

ELPTNet: An Efficient LiDAR-based 3D Pedestrian Tracking Network for Autonomous Navigation Social Robots

Jinzheng Guang, Zhenzhong Cao, YINUO Song, and Jingtai Liu*, *Senior Member, IEEE*

Abstract—Autonomous navigation social robots need to track pedestrian movements in real-time with high precision to optimize path planning and avoid collisions. However, the main challenge of pedestrian tracking lies in the significant variations in human posture, which differ from rigid-body structures like vehicles. In this paper, we propose an Efficient LiDAR-based 3D Pedestrian Tracking Network (ELPTNet). First, our ELPTNet employs a 3D object detector to extract directional 3D pedestrian bounding boxes from LiDAR point clouds. Then, our ELPTNet employs a Constant Acceleration (CA) model and prediction confidence for target trajectory prediction. During the data association process, it integrates geometric, appearance, and motion features to enhance the robustness and real-time performance of 3D MOT when targets are temporarily occluded. Experimental results demonstrate that our ELPTNet achieves the highest ranking on the large-scale JRDB dataset for the 3D tracking task, outperforming previous state-of-the-art (SOTA) methods with improvements of 8.4% in MOTA and 6.6% in HOTA. Additionally, our ELPTNet attains an inference speed of 61 frames per second (FPS) on a single CPU. Therefore, our method enables accurate and real-time tracking of multiple pedestrians. The code is publicly available at <https://github.com/jinzhengguang/ELPTNet>.

I. INTRODUCTION

With the rapid advancement of artificial intelligence technology [1], [2], [3], pedestrian tracking has become a critical task across various domains, such as autonomous driving [4], [5] and mobile robotics [6], [7]. In autonomous navigation social robot systems, accurate pedestrian localization and tracking are key to achieving autonomous navigation and providing comfortable services [8]. However, traditional image-based pedestrian tracking methods [9] are highly sensitive to lighting conditions, showing poor stability under low light and bright light environments, and lack depth perception capabilities. In contrast, LiDAR-based pedestrian tracking methods [10], [11], [12] can effectively overcome these limitations. LiDAR generates high-density 3D point cloud data by emitting laser pulses and receiving reflected signals, providing precise spatial information, thus enabling high-precision pedestrian localization and tracking in dynamic environments, offering significant advantages.

However, LiDAR-based pedestrian tracking methods still face numerous challenges. As illustrated in Fig. 1, one of the primary challenges is the significant variation in pedestrian postures over time. This variation is not only reflected in the

differences in posture among different pedestrians at a given moment but also in the fluctuations in the posture of the same pedestrian over different time intervals. Another challenge lies in the sparsity of LiDAR point cloud data, especially for small-scale pedestrian targets. Thus, the accuracy of pedestrian tracking using LiDAR remains low.

Regarding the aforementioned issues, we propose an Efficient LiDAR-based 3D Pedestrian Tracking Network (ELPTNet) for autonomous navigation social robots. Our ELPTNet is an online approach primarily composed of four key components: 3D pedestrian detection, state prediction, data association, and the state update module. Experimental results indicate that our ELPTNet surpasses the existing state-of-the-art (SOTA) approaches in both accuracy and efficiency. The three contributions of this paper are as follows:

- 1) We propose ELPTNet, a real-time online 3D MOT method that utilizes LiDAR 3D point cloud data for pedestrian tracking, effectively reducing interference caused by lighting variations.
- 2) Our ELPTNet employs a Constant Acceleration (CA) model and prediction confidence for target trajectory prediction. During the data association process, it integrates geometric, appearance, and motion features to enhance the robustness and real-time performance of 3D MOT when targets are temporarily occluded.
- 3) Experimental results demonstrate that our ELPTNet achieves SOTA performance on the large-scale JRDB dataset for the 3D tracking task. Compared to the second-best method, our ELPTNet improves MOTA and HOTA performance by 8.4% and 6.6%, respectively.

II. RELATED WORK

In recent years, 3D Multi-Object Tracking (MOT) has been categorized into two types based on the sensor used: Camera-based 2D MOT and LiDAR-based 3D MOT methods.

A. Camera-based 2D Multi-Object Tracking Methods

Most current 2D Multiple Object Tracking (MOT) methods follow a detection-based tracking framework, where data association is typically addressed through global optimization or online bipartite graph matching methods [13]. Global optimization methods, such as network flow [14], [15] and Bayesian filtering [16], optimize using information from all frames. In contrast, online bipartite graph matching methods rely only on information from the current frame to track targets, employing the Hungarian algorithm [17], [18] or greedy assignment algorithms [19]. Despite the success of many 2D MOT approaches, their ability to represent the dynamics of

*This work was supported by the National Natural Science Foundation of China under Grant 62173189. (Corresponding author: Jingtai Liu.)

The authors are with the Institute of Robotics and Automatic Information System and the TBI Center, Nankai University, Tianjin 300350, China, and also with Tianjin Key Laboratory of Intelligent Robotics, Tianjin 300350, China (e-mail: guangjinzheng@mail.nankai.edu.cn; liujt@nankai.edu.cn).

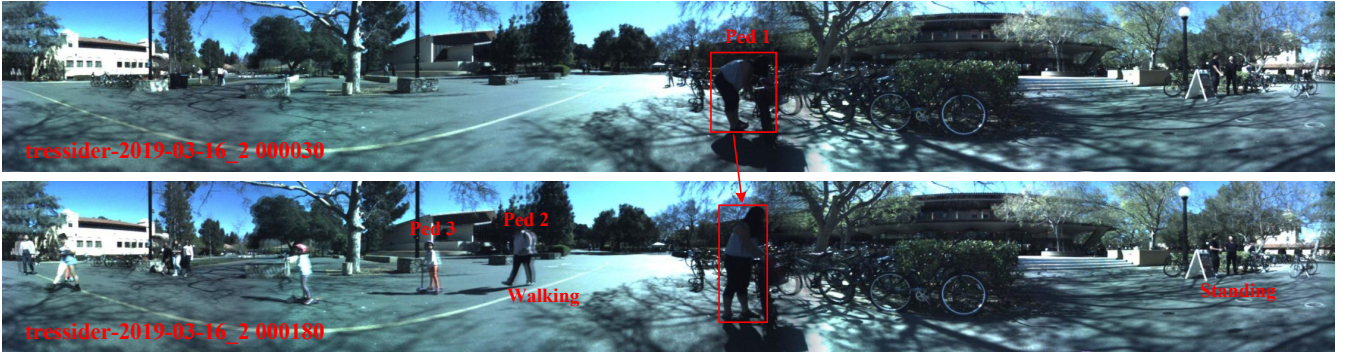


Fig. 1. The main challenge in pedestrian tracking is the significant variation in pedestrian postures. First, unlike rigid bodies (e.g., vehicles), the same individual (e.g., Ped 1) may exhibit different postures or states at different times, such as standing, bending, or sitting. Second, at the same time instant, different pedestrians may be in different postures, such as standing, walking, or bending over. Additionally, pedestrians vary in size, such as adults (e.g., Ped 2) and children (e.g., Ped 3). Here, “tressider-2019-03-16.2 000030” represents the 30th frame (000030) of the tressider-2019-03-16.2 scene.

3D scenes remains limited, leading to suboptimal 3D layout estimation in practical applications.

B. LiDAR-based 3D Multi-Object Tracking Methods

Most 3D MOT methods [20], [21], [22] follow a workflow similar to 2D MOT, adhering to the “detection-association-state estimation” paradigm. The main difference lies in extending the bounding boxes from the image plane into three-dimensional space, which necessitates more precise modeling and management of objects’ 3D dimensions, orientations, and motion states. For example, SimpleTrack [23] emphasizes a unified analysis and improvement of association, motion models, and lifecycle management. Meanwhile, Polymot [24] adopted different motion models and similarity metrics for various object categories in data association and state estimation, reducing missed associations via a “polyhedral” strategy. Lastly, SMAT [25] proposed a self-reinforcing framework that deeply couples dynamic object detection with static mapping in large-scale open environments for simultaneous localization, mapping, and tracking. In summary, although 3D MOT and 2D MOT share the same fundamental components, the more abundant geometric and motion information in three-dimensional scenarios presents new challenges and opportunities for data association and trajectory management.

In addition, several methods that fuse camera and LiDAR data for multi-object tracking [26] have also achieved promising performance, such as CAMO-MOT [27], JRMOT [8], and DeepFusionMOT [28].

III. METHOD

Our pedestrian tracking method is an online approach primarily composed of four components: a 3D pedestrian detection module, a state prediction module, a data association module, and a state update module.

A. 3D Detection

As illustrated in Fig. 2, to achieve better tracking performance, we employed a high-quality pre-trained DCCLA network [29] as our 3D detection module, with the DCCLA

method demonstrating SOTA performance on the 3D object detection task on the JRDB dataset. At time t , the 3D detection module outputs a set of detection results $P = \{P^1, P^2, \dots, P^n\}$, where n denotes the number of detected pedestrians. Since the number of detected pedestrians varies at different times, the value of n will also change accordingly. At frame t , each pedestrian detection result can be represented as a tuple $P_t^i = (x, y, z, l, w, h, \theta, s)$, where (x, y, z) indicates the 3D coordinates of the pedestrian’s center, (l, w, h) represents the dimensions of the pedestrian, θ denotes the heading angle, and s denotes the detection confidence.

B. State Prediction

Our adopts the Constant Acceleration (CA) model and Prediction Confidence to enhance the robustness and accuracy of 3D MOT. Unlike the traditional Constant Velocity (CV) model, which only considers velocity, the CA model further incorporates acceleration, allowing for more precise trajectory prediction when an object accelerates or decelerates. In contrast, the CV model tends to produce larger prediction errors when the object suddenly changes speed or experiences temporary occlusion. The CA model, however, adapts better to dynamic changes in motion by predicting movement using the following equations:

$$\begin{aligned} p_t &= p_{t-1} + v_{t-1} \cdot \delta + \frac{1}{2} a_{t-1} \cdot \delta^2, \\ v_t &= v_{t-1} + a_{t-1} \cdot \delta, \end{aligned} \quad (1)$$

where p_t , v_t , a_t , and δ represent the position, velocity, acceleration of the object at time t , and the time interval between consecutive frames, respectively. Furthermore, to enhance the reliability of predictions, we propose Prediction Confidence, which assesses the uncertainty of the predicted results. The confidence level decays over time according to the following equation: $\gamma_t = \gamma_{t-1} - \eta \gamma_{t-1}$, where γ_t represents the prediction confidence of the current frame and η is the decay factor, which controls the rate at which confidence decreases.

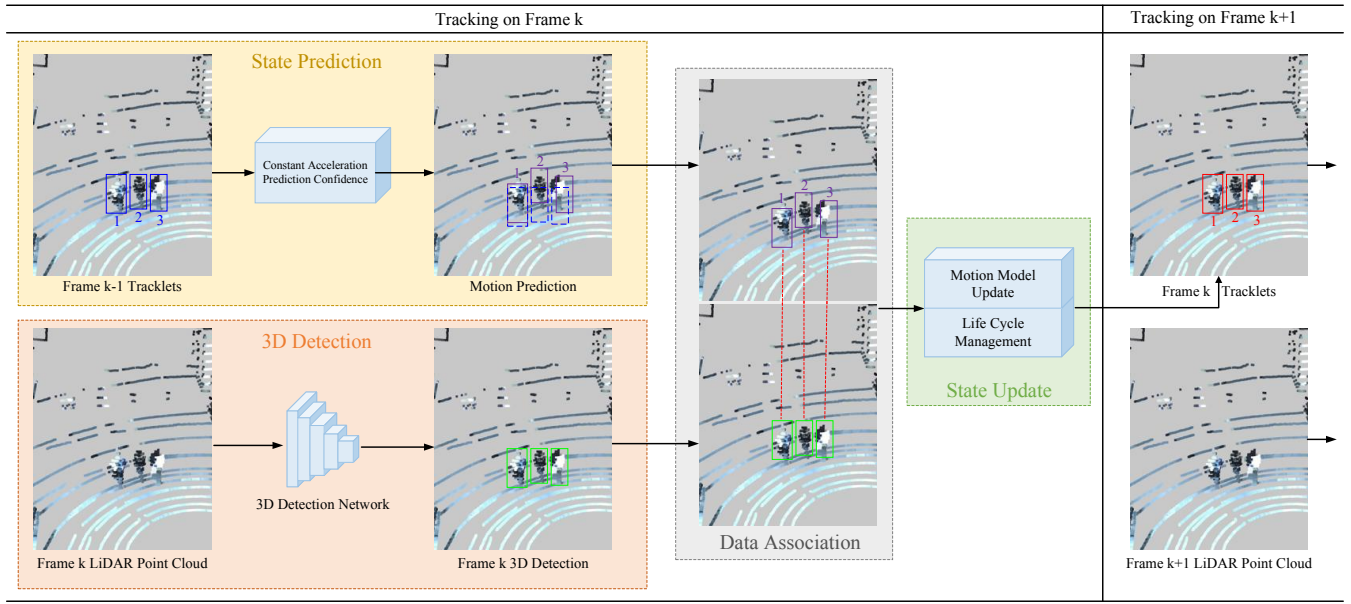


Fig. 2. Our ELPTNet pipeline. Our ELPTNet is composed of four key components: the 3D detection module, which extracts 3D bounding boxes from the LiDAR point cloud; the state prediction module, utilizing a Constant Acceleration (CA) model and prediction confidence to estimate object states in the current frame; the data association module, responsible for linking detections with predicted trajectories; and the state update module, which refines trajectories while handling newly emerging and vanishing objects.

C. Data Association

In 3D MOT tasks, objects may become temporarily undetectable due to occlusion or long distances, leading to tracking failures. Traditional constant velocity (CV) models struggle with abrupt motion changes, while deep learning-based models (e.g., LSTMs), despite their predictive accuracy, are computationally expensive and impractical for real-time applications. To address these challenges, this paper proposes a prediction confidence-guided data association method, which introduces a prediction confidence metric γ to quantify the reliability of predicted states and dynamically adjust the search range in data association. Specifically, our method utilizes a constant acceleration (CA) motion model for state prediction:

$$\hat{Z}_t^j = A\hat{Z}_{t-1}^j; \hat{S}_t^j = A\hat{S}_{t-1}^j A^T + Q, \quad (2)$$

where A is the state transition matrix, $\hat{S}_t$ represents the predicted error covariance, and Q is the process noise covariance. The position and velocity predictions are computed as follows:

$$\begin{aligned} \hat{p}_t^j &= p_{t-1}^j + \delta v_{t-1}^j + \frac{\delta^2 a_{t-1}^j}{2}, \\ \hat{v}_t^j &= v_{t-1}^j + \delta a_{t-1}^j. \end{aligned} \quad (3)$$

To ensure robust data association, the prediction confidence γ decreases over time in the absence of detection and is updated using:

$$\gamma_t^j = \gamma_{t-1}^j - \eta \gamma_{t-1}^j; \gamma_t^j = \gamma_t^j + \eta c_t^j, \quad (4)$$

where c_t^j represents the detection confidence, and η is a parameter that controls the rate of confidence decay. The

data association process is performed by computing an aggregated pairwise cost (C), which integrates geometric (G), appearance (A), and motion (M) features:

$$C^i, j_t = G^i, j_t + A^i, j_t + M^i, j_t. \quad (5)$$

The geometric cost (G) evaluates differences in bounding box dimensions, centroids, and orientations, while the appearance cost (A) is derived from deep features extracted by the detector and computed using the Euclidean distance. Additionally, the motion cost (M) measures the angular and Euclidean distances between velocity vectors. Ultimately, data association is performed using a cost matrix ψ_t weighted by the prediction confidence.

D. State Update

In the target state update process, we employ Kalman Filtering to optimize the state of matched targets. First, the Kalman gain is computed to balance prediction and detection errors, updating the target state and its error covariance matrix to enhance tracking stability. Meanwhile, a Prediction Confidence mechanism is introduced to adaptively adjust the search range for data association based on detection confidence, ensuring reasonable tracking of temporarily lost targets. For unmatched detection targets, a zero-initialization strategy is adopted, adding them to the tracking state set with initial velocity and acceleration set to zero, allowing new targets to be swiftly incorporated into the tracking sequence. Ultimately, the updated target state consists of both matched targets and newly initialized targets, serving as the foundation for tracking in the next frame. By smoothing target trajectories, optimizing data association, and enhancing

TABLE I

COMPARISON OF TRACKING PERFORMANCE BETWEEN OUR ELPTNET AND OTHER STATE-OF-THE-ART (SOTA) METHODS ON THE JRDB TEST SET. OUR ELPTNET ACHIEVES A NEW SOTA PERFORMANCE, IMPROVING HOTA BY 6.6% COMPARED TO THE PREVIOUS SOTA METHOD [25]. Δ REPRESENTS THE ABSOLUTE VALUE OF THE DIFFERENCE BETWEEN OUR ELPTNET AND SMAT.

Method	HOTA ↑ (%)	DetA ↑ (%)	DetRe ↑ (%)	DetPr ↑ (%)	AssA ↑ (%)	AssRe ↑ (%)	AssPr ↑ (%)	IDF1 ↑ (%)	IDR ↑ (%)	IDP ↑ (%)	OSPA ↓	OSPA _{CARD} ↓	OSPA _{LOC} ↓
AB3DMOT [12]	11.812	12.069	12.557	55.234	12.053	12.791	68.369	10.670	6.548	28.803	0.971	0.734	0.237
JRMOT [8]	13.053	12.600	13.198	53.735	13.956	14.940	63.940	12.319	7.672	31.235	0.956	0.627	0.329
PC-DAN [30]	15.569	23.687	25.863	70.013	10.376	10.774	70.629	13.889	9.510	25.743	0.976	0.898	0.079
SS3D-MOT [31]	15.800	23.224	25.196	70.966	10.889	11.184	84.519	14.480	9.811	27.632	0.982	0.911	0.071
SMAT [25]	18.930	17.055	18.417	55.310	21.132	22.342	71.842	21.342	14.224	42.719	0.975	0.693	0.282
ELPTNet (Ours)	25.538	25.991	28.583	64.589	25.429	30.560	53.023	28.686	20.690	46.753	0.783	0.230	0.553
Δ	6.608	8.936	10.166	9.279	4.297	8.218	18.819	7.344	6.466	4.034	0.192	0.463	0.271

TABLE II

COMPARISON OF TRACKING PERFORMANCE BETWEEN OUR ELPTNET AND OTHER SOTA METHODS ON THE JRDB TEST SET. OUR ELPTNET ACHIEVES THE TOP RANKING, IMPROVING MOTA BY 8.4% COMPARED TO THE SECOND-BEST METHOD [25].

Method	MOTA ↑ (%)	MODA ↑ (%)	CLR_Re ↑ (%)	CLR_Pr ↑ (%)	CLR_TP ↑	CLR_FN ↓	CLR_FP ↓	MTR ↑ (%)	PTR ↓	IDSW ↓	MT ↑	PT ↓	ML ↓
AB3DMOT [12]	19.342	19.967	21.350	93.914	211186	777968	13686	4.829	30.264	6179	86	539	1156
JRMOT [8]	20.152	20.578	22.570	91.890	223253	765901	19705	6.064	31.050	4216	108	553	1120
PC-DAN [30]	22.565	25.196	31.068	84.104	307308	681846	58084	4.717	45.929	26022	84	818	879
SS3D-MOT [31]	22.963	24.982	30.243	85.182	299153	690001	52041	4.211	45.536	19973	75	811	895
SMAT [25]	24.210	24.689	28.993	87.073	286790	702364	42577	4.885	43.627	4739	87	777	917
ELPTNet (Ours)	32.653	33.494	38.874	87.844	384524	604630	53212	16.844	37.226	8320	300	663	818
Δ	8.443	8.805	9.881	0.771	97734	97734	10635	11.959	6.401	3581	213	114	99

tracking robustness, this method significantly improves the accuracy and efficiency of 3D multi-object tracking.

IV. EXPERIMENT

The JRDB dataset [32], [8] and its evaluation metrics are first introduced. Next, the performance of ELPTNet is compared with other SOTA methods. Then, an ablation study is conducted to evaluate the performance of the method under various conditions. Finally, we present a visual analysis and discuss the robustness and limitations of the approach.

A. Dataset and Evaluation Metric

1) *Dataset*: The JRDB [32], [8] dataset was collected by the mobile robotic arm JackRabbit while navigating and interacting in complex human environments. JackRabbit is equipped with five cameras and two 16-channel LiDAR sensors, generating 3D point cloud data that mimics a 32-channel LiDAR. This dataset provides 64 minutes of sensor data and includes over 4.2 million pedestrian 3D bounding boxes. In total, it covers 54 indoor and outdoor scenes on a university campus, of which 27 unlabeled scenes are used for the test set and the remaining 27 are used for training and validation. Following the official partitioning scheme, we select 7 of these scenes as our validation set.

2) *Metric*: We used the following evaluation metrics: HOTA (Higher Order Tracking Accuracy), DetA (Detection Accuracy), DetRe (Detection Recall), DetPr (Detection Precision), AssA (Association Accuracy), AssRe (Association Recall), AssPr (Association Precision), IDF1, IDR (ID Recall), IDP (ID Precision), OSPA (Optimal Sub-Pattern As-

signment), OSPA_CARD (Cardinality), OSPA_LOC (Localization), MOTA (Multi-Object Tracking Accuracy), MODA (Multi-Object Detection Accuracy), CLR_Re (Clear Recall), CLR_Pr (Clear Precision), CLR_TP (Clear True Positives), CLR_FN (Clear False Negatives), CLR_FP (Clear False Positives), MTR (Mostly Tracked), PTR (Partially Tracked), MT (Mostly Tracked Targets), PT (Partially Tracked Targets), ML (Missed Targets), IDSW (ID Switches).

B. Comparison with SOTA methods on the JRDB test set

As illustrated in Table I, our ELPTNet achieved the top ranking in the 3D pedestrian tracking task on the JRDB test set¹, outperforming other methods across multiple tracking metrics. In particular, our ELPTNet attains the highest HOTA score (25.538%), outperforming the previous best method SMAT by 6.608%. It also exhibits substantial gains in both detection and association metrics: DetA, DetRe, and DetPr are improved by up to 8.936%, 10.166%, and 9.279%, respectively, whereas AssA, AssRe, and AssPr see improvements of up to 4.297%, 8.218%, and 18.819%. Furthermore, our ELPTNet yields the best IDF1, IDR, and IDP values, underscoring its effectiveness in consistently identifying tracked objects over time. In terms of overall tracking quality measured by OSPA, our ELPTNet reduces the error to 0.783, offering a notable improvement of 0.192 compared to SMAT. Similar advantages are observed in OSPA_CARD and OSPA_LOC, demonstrating that our ELPTNet not only

¹JRDB 3D Tracking: <https://jrdb.erc.monash.edu/leaderboards/tracking>

TABLE III

THE TRACKING PERFORMANCE OF OUR ELPTNET AND SMAT [25] ON THE JRDB TEST SET IS COMPARED ACROSS DIFFERENT SCENARIOS. OUR ELPTNET OUTPERFORMS THE SMAT METHOD IN MOST SCENARIOS.

Sequence	ELPTNet (Ours)							SMAT [25]						
	HOTA ↑ (%)	DetA ↑ (%)	AssA ↑ (%)	IDF1 ↑ (%)	OSPA ↓	MOTA ↑ (%)	MODA ↑ (%)	HOTA ↑ (%)	DetA ↑ (%)	AssA ↑ (%)	IDF1 ↑ (%)	OSPA ↓	MOTA ↑ (%)	MODA ↑ (%)
cubberly-auditorium-2019-04-22.1	40.06	41.45	39.70	47.44	0.69	58.59	59.13	23.18	20.80	26.06	25.41	0.98	35.55	35.98
discovery-walk-2019-02-28.0	36.14	37.95	34.80	43.23	0.84	49.76	50.28	21.12	15.83	30.83	25.70	0.96	26.27	26.59
discovery-walk-2019-02-28.1	37.28	36.67	38.15	48.56	0.81	50.07	50.36	20.49	18.62	22.64	23.56	0.99	32.02	32.36
food-trucks-2019-02-12.0	27.38	38.83	19.55	29.99	0.90	51.82	52.62	17.74	20.53	15.45	21.52	0.98	33.30	33.68
gates-ai-lab-2019-04-17.0	25.55	34.44	19.06	26.62	0.93	37.08	37.78	23.96	23.20	24.75	29.76	0.99	21.64	21.97
gates-basement-elevators-2019-01-17.0	38.58	46.78	31.99	47.01	0.68	58.85	59.93	26.33	24.07	30.44	27.03	0.96	40.04	40.49
gates-foyer-2019-01-17.0	44.36	49.29	39.98	55.90	0.73	63.17	63.37	25.62	31.10	21.17	35.37	0.99	44.88	45.27
gates-to-clark-2019-02-28.0	51.02	46.39	56.18	64.55	0.59	57.57	57.69	31.36	27.18	36.36	48.97	0.96	42.71	43.06
hewlett-class-2019-01-23.0	34.31	44.83	26.63	39.34	0.76	55.46	57.39	27.37	28.58	26.59	31.69	0.97	49.01	49.56
hewlett-class-2019-01-23.1	56.40	57.18	55.69	73.99	0.48	74.06	74.25	35.21	33.97	36.76	38.44	0.97	64.73	65.10
huang-2-2019-01-25.1	43.83	39.06	49.62	54.09	0.73	54.12	54.37	23.35	18.19	29.98	24.78	0.98	22.87	23.24
huang-intersection-2019-01-22.0	30.71	27.23	34.72	40.71	0.83	36.39	36.73	18.21	12.97	27.06	18.77	0.98	23.15	23.43
indoor-coupa-cafe-2019-02-06.0	18.12	22.86	14.45	19.91	0.92	25.66	26.19	11.79	14.42	9.64	15.37	0.99	15.80	16.28
lomita-serra-intersection-2019-01-30.0	38.95	33.49	45.97	51.44	0.76	47.82	47.86	28.58	18.44	46.43	33.48	0.97	25.53	25.86
meyer-green-2019-03-16.1	29.81	19.24	46.56	31.16	0.87	22.64	22.85	17.79	12.13	26.19	12.67	0.99	17.06	17.52
nvidia-aud-2019-01-25.0	25.16	25.26	25.58	30.98	0.81	30.97	31.52	19.55	15.56	24.57	20.69	0.99	19.46	19.93
nvidia-aud-2019-04-18.1	43.49	39.27	48.48	50.98	0.71	49.86	50.11	23.35	16.85	32.40	26.69	0.93	26.91	27.09
nvidia-aud-2019-04-18.2	38.77	35.26	42.90	48.89	0.76	45.59	45.98	28.28	19.91	42.78	30.22	0.97	27.26	27.71
outdoor-coupa-cafe-2019-02-06.0	17.39	17.35	17.71	19.15	0.93	20.28	20.61	13.22	12.29	14.30	14.37	0.99	16.54	16.98
quarry-road-2019-02-28.0	39.53	30.99	50.63	50.73	0.76	37.27	37.41	21.88	14.13	33.97	23.42	0.98	20.00	20.28
serra-street-2019-01-30.0	29.91	20.41	43.91	34.39	0.82	27.85	28.03	17.00	10.03	29.14	15.92	0.99	16.71	17.12
stlc-111-2019-04-19.1	54.25	57.04	51.75	67.32	0.57	70.57	70.98	32.74	32.14	33.55	46.20	0.95	54.13	54.57
stlc-111-2019-04-19.2	45.06	52.14	39.05	57.66	0.68	64.34	64.81	30.71	30.61	31.01	45.22	0.94	52.39	52.86
tressider-2019-03-16.2	25.27	23.49	27.79	31.81	0.86	30.22	30.40	18.79	12.82	28.40	21.29	0.95	20.17	20.38
tressider-2019-04-26.0	16.53	17.26	16.09	20.24	0.91	20.00	21.22	14.20	14.89	13.55	19.63	0.99	19.95	20.60
tressider-2019-04-26.1	17.83	22.51	14.75	19.71	0.91	29.58	30.89	18.82	18.39	19.28	21.05	0.99	25.00	25.48
tressider-2019-04-26.3	17.77	21.37	15.15	19.02	0.89	26.36	27.67	16.97	17.45	16.64	19.38	0.99	24.24	24.92
COMBINED	25.54	25.99	25.43	28.69	0.78	32.65	33.49	18.93	17.06	21.13	21.34	0.98	24.21	24.69

tracks objects more accurately but also achieves more precise localization and robust object cardinality estimates.

As illustrated in Table II, our ELPTNet significantly outperforms existing methods, achieving the highest Multi-Object Tracking Accuracy (MOTA) of 32.653%, which is 8.4% higher than the second-best method, SMAT [25]. In addition, our ELPTNet attains the highest Multi-Object Detection Accuracy (MODA) of 33.494%, an 8.8% improvement over SMAT. Regarding tracking accuracy, our ELPTNet achieves the highest Correct Localization Recall (CLR_Re) of 38.874%, surpassing SMAT by 9.88%, while maintaining a competitive Correct Localization Precision (CLR_Pr) of 87.844%. Our ELPTNet also reports the highest number of CLR_TP (384,524) and the lowest number of CLR_FN (604,630) demonstrating its remarkable ability to accurately detect and track objects while minimizing missed detections. In terms of trajectory consistency metrics, our ELPTNet significantly increases the Mostly Tracked (MTR) percentage to 16.844%, outperforming all other methods. It also achieves a lower Partially Tracked (PTR) percentage of 37.226% and the fewest Mostly Lost (ML) trajectories at 818, indicating strong tracking stability. However, our ELPTNet has a slightly higher ID Switch (IDSW) count of 8,320, suggesting that identity consistency can still be improved. Overall, our ELPTNet substantially advances multi-object tracking performance.

As illustrated in Table III, our ELPTNet consistently outperforms SMAT across various JRDB test sequences in key tracking metrics, including HOTA, DetA, AssA, IDF1, MOTA, and MODA, while maintaining a favorable OSPA value. Notably, the sequences “hewlett-class-2019-01-23.1” and “stlc-111-2019-04-19.1” display substantial improvements in both detection accuracy (DetA) and association accuracy (AssA), underscoring the robustness of our ELPTNet in diverse and challenging scenarios. Overall, these results highlight the effectiveness of our ELPTNet, achieving excellent tracking performance across a wide range of environments.

C. Ablation Study

We validated the impact of various components of our method on the JRDB validation set.

1) *Effect of the number of new objects:* As illustrated in Table IV, increasing the number of new objects generally leads to improvements in both HOTA and MOTA. Introducing 15 new objects yields the highest HOTA 21.894% and MOTA 23.690%, as well as the lowest IDSW 4341, indicating superior overall tracking accuracy and stability. Although MODA 25.525% and IDF1 23.107% achieve their respective peaks when introducing 13 and 11 new objects, respectively, the overall performance is best optimized with 15 new objects. Therefore, selecting 15 new objects is

TABLE IV

THE EFFECT OF THE NUMBER OF NEW OBJECTS. WHEN THE NUMBER OF NEW OBJECTS IS 15, HOTA, MOTA, AND IDSW ACHIEVE OPTIMAL PERFORMANCE.

New object	HOTA ↑ (%)	MOTA ↑ (%)	MODA ↑ (%)	IDF1 ↑ (%)	IDSW ↓
3	21.635	23.646	25.522	22.806	4442
5	21.684	23.676	25.520	22.744	4365
7	21.849	23.643	25.523	22.982	4451
9	21.615	23.645	25.514	22.672	4426
11	21.755	23.656	25.516	23.107	4405
13	21.867	23.675	25.525	22.834	4380
15	21.894	23.690	25.523	23.031	4341

TABLE V

THE EFFECT OF THE MAXIMUM PREDICTED NEW OBJECTS COUNT. WHEN SET TO 25, IT ACHIEVES A BALANCED PERFORMANCE ACROSS MULTIPLE EVALUATION METRICS.

Max pred	HOTA ↑ (%)	MOTA ↑ (%)	MODA ↑ (%)	IDF1 ↑ (%)	IDSW ↓
5	21.708	23.623	25.529	22.654	4514
10	21.760	23.686	25.519	23.107	4340
15	21.644	23.640	25.526	22.862	4466
25	22.030	23.658	25.518	23.432	4404
50	21.957	23.649	25.525	23.230	4442
75	21.946	23.645	25.515	23.027	4429
100	21.975	23.673	25.521	23.040	4376

recommended to achieve better performance across multiple evaluation metrics.

2) *Effect of maximum predicted new objects count:* As illustrated in Table V, it can be observed that different settings for the maximum number of newly predicted objects (Max pred) produce varied effects on each metric. When Max pred is set to 25, HOTA reaches 22.030% and IDF1 reaches 23.432%, both at their highest levels. MOTA peaks at 23.686% with Max pred = 10. MODA achieves 25.529% when Max pred is set to 5, while IDSW is minimized at 4340 with Max pred = 10. These findings suggest that selecting the maximum number of newly predicted objects should balance HOTA, MOTA, MODA, IDF1, and IDSW. If HOTA and IDF1 are priorities, setting Max pred to 25 is advisable. If the focus is on MOTA or MODA, reducing Max pred to 10 or 5 may be more appropriate. It is also crucial to track changes in IDSW to minimize identity switches during the tracking process.

3) *Effect of post score:* As illustrated in Fig. 3, adjusting the post score threshold from 0.1 to 0.9 significantly influences tracking performance. At a threshold of 0.1, HOTA is 17.880% and IDF1 is 16.696%, indicating a high number of false positives. Increasing the threshold to 0.5 improves HOTA to 20.949% and IDF1 to 21.310%. At 0.7, HOTA rises to 21.635% and IDF1 to 22.806%, representing the best overall performance. Setting the threshold as high as 0.9, however, decreases HOTA to 14.080% and IDF1 to 15.692%, implying that many valid detections are lost. Therefore, when the post score is 0.7, both HOTA and IDF1 reach their optimal values.

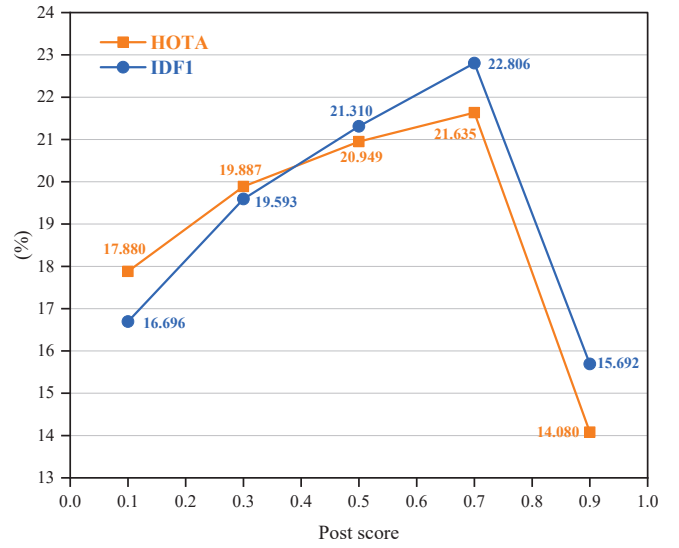


Fig. 3. The effect of the post score. When the post score is 0.7, both HOTA and IDF1 reach their highest values.

TABLE VI

THE EFFECT OF TRACKING DELAY TIME. WHEN THE LATENCY IS 0, IT INDICATES REAL-TIME ONLINE TRACKING. WHEN THE LATENCY IS 1, MEANING ONLINE TRACKING WITH A 1-SECOND DELAY, THE OVERALL TRACKING PERFORMANCE IS BETTER. THE FPS WAS TESTED ON THE INTEL CORE I9 10900K CPU.

Latency	HOTA ↑ (%)	MOTA ↑ (%)	MODA ↑ (%)	IDF1 ↑ (%)	IDSW ↓	FPS ↑
0	21.635	23.646	25.522	22.806	4442	61.132
1	22.114	27.324	28.187	23.926	2045	62.378
2	22.067	26.663	27.500	24.035	1982	63.681
3	21.961	25.702	26.558	24.008	2028	64.842

4) *Effect of tracking delay time:* As illustrated in Table VI, when the latency is 0 (real-time online tracking), the HOTA, MOTA, MODA, and IDF1 scores reach 21.635%, 23.646%, 25.522%, and 22.806%, respectively. More importantly, the frame rate achieves 61.132 FPS. As the latency increases, overall performance gradually improves. Notably, under a 1-second latency, MOTA and MODA exhibit the most significant improvements, while the IDSW count markedly decreases, and the FPS slightly increases to 62.378, resulting in the best overall tracking performance. Further increases in latency lead to continued improvements in certain metrics (e.g., IDF1 and FPS), but also cause slight declines in accuracy metrics such as MOTA and MODA, indicating that additional latency does not confer an overall benefit. In summary, our approach achieves the most balanced trade-off between computational demands and tracking accuracy, making it particularly suitable for real-time applications, including autonomous social navigation robots.

5) *Robustness analysis:* As illustrated in Table VII, we assessed the robustness of our multi-object tracking method in various scenarios. Despite significant differences among

TABLE VII

TRACKING PERFORMANCE UNDER DIFFERENT SCENARIOS. NOTE THAT OUR APPROACH DEMONSTRATES STABLE TRACKING PERFORMANCE ACROSS DIVERSE ENVIRONMENTS.

Sequence	HOTA ↑ (%)	DetA ↑ (%)	AssA ↑ (%)	MOTA ↑ (%)	CLR_Re ↑ (%)	CLR_Pr ↑ (%)	MODA ↑ (%)	PTR ↓	MTR ↑ (%)	IDSW ↓	IDF1 ↑ (%)	OSPA ↓
clark-center-2019-02-28_1	20.299	24.008	17.433	17.741	40.469	66.596	20.170	67.480	6.504	1453	22.001	0.979
gates-ai-lab-2019-02-08_0	27.767	27.306	29.251	2.318	59.243	51.518	3.491	56.000	28.000	123	26.815	0.968
huang-2-2019-01-25_0	30.359	31.906	29.020	30.733	53.333	71.253	31.816	63.415	19.512	107	37.352	0.952
meyer-green-2019-03-16_0	34.224	29.527	40.013	25.508	52.999	66.775	26.629	73.333	13.333	65	40.607	0.982
nvidia-aud-2019-04-18_0	30.335	33.510	27.525	38.761	56.122	77.391	39.726	62.500	22.500	187	38.200	0.959
tressider-2019-03-16_1	25.724	18.313	37.020	18.893	27.871	77.155	19.619	36.842	7.895	116	24.664	0.926
tressider-2019-04-26_2	16.755	27.049	10.704	26.065	46.202	71.888	28.135	62.802	10.145	2391	17.511	0.966
COMBINED	21.635	26.535	18.056	23.646	45.369	69.567	25.522	61.963	11.861	4442	22.806	0.962

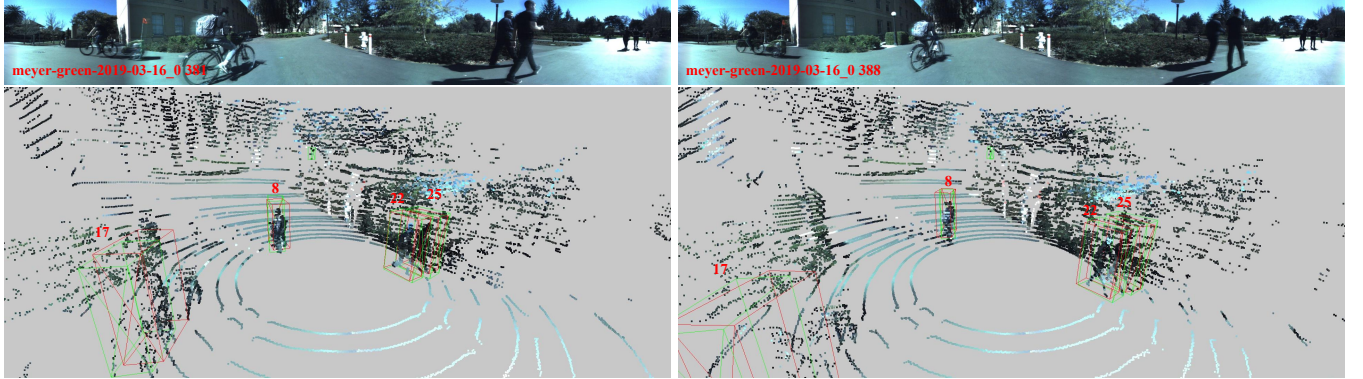


Fig. 4. Visualization of meyer-green-2019-03-16.0 scene. Our ELPTNet can effectively track pedestrians. The green and red 3D bounding boxes represent the ground truth and our predicted results, respectively.

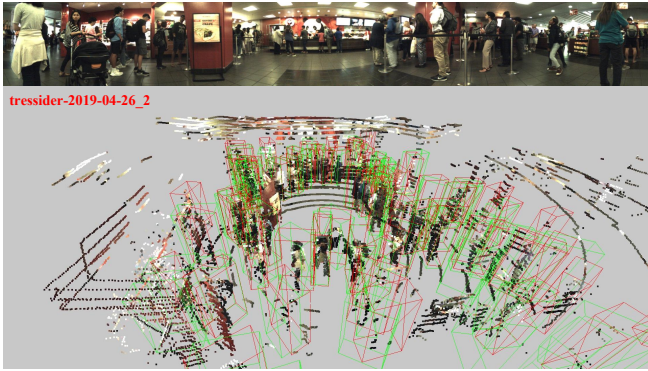


Fig. 5. Visualization of the tressider-2019-04-26.2 scenario. This indoor scene features high crowd density along with frequent mutual occlusions and interactions, which result in reduced tracking performance.

the scenes in terms of environmental layout, crowd density, and viewing angles, our ELPTNet demonstrated satisfactory performance in metrics such as HOTA, DetA, AssA, MOTA, MODA, and IDF1. Notably, as illustrated in Fig. 4, in the meyer-green-2019-03-16.0 scenario, due to dispersed individuals, fewer occlusions, and more stable lighting conditions, the algorithm achieved the best overall performance, with a HOTA of 34.224%, an IDF1 of 40.607%, and only 65 IDSW. Other scenarios also showed stable tracking performance. These results fully demonstrate the method's adaptability and robustness in various complex environments.

D. Limitations

As illustrated in Fig. 5, indoor scenes (e.g., tressider-2019-04-26.2) typically have high crowd density and frequent occlusions and interactions, which significantly increase the difficulty of tracking. Consequently, this scene exhibits the highest number of ID switches (2391) in Table VII, while the overall HOTA score is only 16.755%. These results indicate that although the proposed method demonstrates a certain level of robustness across various indoor and outdoor scenarios, it still encounters a large number of identity switches and tracking discontinuities in the presence of extensive occlusions and dense crowds. Therefore, further improvements and optimizations are necessary to address these challenges in complex indoor environments.

V. CONCLUSIONS

This paper presents a simple yet Efficient LiDAR-based 3D Pedestrian Tracking Network (ELPTNet) for autonomous navigation social robots. Our ELPTNet first utilizes a 3D object detector to obtain pedestrian 3D bounding boxes from LiDAR point clouds. In addition, our ELPTNet employs a Constant Acceleration (CA) model and prediction confidence for target trajectory prediction. During the data association process, it integrates geometric, appearance, and motion features to enhance the robustness and real-time performance of 3D MOT when targets are temporarily occluded. Experimental results on the large-scale JRDB pedestrian dataset

demonstrate that our method ranks first in the JRDB 3D pedestrian tracking task, achieving an 8.4% improvement in MOTA and a 6.6% increase in HOTA compared to previous state-of-the-art (SOTA) methods. Additionally, our ELPTNet achieves an inference speed of 61 Frames Per Second (FPS) on a single CPU, enabling accurate and real-time multi-pedestrian tracking. Since LiDAR point cloud data is sparse and less accurate for tracking distant or small objects, whereas camera images excel at capturing color and fine details, future work will focus on integrating LiDAR and camera sensor data to improve tracking accuracy and robustness in complex environments for intelligent robots.

REFERENCES

- [1] S. M. Marvasti-Zadeh, L. Cheng, H. Ghanei-Yakhdan, and S. Kasaei, "Deep learning for visual tracking: A comprehensive survey," *IEEE Transactions on Intelligent Transportation Systems*, vol. 23, no. 5, pp. 3943–3968, 2021.
- [2] Z. Cao, Q. Zhang, J. Guang, S. Wu, Z. Hu, and J. Liu, "Semantictopology: Semantic loop closure with 3d topological graph based on quadric-level object map," *IEEE Robotics and Automation Letters*, vol. 9, no. 5, pp. 4257–4264, 2024.
- [3] J. Guang and J. Liang, "Cmse: Compound model scaling with efficient attention for fine-grained image classification," *IEEE Access*, vol. 10, pp. 18 222–18 232, 2022.
- [4] M. Lyssenko, C. Gladisch, C. Heinzemann, M. Woehrle, and R. Triebel, "Towards safety-aware pedestrian detection in autonomous systems," in *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2022, pp. 293–300.
- [5] Y. Zeng, C. Ma, M. Zhu, Z. Fan, and X. Yang, "Cross-modal 3d object detection and tracking for auto-driving," in *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2021, pp. 3850–3857.
- [6] J. Guang, Z. Hu, S. Wu, Q. Zhang, and J. Liu, "Rpea: A residual path network with efficient attention for 3d pedestrian detection from lidar point clouds," *Expert Systems with Applications*, vol. 249, p. 123497, 2024.
- [7] D. Jia, A. Hermans, and B. Leibe, "2d vs. 3d lidar-based person detection on mobile robots," in *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2022, pp. 3604–3611.
- [8] A. Sheno, M. Patel, J. Gwak, P. Goebel, A. Sadeghian, H. Rezatofighi, R. Martin-Martin, and S. Savarese, "Jrmot: A real-time 3d multi-object tracker and a new large-scale dataset," in *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2020, pp. 10 335–10 342.
- [9] Z. Wang, Z. Li, J. Leng, M. Li, and L. Bai, "Multiple pedestrian tracking with graph attention map on urban road scene," *IEEE Transactions on Intelligent Transportation Systems*, vol. 24, no. 8, pp. 8567–8579, 2022.
- [10] Y. Yang, B. Fan, Y. Jiang, W. Zhou, D. Chen, and H. Xu, "Enhancing 3d single object tracking with efficient point cloud segmentation," in *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2024, pp. 11 631–11 638.
- [11] J. Shan, S. Zhou, Z. Fang, and Y. Cui, "Ptt: Point-track-transformer module for 3d single object tracking in point clouds," in *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2021, pp. 1310–1316.
- [12] X. Weng, J. Wang, D. Held, and K. Kitani, "3d multi-object tracking: A baseline and new evaluation metrics," in *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2020, pp. 10 359–10 366.
- [13] J. Wu, J. Cao, L. Song, Y. Wang, M. Yang, and J. Yuan, "Track to detect and segment: An online multi-object tracker," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 12 352–12 361.
- [14] P. Lenz, A. Geiger, and R. Urtasun, "Followme: Efficient online min-cost flow tracking with bounded memory and computation," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 4364–4372.
- [15] V. Chari, S. Lacoste-Julien, I. Laptev, and J. Sivic, "On pairwise costs for network flow multi-object tracking," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 5537–5545.
- [16] B. Lee, E. Erdenee, S. Jin, M. Y. Nam, Y. G. Jung, and P. K. Rhee, "Multi-class multi-object tracking using changing point detection," in *Computer Vision–ECCV 2016 Workshops: Amsterdam, The Netherlands, October 8–10 and 15–16, 2016, Proceedings, Part II 14*. Springer, 2016, pp. 68–83.
- [17] S. Sharma, J. A. Ansari, J. K. Murthy, and K. M. Krishna, "Beyond pixels: Leveraging geometry and shape cues for online multi-object tracking," in *2018 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2018, pp. 3508–3515.
- [18] H. Karunasekera, H. Wang, and H. Zhang, "Multiple object tracking with attention to appearance, structure, motion and size," *IEEE Access*, vol. 7, pp. 104 423–104 434, 2019.
- [19] K. Fang, Y. Xiang, X. Li, and S. Savarese, "Recurrent autoregressive networks for online multi-object tracking," in *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 2018, pp. 466–475.
- [20] J. Jin, J. Zhang, K. Zhang, Y. Wang, Y. Ma, and D. Pan, "3d multi-object tracking with boosting data association and improved trajectory management mechanism," *Signal Processing*, vol. 218, p. 109367, 2024.
- [21] H. Wu, W. Han, C. Wen, X. Li, and C. Wang, "3d multi-object tracking in point clouds based on prediction confidence-guided data association," *IEEE Transactions on Intelligent Transportation Systems*, vol. 23, no. 6, pp. 5668–5677, 2021.
- [22] Y. Ma, J. Zhang, G. Qin, J. Jin, K. Zhang, D. Pan, and M. Chen, "3d multi-object tracking based on dual-tracker and ds evidence theory," *IEEE Transactions on Intelligent Vehicles*, vol. 8, no. 3, pp. 2426–2436, 2022.
- [23] Z. Pang, Z. Li, and N. Wang, "Simpletrack: Understanding and rethinking 3d multi-object tracking," in *European Conference on Computer Vision*. Springer, 2022, pp. 680–696.
- [24] X. Li, T. Xie, D. Liu, J. Gao, K. Dai, Z. Jiang, L. Zhao, and K. Wang, "Poly-mot: A polyhedral framework for 3d multi-object tracking," in *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2023, pp. 9391–9398.
- [25] T. Fan, B. Shen, Y. Zhang, C. Zhang, L. Yang, H. Chen, W. Zhang, and J. Pan, "S2mat: Simultaneous and self-reinforced mapping and tracking in dynamic urban scenarios: A framework for simultaneous mapping and tracking in unbounded urban environments," *arXiv preprint arXiv:2304.14356*, 2023.
- [26] S. Xin, Z. Zhang, L. Liu, X. Hou, D. Zhu, M. Wang, and Y. Liu, "A robotic-centric paradigm for 3d human tracking under complex environments using multi-modal adaptation," in *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2024, pp. 4934–4940.
- [27] L. Wang, X. Zhang, W. Qin, X. Li, J. Gao, L. Yang, Z. Li, J. Li, L. Zhu, H. Wang *et al.*, "Camo-mot: Combined appearance-motion optimization for 3d multi-object tracking with camera-lidar fusion," *IEEE Transactions on Intelligent Transportation Systems*, vol. 24, no. 11, pp. 11 981–11 996, 2023.
- [28] X. Wang, C. Fu, Z. Li, Y. Lai, and J. He, "Deepfusionmot: A 3d multi-object tracking framework based on camera-lidar fusion with deep association," *IEEE Robotics and Automation Letters*, vol. 7, no. 3, pp. 8260–8267, 2022.
- [29] J. Guang, S. Wu, Z. Hu, Q. Zhang, P. Wu, and J. Liu, "Dccla: Dense cross connections with linear attention for lidar-based 3d pedestrian detection," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 35, no. 5, pp. 4535–4548, 2025.
- [30] A. Kumar, J. Kini, M. Shah, and A. Mian, "Pc-dan: Point cloud based deep affinity network for 3d multi-object tracking," *arXiv preprint arXiv:2106.07552*, 2021.
- [31] C. Liu, C. Gao, F. Liu, J. Liu, D. Meng, and X. Gao, "Ss3d: Sparsely-supervised 3d object detection from point cloud," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 8428–8437.
- [32] R. Martin-Martin, M. Patel, H. Rezatofighi, A. Sheno, J. Gwak, E. Frankel, A. Sadeghian, and S. Savarese, "Jrdb: A dataset and benchmark of egocentric robot visual perception of humans in built environments," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 6, pp. 6748–6765, 2023.