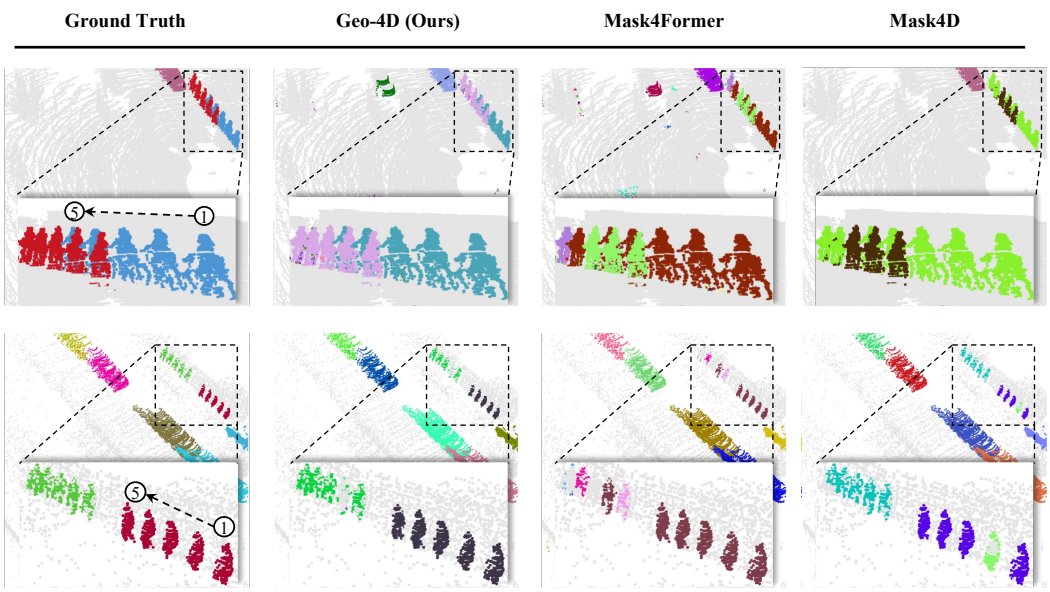

Figure 2: **Qualitative comparison on SemanticKITTI**. We visualize the association results over five consecutive scans for both the baselines and our GEO-4D. Different colors represent different instances. The dotted boxes zoom in for clear comparison. For clarity, we use circled numbers (①–⑤) to represent the frame indices.

classes (8 Things / 11 Stuff). In contrast, panoptic nuScenes (Fong et al., 2022) uses a 32-beam LiDAR at 2 Hz and offers point-wise annotations for 16 semantic classes (10 Things / 6 Stuff).

Evaluation Metrics. To evaluate tracking quality, we adopt the widely used LSTQ (Aygun et al., 2021), defined as LSTQ $= \sqrt{S_{cls} \times S_{assoc}}$. Here, S_{cls} measures how accurately semantic labels are assigned and S_{assoc} evaluates how consistently instance identities are maintained throughout the sequence, independently of semantic accuracy. A detailed formulation, along with descriptions of the additional evaluation metrics and implementation details, is provided in Appendix A.

4.2 4D LIDAR PANOPTIC SEGMENTATION RESULTS

Addressing Q1. Table 2 reports the experimental results on the SemanticKITTI dataset. Beyond the standard LSTQ and S_{assoc} metrics, we report additional tracking metrics to provide a more comprehensive evaluation of instance association quality. These results reveal two key observations.

First, GEO-4D consistently exhibits strong improvements over competing methods across all tracking metrics, regardless of the number of input scans. Specifically, on the validation set, GEO-4D improves S_{assoc} by 2.0 points over the strongest competitor (78.3 vs. 76.3), while achieving substantially larger gains of 21.2 and 17.6 points in MOTSA and sMOTSA, respectively. Furthermore, GEO-4D maintains strong performance on both the validation and test sets, whereas learning-based methods

Table 3: **Association analysis on SemanticKITTI.** (a) Comparison with classic association baselines. (b) Controlled comparison under identical frame-wise predictions.

(a) Classic association baselines

Method	#Scan	LSTQ↑	S_{assoc} ↑
Centroid + Hungarian	1	55.4	45.4
3D Box IoU + Hungarian	1	63.6	59.9
Kalman + 3D Box IoU + Hungarian	1	64.6	61.9
GEO-4D (Ours)	**1**	**72.7**	**78.3**

All baselines use the same Mask4Former predictions.

(b) Same-mask comparison

Input Masks	Association Strategy	#Scan	LSTQ↑	S_{assoc} ↑
Eq-4D-StOP	IoU-based	2	65.0	71.5
	Geo-4D (Ours)	1	**65.2**	**71.9**
Mask4Former	IoU-based	2	72.0	76.3
	Geo-4D (Ours)	1	**72.8**	**77.8**

Each pair shares identical frame-wise predictions.

(Mask4D and CA-Net) exhibit noticeable performance degradation on the test set. This highlights the robustness of direct geometric association on raw point clouds.

Second, GEO-4D offers the most well-balanced trade-off between performance and computational overhead. Compared with methods using multiple input scans, GEO-4D requires 69.7%–73.8% less memory and reduces inference time by 61.9%–70.0%, while achieving superior performance. Although Mask4D shows competitive runtime, it incurs substantially higher memory overhead due to its transformer-based architecture. CA-Net is also computationally efficient, but yields considerably weaker association performance, with the gap becoming even more pronounced on the hidden test set.

Moreover, we qualitatively evaluate GEO-4D in Figure 2. In the first row, GEO-4D consistently preserves the identities of two closely spaced bicyclists, whereas Mask4Former struggles to preserve identity consistency when objects are spatially adjacent or overlapping due to its IoU-based association. Mask4D also becomes less reliable when objects share similar visual or spatial characteristics, as its association relies on learnable tracking queries. In the second row, we demonstrate the benefit of our memory bank for temporarily missing instances. Although the 3D panoptic model misses the person shown in green at the second time step, our memory bank successfully retrieves the instance and maintains its identity, while Mask4Former experiences ID switching when it fails to detect an instance due to the absence of overlapping regions. Consequently, these observations underscore the importance of explicit geometric matching in our approach and a well-structured association pipeline.

Table 4: Quantitative results on panoptic nuScenes (Fong et al., 2022).

Method	#Sc.	LSTQ↑	S_{assoc} ↑	S_{cls} ↑
4D-StOP (Kreuzberg et al., 2022)	4	60.5	62.5	58.6
Eq-4D-StOP (Zhu et al., 2023)	2	58.9	62.3	55.7
Mask4Former (Yilmaz et al., 2024)	2	67.6	63.0	**72.5**
Mask4D (Marcuzzi et al., 2023)	1	58.5	54.1	63.1
Mask4Former	1	–	–	71.0
+ CA-Net (Marcuzzi et al., 2022)	1	63.4	56.6	71.0
+ **GEO-4D (Ours)**	**1**	**67.9**	**65.0**	71.0

Addressing Q2. Table 3(a) highlights the necessity of directly leveraging raw object geometry rather than high-level geometric descriptors (e.g., centroid and 3D bbox). The centroid computes the mean of object points and uses Euclidean distance as the matching cost, while the 3D bounding box defines the cost based on 3D IoU. We further enhance the baseline with a Kalman filter for temporal state estimation. We find that our GEO-4D outperforms classic baselines, yielding 26.5%–72.5% improvements in S_{assoc}. This finding suggests that exploiting the geometric structure of objects significantly reduces identity ambiguity, resulting in reliable instance association.

Addressing Q3. Table 3(b) evaluates GEO-4D against an IoU-based strategy while holding the segmentation masks fixed. We clearly observe that GEO-4D improves association performance under both mask settings, using predictions from EQ-4D-StOP and Mask4Former. Notably, these gains are achieved even when only single frame segmentation masks are available. Moreover, Table 4 shows the 4D panoptic results on panoptic nuScenes dataset. Consistent with the observations on the SemanticKITTI, GEO-4D delivers notably stronger instance association performance, outperforming existing methods in terms of LSTQ and S_{assoc}. Together, these results imply that GEO-4D generalizes across different segmentation masks and datasets, while remaining highly effective under sparse LiDAR observations.

4.3 FURTHER ANALYSIS OF THE GLOBAL GEOMETRIC ASSOCIATION

Addressing Q4. The Global Geometry-aware Soft Matching (GGSM) forms instance-aware point correspondence between source and target points. Owing to this strategy, accurate error measurements

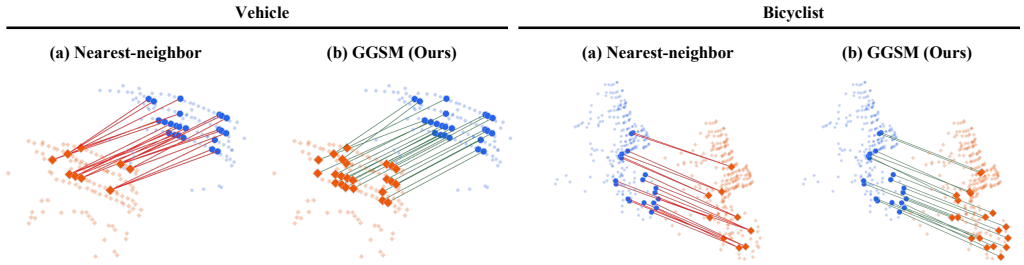

Figure 3: Visualization of point-wise correspondences. (a) Nearest-neighbor matching. (b) Our Global Geometry-aware Soft Matching. ◆ and ● represent the target and source points, respectively.

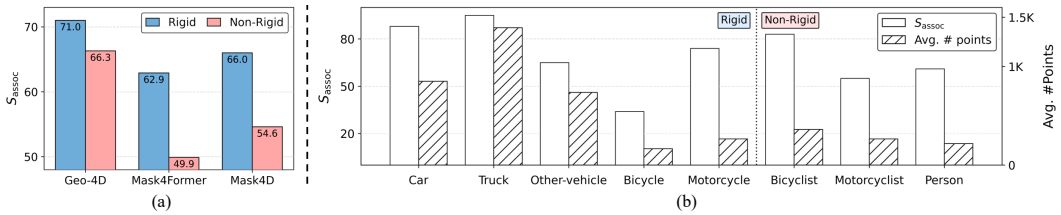

Figure 4: (a) Comparison with representative baselines across different association strategy categories on rigid and non-rigid object classes. (b) Class-wise performance with respect to object class types and the average number of points per instance.

can be obtained even in the presence of outliers, as the estimation is not dominated by local proximity. Figure 3 visualizes the point-wise correspondences obtained from two consecutive scans for 20 randomly selected points, where the current scan is used as the source and the previous scan as the target. For this analysis, we extract each point segment at the initial stage of ICP, as this stage retains its geometric structure intact. We observe that the nearest-neighbor matching is biased toward the closest point, ignoring the underlying point-wise relationships of the instance (red lines). In contrast, applying GGSM on the baseline maintains the geometric structure, yielding correspondences that align consistently with the instance's motion direction (green lines).

Addressing Q5. We categorize object classes into rigid (car, truck, other-vehicle, bicycle, motorcycle) and non-rigid (bicyclist, motorcyclist, person) groups based on their physical characteristics. As shown in Figure 4(a), GEO-4D outperforms representative IoU-based and query-based association baselines across both rigid and non-rigid classes. Furthermore, GEO-4D consistently maintains strong association performance across all object classes (see Figure 4(b)). These results are attributed to the robustness of our proposed global geometry-aware association strategy, which captures holistic geometric structure and enables reliable association under partial non-rigid motion and significant variations in object size.

4.4 ABLATION STUDY

We conduct a series of ablation studies to validate the effectiveness of each component in our framework. All experiments are performed on the SemanticKITTI (Behley et al., 2019) validation set, with GEO-4D integrated into Mask4Former (Yilmaz et al., 2024).

Module Design. Table 5(a) shows the effects of our proposed methods. Exp. #1 directly applies the vanilla Iterative Closest Point (ICP) for associating instances, serving as a baseline to examine the feasibility of ICP in an unsupervised manner. Despite its strong matching capability, ICP is highly sensitive to the quality of initial alignment. In Exp. #2, we observe a little performance gain when applying histogram-based initialization proposed by (Lin & Caesar, 2024), which stabilizes the initial alignment. Notably, our Global Geometry-aware Soft Matching (GGSM) brings further improvements without the influence of initial pose (see Exp. #3). The key takeaway is that instance-aware correspondence to alleviate the instability of local point matching is highly beneficial for obtaining reliable matching points between noisy instance segments, and as a result helps to estimate an accurate transformation. Exp. #4 demonstrates that coupling these components yields a complementary effect, leading to more robust instance association and improved overall performance. Lastly, Exp. #5

Table 5: (a) Effect of each module in GEO-4D. (b) Effect of handling different instance types.

(a) Module design

Exp.	Components			Metric	
	Hist. Init.	GGSM	Memory	LSTQ↑	S_{assoc} ↑
1	–	–	–	70.1	73.4
2	✓	–	–	71.3	75.4
3	–	✓	–	71.4	75.6
4	✓	✓	–	72.0	76.8
5	✓	✓	✓	**72.7**	**78.3**

(b) Instance-type handling

Instance Type			Metric		
Static	Dynamic	Missing	LSTQ↑	S_{assoc} ↑	Time (s)↓
✓	–	–	35.4	18.5	0.82
–	✓	–	71.8	76.7	1.86
✓	✓	–	72.0	76.8	0.96
✓	✓	✓	**72.7**	**78.3**	1.10

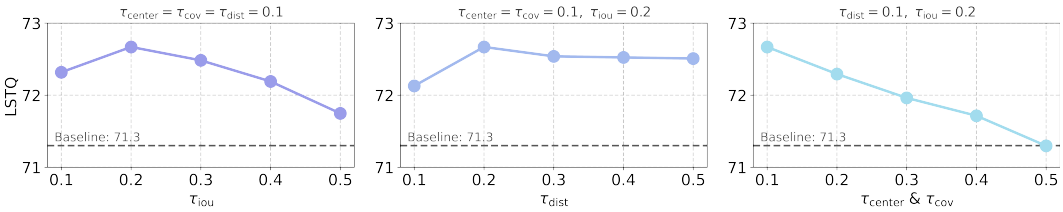

Figure 5: We illustrate the performance variation w.r.t. the hyperparameters of GEO-4D. The dashed line denotes the average performance of three different association methods: Mask4Former (Yilmaz et al., 2024), Mask4D (Marcuzzi et al., 2023), and CA-Net (Marcuzzi et al., 2022).

demonstrates the effectiveness of incorporating a memory bank for handling occluded and missing instances, resulting in superior performance.

Effect of Different Types of Instances. To validate our carefully designed pipeline with respect to different instance types, Table 5(b) presents the association results and average inference time per a single scan. First, we employ a statistics-based geometric filtering for static instances to match candidates over all detected objects. Despite requiring minimal computation, only a small subset of instances remain consistently tracked, 18.5 percent points indicating that static cues alone are insufficient for reliable association. In contrast, relying solely on global geometry association achieves high performance, but is computationally expensive. Thus, combining these strategies reduces the runtime by 48.4% while simultaneously improving performance, as it significantly narrows the matching candidates for global geometry association. Consequently, our full GEO-4D enhances the tracking continuity without sacrificing computational efficiency.

Hyperparameter Analysis In this section, we examine the impact of different matching thresholds on instance association performance (see Figure 5). We empirically determine the matching thresholds $\{\tau_{iou}, \tau_{dist}, \tau_{center}, \tau_{cov}\}$ over $\{0.1, 0.2, 0.3, 0.4, 0.5\}$. Here, we observe three insights. First, low value of threshold τ_{iou} often yields decent performance due to over-association, whereas high thresholds lead to over-filtering and degraded matching quality. Second, large τ_{center} and τ_{cov} cause static instances to be matched too liberally, leading to unreliable association. Lastly, our proposed GEO-4D delivers performance gains over the baselines, regardless of the hyperparameter choices.

5 CONCLUSION

In this paper, we introduce a novel instance association strategy GEO-4D for 4D LiDAR panoptic segmentation, the first training-free framework tailored for a single scan setting. Our proposed Global Geometry-aware Soft Matching forming instance-aware correspondence enables ICP procedure to precisely align between two consecutive instance point segments, thereby improving temporal continuity. Additionally, we systematically design the overall pipeline, underscoring the instance types to alleviate the inherent challenges of ICP and LiDAR sensors. Through experimental evaluations on both SemanticKITTI and nuScenes, our GEO-4D showcases its superiority not only on the validation sets but also on the hidden test sets, demonstrating robustness across various sensor configurations. Ultimately, our approach offers a promising direction for 4D perception by solely leveraging strong 3D panoptic models to support temporal understanding. **Limitations.** Although GEO-4D shows substantial promise in advancing 4D perception, several limitations persist that warrant further investigation, particularly in real-time deployment, handling partial LiDAR observations, and dependence on accurate ego pose estimation. We provide additional discussions in the supplementary materials.

REFERENCES

Ali Athar, Enxu Li, Sergio Casas, and Raquel Urtasun. 4d-former: Multimodal 4d panoptic segmentation. In *Conference on Robot Learning*, pp. 2151–2164. PMLR, 2023.

Mehmet Aygun, Aljosa Osep, Mark Weber, Maxim Maximov, Cyrill Stachniss, Jens Behley, and Laura Leal-Taixe. 4d panoptic lidar segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5527–5537, June 2021.

Mokhtar S Bazaraa, John J Jarvis, and Hanif D Sherali. *Linear programming and network flows*. John Wiley & Sons, 2011.

Jens Behley, Martin Garbade, Andres Milioto, Jan Quenzel, Sven Behnke, Cyrill Stachniss, and Jurgen Gall. Semantickitti: A dataset for semantic scene understanding of lidar sequences. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 9297–9307, 2019.

Jon Louis Bentley. Multidimensional binary search trees used for associative searching. *Communications of the ACM*, 18(9):509–517, 1975.

Paul J Besl and Neil D McKay. Method for registration of 3-d shapes. In *Sensor fusion IV: control paradigms and data structures*, volume 1611, pp. 586–606. Spie, 1992.

Alex Bewley, Zongyuan Ge, Lionel Ott, Fabio Ramos, and Ben Upcroft. Simple online and realtime tracking. In *2016 IEEE international conference on image processing (ICIP)*, pp. 3464–3468. Ieee, 2016.

Sofien Bouaziz, Andrea Tagliasacchi, and Mark Pauly. Sparse iterative closest point. In *Computer graphics forum*, volume 32, pp. 113–123. Wiley Online Library, 2013.

Alvaro Parra Bustos and Tat-Jun Chin. Guaranteed outlier removal for point cloud registration with correspondences. *IEEE transactions on pattern analysis and machine intelligence*, 40(12): 2868–2882, 2017.

Ricardo JGB Campello, Davoud Moulavi, and Jörg Sander. Density-based clustering based on hierarchical density estimates. In *Pacific-Asia conference on knowledge discovery and data mining*, pp. 160–172. Springer, 2013.

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

Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. *Advances in neural information processing systems*, 26, 2013.

Whye Kit Fong, Rohit Mohan, Juana Valeria Hurtado, Lubing Zhou, Holger Caesar, Oscar Beijbom, and Abhinav Valada. Panoptic nuscenes: A large-scale benchmark for lidar panoptic segmentation and tracking. *IEEE Robotics and Automation Letters*, 7(2):3795–3802, 2022.

Fangzhou Hong, Hui Zhou, Xinge Zhu, Hongsheng Li, and Ziwei Liu. Lidar-based panoptic segmentation via dynamic shifting network. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 13090–13099, 2021.

Fangzhou Hong, Lingdong Kong, Hui Zhou, Xinge Zhu, Hongsheng Li, and Ziwei Liu. Unified 3d and 4d panoptic segmentation via dynamic shifting networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(5):3480–3495, 2024. doi: 10.1109/TPAMI.2023.3349304.

Juana Valeria Hurtado, Rohit Mohan, Wolfram Burgard, and Abhinav Valada. Mopt: Multi-object panoptic tracking. *arXiv preprint arXiv:2004.08189*, 2020.

Rudolph Emil Kalman. A new approach to linear filtering and prediction problems. 1960.

Alexander Kirillov, Kaiming He, Ross Girshick, Carsten Rother, and Piotr Dollár. Panoptic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9404–9413, 2019.

Lars Kreuzberg, Idil Esen Zulfikar, Sabarinath Mahadevan, Francis Engelmann, and Bastian Leibe. 4d-stop: Panoptic segmentation of 4d lidar using spatio-temporal object proposal generation and aggregation. In *European Conference on Computer Vision*, pp. 537–553. Springer, 2022.

Harold W Kuhn. The hungarian method for the assignment problem. *Naval research logistics quarterly*, 2(1-2):83–97, 1955.

Xin Lai, Jianhui Liu, Li Jiang, Liwei Wang, Hengshuang Zhao, Shu Liu, Xiaojuan Qi, and Jiaya Jia. Stratified transformer for 3d point cloud segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 8500–8509, 2022.

Jiayuan Li. A practical $o(n^2)$ outlier removal method for correspondence-based point cloud registration. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(8):3926–3939, 2021.

Jinke Li, Xiao He, Yang Wen, Yuan Gao, Xiaoqiang Cheng, and Dan Zhang. Panoptic-phnet: Towards real-time and high-precision lidar panoptic segmentation via clustering pseudo heatmap. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 11809–11818, 2022.

Xueqian Li, Jhony Kaesemodel Pontes, and Simon Lucey. Neural scene flow prior. *Advances in neural information processing systems*, 34:7838–7851, 2021.

Yang Li and Tatsuya Harada. Lepard: Learning partial point cloud matching in rigid and deformable scenes. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 5554–5564, 2022a.

Yang Li and Tatsuya Harada. Non-rigid point cloud registration with neural deformation pyramid. *Advances in Neural Information Processing Systems*, 35:27757–27768, 2022b.

Yancong Lin and Holger Caesar. Icp-flow: Lidar scene flow estimation with icp. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15501–15511, 2024.

Huanyu Liu, Chao Peng, Changqian Yu, Jingbo Wang, Xu Liu, Gang Yu, and Wei Jiang. An end-to-end network for panoptic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 6172–6181, 2019.

Zhijian Liu, Xinyu Yang, Haotian Tang, Shang Yang, and Song Han. Flatformer: Flattened window attention for efficient point cloud transformer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 1200–1211, 2023.

Rodrigo Marcuzzi, Lucas Nunes, Louis Wiesmann, Ignacio Vizzo, Jens Behley, and Cyrill Stachniss. Contrastive instance association for 4d panoptic segmentation using sequences of 3d lidar scans. *IEEE Robotics and Automation Letters*, 7(2):1550–1557, 2022.

Rodrigo Marcuzzi, Lucas Nunes, Louis Wiesmann, Elias Marks, Jens Behley, and Cyrill Stachniss. Mask4d: End-to-end mask-based 4d panoptic segmentation for lidar sequences. *IEEE Robotics and Automation Letters*, 8(11):7487–7494, 2023.

Tim Meinhardt, Alexander Kirillov, Laura Leal-Taixe, and Christoph Feichtenhofer. Trackformer: Multi-object tracking with transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 8844–8854, 2022.

Andres Milioto, Jens Behley, Chris McCool, and Cyrill Stachniss. Lidar panoptic segmentation for autonomous driving. In *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 8505–8512. IEEE, 2020.

Rohit Mohan and Abhinav Valada. Efficientps: Efficient panoptic segmentation. *International Journal of Computer Vision*, 129(5):1551–1579, 2021.

Richard A Newcombe, Dieter Fox, and Steven M Seitz. Dynamicfusion: Reconstruction and tracking of non-rigid scenes in real-time. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 343–352, 2015.

Ziqi Pang, Zhichao Li, and Naiyan Wang. Simpletrack: Understanding and rethinking 3d multi-object tracking. In *European Conference on Computer Vision*, pp. 680–696. Springer, 2022.

Zheng Qin, Le Wang, Sanping Zhou, Panpan Fu, Gang Hua, and Wei Tang. Towards generalizable multi-object tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18995–19004, 2024.

Ryan Razani, Ran Cheng, Enxu Li, Ehsan Taghavi, Yuan Ren, and Liu Bingbing. Gp-s3net: Graph-based panoptic sparse semantic segmentation network. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 16076–16085, 2021.

Szymon Rusinkiewicz and Marc Levoy. Efficient variants of the icp algorithm. In *Proceedings third international conference on 3-D digital imaging and modeling*, pp. 145–152. IEEE, 2001.

Radu Bogdan Rusu, Nico Blodow, and Michael Beetz. Fast point feature histograms (fpfh) for 3d registration. In *2009 IEEE international conference on robotics and automation*, pp. 3212–3217. IEEE, 2009.

Aleksandr Segal, Dirk Haehnel, and Sebastian Thrun. Generalized-icp. In *Robotics: science and systems*, volume 2, pp. 435. Seattle, WA, 2009.

Mattia Segu, Luigi Piccinelli, Siyuan Li, Yung-Hsu Yang, Bernt Schiele, and Luc Van Gool. Samba: Synchronized set-of-sequences modeling for multiple object tracking. *arXiv preprint arXiv:2410.01806*, 2024.

Richard Sinkhorn. Diagonal equivalence to matrices with prescribed row and column sums. *The American Mathematical Monthly*, 74(4):402–405, 1967.

Cédric Villani et al. *Optimal transport: old and new*, volume 338. Springer, 2008.

Paul Voigtlaender, Michael Krause, Aljosa Osep, Jonathon Luiten, Berin Balachandar Gnana Sekar, Andreas Geiger, and Bastian Leibe. Mots: Multi-object tracking and segmentation. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition*, pp. 7942–7951, 2019.

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*, pp. 3523–3532, 2019.

Xinshuo Weng, Jianren Wang, David Held, and Kris Kitani. Ab3dmot: A baseline for 3d multi-object tracking and new evaluation metrics. *arXiv preprint arXiv:2008.08063*, 2020.

Nicolai Wojke, Alex Bewley, and Dietrich Paulus. Simple online and realtime tracking with a deep association metric. In *2017 IEEE international conference on image processing (ICIP)*, pp. 3645–3649. IEEE, 2017.

Xiaoyang Wu, Yixing Lao, Li Jiang, Xihui Liu, and Hengshuang Zhao. Point transformer v2: Grouped vector attention and partition-based pooling. *Advances in Neural Information Processing Systems*, 35:33330–33342, 2022.

Xiaoyang Wu, Li Jiang, Peng-Shuai Wang, Zhijian Liu, Xihui Liu, Yu Qiao, Wanli Ouyang, Tong He, and Hengshuang Zhao. Point transformer v3: Simpler faster stronger. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 4840–4851, 2024.

Yuwen Xiong, Renjie Liao, Hengshuang Zhao, Rui Hu, Min Bai, Ersin Yumer, and Raquel Urtasun. Upsnet: A unified panoptic segmentation network. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 8818–8826, 2019.

Jiaolong Yang, Hongdong Li, Dylan Campbell, and Yunde Jia. Go-icp: A globally optimal solution to 3d icp point-set registration. *IEEE transactions on pattern analysis and machine intelligence*, 38(11):2241–2254, 2015.

Kadir Yilmaz, Jonas Schult, Alexey Nekrasov, and Bastian Leibe. Mask4former: Mask transformer for 4d panoptic segmentation. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 9418–9425. IEEE, 2024.

Fangao Zeng, Bin Dong, Yuang Zhang, Tiancai Wang, Xiangyu Zhang, and Yichen Wei. Motr: End-to-end multiple-object tracking with transformer. In *European conference on computer vision*, pp. 659–675. Springer, 2022.

Tianyuan Zhang, Xuanyao Chen, Yue Wang, Yilun Wang, and Hang Zhao. Mutr3d: A multi-camera tracking framework via 3d-to-2d queries. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4537–4546, 2022a.

Xiyu Zhang, Jiaqi Yang, Shikun Zhang, and Yanning Zhang. 3d registration with maximal cliques. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 17745–17754, 2023a.

Yifu Zhang, Peize Sun, Yi Jiang, Dongdong Yu, Fucheng Weng, Zehuan Yuan, Ping Luo, Wenyu Liu, and Xinggang Wang. Bytetrack: Multi-object tracking by associating every detection box. In *European conference on computer vision*, pp. 1–21. Springer, 2022b.

Yuang Zhang, Tiancai Wang, and Xiangyu Zhang. Motrv2: Bootstrapping end-to-end multi-object tracking by pretrained object detectors. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 22056–22065, 2023b.

Yushan Zhang, Aljoša Ošep, Laura Leal-Taixé, and Tim Meinhardt. Zero-shot 4d lidar panoptic segmentation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 24506–24517, 2025.

Xingyi Zhou, Vladlen Koltun, and Philipp Krähenbühl. Tracking objects as points. In *European conference on computer vision*, pp. 474–490. Springer, 2020.

Zixiang Zhou, Yang Zhang, and Hassan Foroosh. Panoptic-polarnet: Proposal-free lidar point cloud panoptic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 13194–13203, 2021.

Minghan Zhu, Shizhong Han, Hong Cai, Shubhankar Borse, Maani Ghaffari, and Fatih Porikli. 4d panoptic segmentation as invariant and equivariant field prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 22488–22498, 2023.

A Implementation Details

A.1 Baseline Details

In this section, we provide a detailed overview of the strong baselines considered in our comparison: 4D-PLS (Aygun et al., 2021) (CVPR'21), 4D-Stop (ECCVW'22) (Kreuzberg et al., 2022), EQ-4D-Stop (ICCV'23) (Zhu et al., 2023), CA-Net (R-AL'22) (Marcuzzi et al., 2022), Mask4D (R-AL'23) (Marcuzzi et al., 2023), and Mask4Former (ICRA'24) (Yilmaz et al., 2024). Note that we report the official results as stated in the original papers, and reproduce the results when formal reports are not available. The following paragraph details the experimental settings and notable modifications for each baseline.

First, 4D-PLS, 4D-Stop, and EQ-4D-Stop follow a two-stage design, where instance tracklets are first generated independently for all scans and then stitched across consecutive frames. Accordingly, we measure the inference time of each stage separately and report their average as the overall runtime. Second, Mask4Former is originally developed for 4D LiDAR panoptic segmentation with superimposed point clouds. To establish a strong baseline for our experiments, we train it as a 3D panoptic network without any modifications, except setting the number of input scans to one. Third, CA-Net employs contrastive learning for instance association on top of an off-the-shelf 3D panoptic segmentation network. For a fair comparison, we integrate the single scan Mask4Former into the CA-Net framework. Since CA-Net requires point-wise instance features for contrastive learning, we extract them by computing the dot product between instance queries and instance masks. Lastly, we additionally reproduce the baselines on panoptic nuScenes Fong et al. (2022) using their official repositories, following the experimental protocol of EQ-4D-StOP.

A.2 METRICS

In this section, we provide detailed explanations of the metrics employed for evaluating association performance in our experiments. While LSTQ is the most widely used metric, we additionally report two complementary metrics, MOTSA and PTQ, to provide a more comprehensive assessment.

LSTQ (Aygun et al., 2021) is delicately designed to evaluate 4D panoptic segmentation quality. First, to assess the semantic segmentation performance over the classes C, the classification score S_{cls} is defined as:

$$S_{cls} = \frac{1}{|C|} \sum_{c=1}^{C} \frac{|\mathrm{TP}_c|}{|\mathrm{TP}_c| + |\mathrm{FN}_c| + |\mathrm{FP}_c|} = \frac{1}{C} \sum_{c=1}^{C} IoU(c), \tag{11}$$

where TP, FN, and FP denote the true positives, false negatives, and false positives, respectively. Furthermore, the association score S_{assoc}, which is responsible for measuring temporal continuity, is defined as:

$$S_{assoc} = \frac{1}{|I|} \sum_{t \in I} \frac{1}{|gt_{id}(t)|} \sum_{s \in S,\ s \cap t \neq 0} TPA(s,t) IoU(s,t), \tag{12}$$

where I and S refer to the ground-truth and predicted segments. Note that TPA represents the points that maintain consistent identities across scans, regardless of class. Consequently, LSTQ is computed as the geometric mean of the classification score S_{cls} and association score S_{assoc}.

MOTSA & sMOTSA (Voigtlaender et al., 2019) are mask-based variants of the widely used Multi-Object Tracking Accuracy (MOTA) metric, replacing bounding-box overlaps with mask IoU to better evaluate segmentation-level tracking performance. Formally,

$$\text{MOTSA} = \frac{|TP| - |FP| - |IDSW|}{|M|}, \tag{13}$$

where $|IDSW|$ denotes the total number of ID switches, and $|M|$ is the number of ground truth segmentation masks. In addition, sMOTSA replaces the true positive (TP) with soft true positive (sTP), obtained by accumulating the mask IoUs over all matched predictions.

PTQ & sPTQ (Hurtado et al., 2020) are formulated to satisfy the fundamental requirement that panoptic tracking metrics must jointly capture segmentation accuracy and tracking consistency over time. Thus, they incorporate an ID-switch penalty into the original Panoptic Quality (PQ) measure, as follows:

$$\text{PTQ} = \frac{1}{|C|} \sum_{c=1}^{C} \frac{\sum_{(s,t) \in TP_c} IoU(s,t) - |IDSW_c|}{|TP_c| + \frac{1}{2}|FP_c| + \frac{1}{2}|FN_c|}. \tag{14}$$

Furthermore, sPTQ leverages the soft version of IDSW (sIDSW) instead of IDSW. This soft variant aggregates the IoU scores for segments in which the predicted instance ID ID changes:

$$sIDSW_c = \{IoU(s,\ t)\ \mid\ (s,\ t) \in TP_c \wedge ID_s \neq ID_t\}. \tag{15}$$

B METHOD DETAILS

B.1 PIPELINE DETAILS

The overall association pipeline of our proposed GEO-4D is summarized in Algorithm 1. We apply GEO-4D sequentially across all scans within each sequence, and reset the memory bank $\mathcal{B}$ at the beginning of every new sequence.

B.2 COMPLEXITY ANALYSIS

In this section, we provide a computational complexity comparison between Global Geometry-aware Soft Matching (GGSM) and nearest neighbor-based matching. As described in Section.3.2, we utilize

Algorithm 1 Overall process of GEO-4D

Require: Panoptic prediction (s, l), memory bank $\mathcal{B}$
1: **for** sequence $t \in \{1, 2, \ldots, T-1\}$ **do**
2: Construct the candidate set $\mathcal{M}' = \mathcal{G}'_{\text{src}} \times \mathcal{G}'_{\text{dst}}$ filtered by semantic predictions s.
// Static instances
3: Associate instances using the centroid μ of each point set (Eq. 5).
4: Refine the matched pairs using the covariance descriptor σ (Eq. 7).
// Dynamic instances
5: Perform global geometric association using ICP enhanced with global geometry-aware soft matching.
// Missing instances
6: **if** $\mathcal{B}$ is not empty **then**
7: Apply the dynamic matching procedure (Lines 5) to missing instances.
8: **end if**
9: Update the memory bank $\mathcal{B}$.
10: **end for**
Ensure: Consistent instance IDs across the entire sequence.

the entropy-regularized optimal transport problem (Cuturi, 2013) to obtain instance-aware point correspondences. Before analyzing the regularized form, we briefly review the optimal transport problem. Given two probability vectors $r \in \mathbb{R}_+^n$ and $c \in \mathbb{R}_+^m$, the original optimal transport problem (Villani et al., 2008) aims to find the optimal transport plan $\mathcal{Q} \in U(r,\ c)$. The entries of $\mathcal{Q}$ represent the amount of mass transported from the i-th element of r to the j-th element of c. The transport polytope U is defined as:

$$U(r,\ c) = \{Q \in \mathbb{R}_+^{n \times m} | \sum_j Q_{ij} = r_i,\ \forall i,\ \sum_i Q_{ij} = c_j,\ \forall j\}. \tag{16}$$

Let $\mathcal{Z}$ denote the transport cost matrix. The optimal transport plan is then computed as the Frobenius inner product between $\mathcal{Q}$ and $\mathcal{Z}$:

$$\text{OT}(r, c) := \min_{\mathcal{Q}} \langle \mathcal{Z},\ \mathcal{Q} \rangle_F. \tag{17}$$

This optimization is a linear program, for which the network simplex (Bazaraa et al., 2011) is the most widely used exact solver. In the symmetric case $n = m$, its worst-case computational complexity scales as $O(n^3 \log n)$. However, due to its heavy computational cost, adding an entropic regularization is widely adopted to accelerate optimal transport. By incorporating the Shannon entropy $H(\mathcal{Q}) := -\sum_{i,j} \mathcal{Q}_{ij} \log \mathcal{Q}_{ij}$ into Eq. 17, we obtain the entropy-regularized optimal transport formulation. To handle marginal constraints, we introduce dual variables $f \in \mathbb{R}^n$ and $g \in \mathbb{R}^m$ and form the corresponding Lagrangian. Then, by taking the first-order optimality condition, we find that optimal solution $\hat{\mathcal{Q}}$ has matrix form $\hat{\mathcal{Q}} = \text{diag}(u) K \text{diag}(v)$, where $u_i = e^{f_i/\epsilon}$, $v_j = e^{g_j/\epsilon}$, and $K_{ij} = e^{-Z_{ij}/\epsilon}$. Since u and v must satisfy the marginal constraints, the entropic regularization transforms the classical optimal transport problem from a linear program into a matrix scaling problem. Thus, when $n = m$, the computational complexity is reduced to $O(n^2)$. However, the nearest neighbor algorithm operates in $O(n \log n)$ on average and up to $O(n^2)$ in the worst case in practice, as implemented in the scikit-learn library (e.g., using KD-Tree (Bentley, 1975)). Consequently, our GGSM solution is theoretically comparable to the nearest neighbor algorithm.

B.3 OPTION: BIJECTIVE ASSIGNMENT.

In global geometric association, a naive approach is to pair each source instance with its most similar target instance; however, it fails to enforce one-to-one correspondence. Thus, we derive a transformation-aware matching cost that integrates the ICP results—rotation $\mathbf{R}$, translation $\mathbf{t}$, and the IoU score s_{iou}. Formally, the matching cost $\mathcal{C}$ of dynamic matching pairs $(i, j) \in \mathcal{M}''$ is defined as:

$$\begin{aligned} \mathcal{C}_{ij} &= \gamma_t c_{ij}^{(t)} + \gamma_r\, c_{ij}^{(r)} + \gamma_s\, c_{ij}^{(s)}, \\ c_{ij}^{(t)} = \|\mathbf{t}_{ij}\|_2, &\quad c_{ij}^{(r)} = \|\theta_{ij}\|_2, \quad c_{ij}^{(s)} = \|1 - s_{iou}\|_2, \end{aligned} \tag{18}$$

Table A.1: SemanticKITTI (Behley et al., 2019) Test Set Leaderboard (Top 5). mAQ and mIoU denote S_{assoc} and S_{cls}, respectively.

Rank	User	LSTQ ↑	mAQ ↑	mIoU ↑
1	**GEO-4D (Ours)**	**70.3**	70.0	**70.5**
2	guysh	70.2	70.6	69.7
2	YAKD	68.6	**70.8**	66.5
3	hphnngcquan	68.5	70.4	66.7
4	KadirYilmaz	68.4	67.3	69.6

where θ denotes the axis–angle rotation vector obtained from $\mathbf{R}$. We normalize each cost term to comparable scales before weighting so that γ_t, γ_r, and γ_s control their relative contributions. By minimizing $\mathcal{C}$ with the Hungarian algorithm (Kuhn, 1955), we obtain a bijective assignment that considers the geometric alignment.

B.4 LEADERBOARD RESULTS

Table A.1 shows the SemanticKITTI (Behley et al., 2019) challenge leaderboard ranks on the hidden test set, ordered by LSTQ. In this leaderboard, we observe that our proposed GEO-4D outperforms not only all published methods but also every competition participant. Specifically, we achieve 1st place in LSTQ as of submission date. These results demonstrate the strong generalization capability and practical effectiveness of our method under challenging scenarios.

C ADDITIONAL EXPERIMENTAL RESULTS

C.1 ANALYSIS OF MATCHING STRATEGY.

In Table A.2, we compare the performance of GEO-4D with one-to-one or greedy matching. The evaluation is done using SemanticKITTI validation set and set all $\gamma_{\{t,r,s\}} = 1$ for the Hungarian matching. We observe that greedy matching surpasses the bijective matching by a small margin in S_{assoc}. We interpret this gap as being caused by segmentation noise, where a single object is often predicted as multiple disjoint segments. In such case, the greedy strategy adapts to this variability, resulting in more robust temporal consistency. However, as segmentation quality improves, bijective matching becomes advantageous for maintaining unique and consistent identities.

Table A.2: Ablation on matching strategy.

Method	LSTQ	S_{assoc}	S_{cls}
Hungarian Matching	72.3	77.6	67.5
Greedy Matching	72.7	78.3	67.5

C.2 EFFECT OF THE NUMBER OF SCANS.

In Table A.3, we examine how well our proposed GEO-4D performs instance association with only a single scan. The number of superimposed scans is a pivotal factor influencing the association performance of the IoU-based method. Since formal reports for specific configurations are unavailable, we reproduce the results using official implementation—namely, Eq-4D-Stop (Zhu et al., 2023) with 2 scans and Mask4Former (Yilmaz et al., 2024) with 4 scans. As clearly observed in Table A.3, leveraging four superimposed scans yields a substantial improvement in S_{assoc}, with a notable sacrifice in computational efficiency. In contrast, we observe that GEO-4D exhibits superior performance, despite the absence of explicit 4D spatio-temporal volume constructed from stacked point clouds. This observation demonstrates that GEO-4D effectively associates instances using only a single scan and a 3D panoptic network, while remaining highly efficient.

C.3 CONTROLLED SAME-MASK COMPARISON WITH DETECT & TRACK ASSOCIATION.

We further conduct a controlled same-mask comparison with CA-Net (Marcuzzi et al., 2022) on SemanticKITTI dataset (Behley et al., 2019). For each segmentation backbone, we fix the frame-wise

Table A.3: Analysis on performance over the number of superimposed scans.

Method	# Scan	Memory (GB)	Time (s)	Validation set LSTQ	S_{assoc}	S_{cls}	IoU^{St}	IoU^{Th}	Test set LSTQ	S_{assoc}	S_{cls}	IoU^{St}	IoU^{Th}
(a) IoU-based Association													
4D-PLS (Aygun et al., 2021)	2	-	-	59.9	58.8	61.0	65.0	63.1	-	-	-	-	-
	4	8.0	3.48	62.7	65.1	60.5	65.4	61.3	56.9	56.4	57.4	66.9	51.6
4D-Stop (Kreuzberg et al., 2022)	2	15.9	3.67	66.4	71.8	61.4	64.9	64.1	62.9	67.3	58.8	68.3	53.3
	4	18.9	4.32	67.0	74.4	60.3	65.3	60.9	63.9	69.5	58.8	67.7	53.8
Eq-4D-Stop (Zhu et al., 2023)	2	11.3	2.89	68.9	74.8	63.5	65.7	68.4	65.4	69.7	61.5	68.7	59.1
	4	12.1	3.18	70.1	77.6	63.4	66.4	67.1	67.0	**72.0**	62.4	69.1	60.9
Mask4Former (Yilmaz et al., 2024)	2	10.4	3.51	72.0	76.3	68.0	66.2	70.5	68.4	67.3	69.6	72.7	65.3
	4	17.1	8.56	71.9	76.3	67.8	66.9	69.0	68.0	69.7	66.3	70.7	60.3
(c) Detect & Track Association													
GEO-4D (Ours)	1	**3.2**	**1.10**	**72.7**	**78.3**	67.5	65.2	70.6	**70.3**	70.0	70.5	72.7	67.4

Table A.4: Comparison of Geo-4D with CA-Net using different frame-wise segmentation backbones.

Method	# Scan	Assoc. Training	Validation set LSTQ	S_{assoc}	S_{cls}	IoU^{St}	IoU^{Th}	Test set LSTQ	S_{assoc}	S_{cls}	IoU^{St}	IoU^{Th}
DS-Net (Hong et al., 2021)	1	-	-	-	63.5	64.6	62.0	-	-	60.6	66.9	52.0
+ CA-Net (Marcuzzi et al., 2022)	1	✓	67.4	71.6	63.5	64.6	62.0	62.1	63.7	60.6	66.9	52.0
+ **GEO-4D *(Ours)***	1	✗	68.5 (+1.1)	74.0 (+2.4)	63.5	64.6	62.0	62.9 (+0.8)	65.3 (+1.6)	60.6	66.9	52.0
Mask4Former* (Yilmaz et al., 2024)	1	-	-	-	67.5	65.2	70.6	-	-	70.5	72.7	67.4
+ CA-Net*	1	✓	70.4	73.4	67.5	65.2	70.6	66.5	62.8	70.5	72.7	67.4
+ **GEO-4D *(Ours)***	1	✗	**72.7** (+2.3)	**78.3** (+4.9)	67.5	65.2	70.6	**70.3** (+3.8)	**70.0** (+7.2)	70.5	72.7	67.4

predictions and apply either CA-Net or Geo-4D as the association strategy. CA-Net is trained following the official implementation and its recommended training setup. As shown in Table A.4, Geo-4D consistently improves association quality under identical frame-wise predictions, demonstrating that the gain comes from our association method rather than stronger segmentation masks, consistent with the findings reported in the main paper.

C.4 Analysis of Global Geometry-aware Soft Matching.

In the main paper, we argue that the proposed Global Geometry-aware Soft Matching (GGSM) improves the robustness of the matching strategy against noisy measurements and predictions. To isolate the contribution of GGSM, we additionally evaluate its effectiveness on non-rigid point cloud registration benchmarks (Li & Harada, 2022a;b) independent of tracking performance. This allows us to analyze whether GGSM improves geometric correspondence estimation under challenging conditions such as partial overlap, occlusion, and large motion.

Baselines. While more advanced baselines exist for non-rigid point cloud registration, we compare against several representative registration methods to better understand the effectiveness of GGSM. Specifically, we include Iterative Closest Point (ICP) (Besl & McKay, 1992) as a classical baseline, Sinkhorn-based registration which is closely related to our global geometry association, Non-rigid ICP (NICP) (Newcombe et al., 2015) as an extension of ICP for non-rigid registration, and Neural Scene Flow Prior (NSFP) (Li et al., 2021) which predicts point-wise motion using neural networks. ICP determines the global rotation and translation parameters, which are designed for rigid transformations. Optimal Transport (OT) (Cuturi, 2013) iteratively estimates point displacements by optimizing a transport plan via the Sinkhorn-Knopp algorithm (Sinkhorn, 1967) and updating point locations according to the gradients of the OT objective. NICP estimates the movement of all points individually rather than relying on a single rigid transformation. We adopt a typical solution that optimizes a deformation graph with regularization constraints to enforce smooth deformations. NSFP optimizes a multilayer perceptron (MLP) to predict point-wise motion by minimizing the Chamfer distance between two point clouds.

Metrics. We adopt widely used metrics to evaluate how well the registered points align with the ground truth. End-Point Error (EPE) measures the average 3D warping error across all points.

Table A.5: Quantitative comparison on 4DMatch-F (Li & Harada, 2022b) and 4DLoMatch-F (Li & Harada, 2022b).

	4DMatch-F				4DLoMatch-F				
Method	EPE↓	AccS↑	AccR↑	Outlier↓	EPE↓	AccS↑	AccR↑	Outlier↓	Time*
ICP (Besl & McKay, 1992)	0.296	2.96	12.06	71.50	0.565	0.14	0.74	90.87	0.10
Sinkhorn (Sinkhorn, 1967)	0.308	2.76	8.13	79.86	0.505	0.20	0.86	89.47	3.76
NICP (Newcombe et al., 2015)	0.325	6.44	12.10	80.40	0.517	0.18	0.73	92.37	4.80
NSFP (Li et al., 2021)	0.265	**8.66**	**18.65**	64.96	0.495	**0.38**	**1.56**	84.77	39.54
Ours (ICP+GGSM)	**0.210**	2.51	11.26	**62.58**	**0.431**	0.18	1.19	**78.80**	1.62

* Runtime measured on a single NVIDIA A100.

AccS and AccR represent the percentage of points whose relative errors are below 2.5% and 5.0%, respectively. Outlier denotes the percentage of points whose relative error exceeds 30%. EPE and Outlier evaluate the overall registration performance, while AccS and AccR measure the point-wise accuracy.

Experimental Analysis. As shown in Tab. A.5, two key observations emerge across both 4DMatch-F and 4DLoMatch-F. First, GGSM achieves the lowest EPE and Outlier rates despite relying only on rigid transformations. Although non-rigid methods achieve higher point-wise accuracy (AccS and AccR), they often introduce globally inconsistent deformations, which can degrade the overall geometric alignment. In contrast, GGSM leverages global geometric structure to estimate a coherent transformation, leading to more reliable overall registration performance. Second, GGSM provides a well-balanced trade-off between efficiency and performance. NSFP models point-wise motion by optimizing a neural network for each point coordinate, which offers strong flexibility but incurs significant computational overhead. Moreover, NICP estimates per-vertex deformation parameters through iterative non-rigid optimization, which also limits efficiency. On the other hand, our GGSM remains computationally efficient as it inherits the efficiency of rigid ICP, resulting in strong overall registration quality by improving geometric consistency.

Takeaway: In this section, we validate the effect of GGSM in a controlled setting by excluding other factors (e.g., rigid objects and driving surface in on-road environments). It is important to note that the primary scope of this work is 4D LiDAR panoptic segmentation in driving scenes, rather than non-rigid point cloud registration itself. These experimental results support our claim that GGSM plays a crucial role in enabling robust and reliable instance association in driving scenes with both rigid and non-rigid instances. For additional discussions, please refer to the corresponding analyses presented in the main manuscript.

C.5 Analysis of DBSCAN Instance Refinement.

In this specific experiment of integrating GEO-4D with Mask4Former, we apply DBSCAN clustering (Campello et al., 2013) as done in the original Mask4Former setting. Table A.6 summarizes the results, where we derive two major insights. First, GEO-4D achieves performance comparable to existing state-of-the-art methods without the additional clustering step, and further surpasses them when clustering is applied. Furthermore, we discover that the increase of runtime primarily results from the post-processing step. Without this step, GEO-4D delivers the most efficient runtime of 340 ms compared to all other methods, with only a marginal loss in performance. These observations clearly highlight that GEO-4D offers both high efficiency and strong performance, and can further benefit from improvements in the backbone network.

Table A.6: **Instance refinement.**

Setting	LSTQ↑	S_{assoc} ↑	S_{cls} ↑	Time (s)↓
w/o DBSCAN	71.8	75.8	67.5	**0.34**
w/ DBSCAN	**72.7**	**78.3**	67.5	1.10

C.6 Quantitative Results

We additionally provide the results of each semantic class on panoptic nuScenes (Fong et al., 2022) to facilitate a more comprehensive understanding. In Table A.7, our method yields overall performance improvements, particularly in safety-critical *thing* classes (e.g., bicycle, car, and pedestrian). Fur-

Table A.7: Comparison of the quantitative results on panoptic nuScenes (Fong et al., 2022).

Model	LSTQ	S_{assoc}	S_{cls}	Thing classes (AQ / IoU)										Stuff classes (IoU only)					
				barrier	bicycle	bus	car	const. veh.	motorcycle	pedestrian	traffic cone	trailer	truck	drive. surf.	other flat	sidewalk	terrain	manmade	vegetation
Mask4Former (Yilmaz et al., 2024)	67.6	63.0	72.5	48.3/76.9	42.6/37.0	60.5/90.2	76.7/92.7	42.7/44.9	61.1/77.0	53.5/74.0	48.9/55.7	38.7/47.8	62.2/77.8	95.6	69.2	71.5	74.6	88.5	87.0
Mask4D (Marcuzzi et al., 2023)	58.5	54.1	63.1	29.7/64.3	27.8/20.5	60.5/79.2	72.5/83.5	29.5/21.8	52.5/72.0	42.3/65.9	39.5/43.0	19.2/35.2	58.9/71.8	94.2	53.9	64.9	71.3	85.2	84.5
CA-Net (Marcuzzi et al., 2022)	63.4	56.6	71.0	30.1/74.5	31.9/29.1	56.5/88.5	73.3/91.2	38.6/43.2	63.2/78.7	53.6/71.2	34.6/54.6	39.5/48.3	62.1/77.3	95.6	68.4	70.3	72.7	86.7	85.2
Geo-4D (Ours)	**67.9**	**65.0**	71.0	40.7/74.5	44.1/29.1	58.9/88.5	80.9/91.2	40.1/43.2	65.9/78.7	67.2/71.2	52.0/54.6	40.6/48.3	65.7/77.3	95.6	68.4	70.3	72.7	86.7	85.2

thermore, classes with relatively fewer points, such as pedestrians and traffic cones, exhibit stronger association performance against the baselines. This observation aligns with our earlier analysis on SemanticKITTI (Behley et al., 2019) presented in the main manuscript.

C.7 Qualitative Results

In this section, we present additional qualitative analysis by illustrating the results of SemanticKITTI (Behley et al., 2019) and panoptic nuScenes (Fong et al., 2022) dataset. As clearly shown in Figure A.1 and A.2, Geo-4D maintains robust ID consistency even in challenging scenarios. Our method explicitly incorporates the inherent geometric relations between instances and adopts a delicately designed pipeline that is resilient to occlusion and noisy predictions. In contrast, Mask4Former (Yilmaz et al., 2024) associates instances by selecting the majority ID among spatially overlapped points, making it vulnerable to imprecise predictions and leading to frequent ID switches (see the 1st, 2nd, and 3rd rows of Figure A.1). Apart from this, Mask4D (Marcuzzi et al., 2023) struggles with separating nearby objects and accumulates errors from early mispredictions (see the 1st, 2nd, and 4th rows of Figure A.1). Furthermore, Geo-4D also operates well on the panoptic nuScenes dataset, which is sparse and large temporal gaps between consecutive scans. Specifically, in the 2nd and 4th rows of Figure A.2, we observe that the instances have no overlap and sparse representations. The baselines confuse the instance IDs due to their partial observations and rapid motions. However, Geo-4D successfully finds the corresponding instances even when they are distant. Also, from the 1st row of Figure A.2, we observe that Geo-4D can effectively associate instances in complex scene where pedestrians are heavily entangled.

D Discussion

D.1 Limitation and Future work

We identify several limitations of our proposed approach, which inherits issues from both pretrained 3D panoptic models and LiDAR sensors. While Geo-4D demonstrates strong association performance with a cost efficient pipeline, real-time deployment remains challenging. Our method highly depends on the performance of the underlying 3D panoptic model; thus, its runtime can be further improved as point cloud processing continues to advance. In addition, partial observation of instances in LiDAR sensing make it difficult to reliably associate instances. Incorporating trajectory cues represents a promising direction to further stabilize instance association. Lastly, our pipeline forces the requirement of camera parameters as a prerequisite for transforming the point coordinates into the world coordinates. Relaxing this requirement would improve the applicability of our method in broader real-world settings. We hope these insights encourage continued progress toward scalable 4D perception systems.

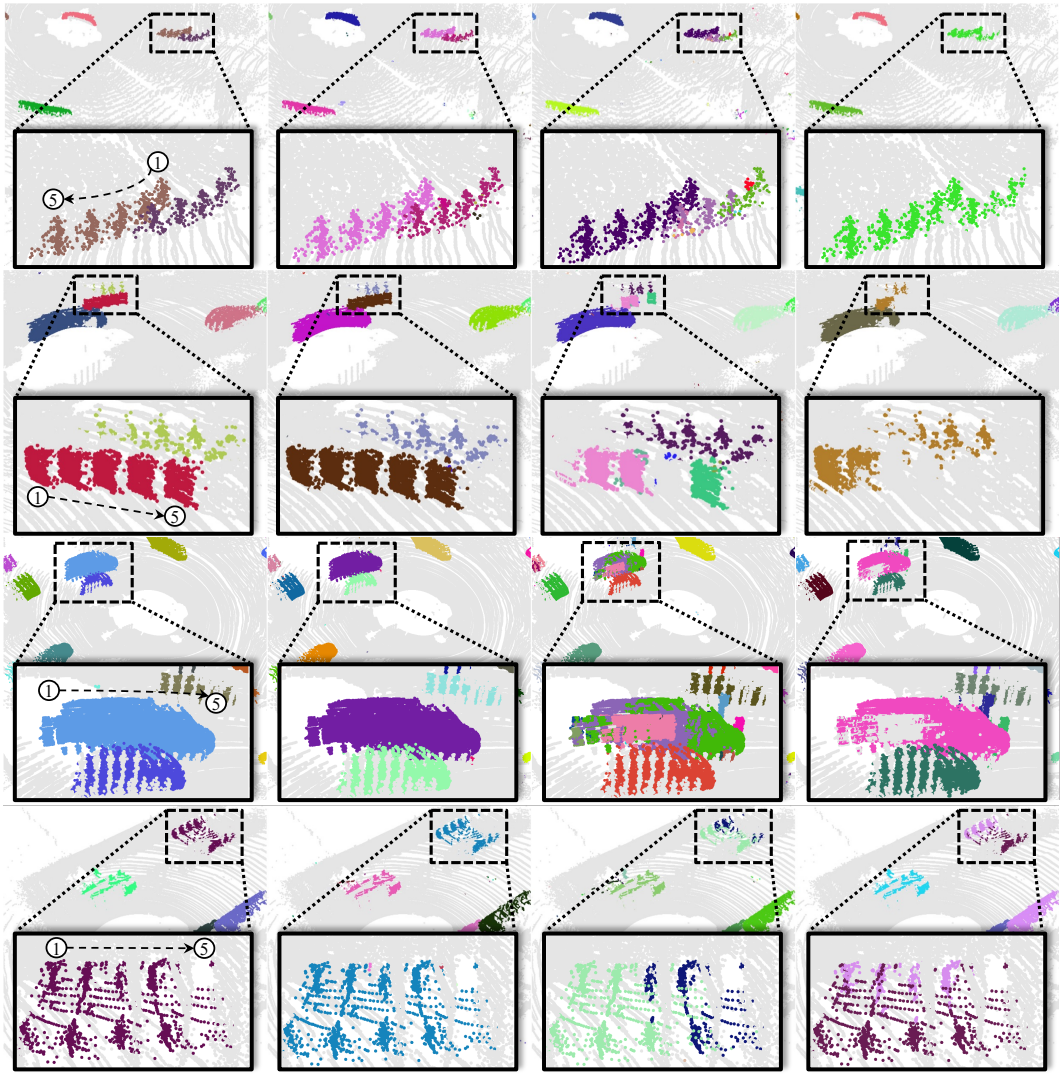

Figure A.1: Qualitative comparison on semanticKITTI (Behley et al., 2019) validation set. We visualize the association results over five consecutive scans for both the baselines and our GEO-4D. Different colors represent different instances. The dotted boxes zoom in for clear comparison. For clarity, we use circled numbers (①–⑤) to represent the frame indices.

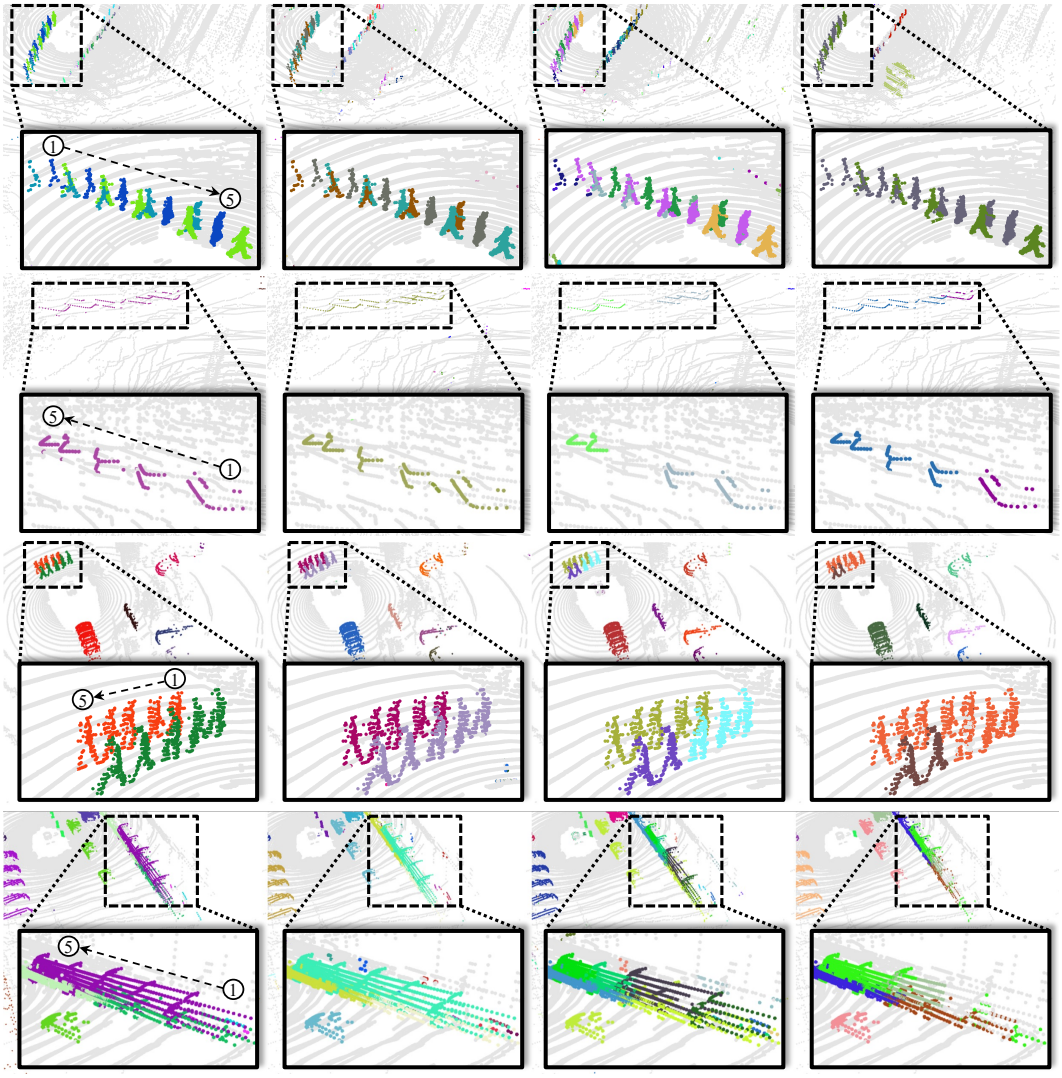

Figure A.2: Qualitative comparison on panoptic nuScenes (Fong et al., 2022). We visualize the association results over five consecutive scans for both the baselines and our GEO-4D. Different colors represent different instances. The dotted boxes zoom in for clear comparison. For clarity, we use circled numbers (①–⑤) to represent the frame indices.