2026 IEEE International Conference on Robotics and Automation (ICRA 2026)
June 1-5, 2026. Vienna, Austria

Robust 3D Multi-Object Tracking for Autonomous Driving with Adaptive LiDAR-Visual Fusion and Multilevel Data Association

Chao Jiang, Chao Wang, Liang Nie, Mingyue Zhang, and Yuting Zhou

Abstract— **To increase the safety and reliability of autonomous driving systems in complex traffic environments, this paper proposes a novel 3D multiobject tracking (MOT) method that integrates center-plane adaptive multisensor fusion, motion compensation, and multilevel data association. Unlike traditional methods, our approach employs a center-plane adaptive fusion strategy to align LiDAR and visual data precisely, mitigating errors in the target width caused by pose variations, and improving tracking accuracy. To address vehicle motion-induced association errors in dynamic scenarios, we incorporate IMU and GPS data for high-frequency vehicle pose estimation and compensation, ensuring stable and robust target association. Additionally, a rotational geometric distance intersection-over-union (RGDIoU) cost function is introduced, combined with multilevel spatial indexing, to optimize the data association efficiency and accuracy. The experimental results on benchmark datasets, including KITTI and nuScenes, demonstrate that our method achieves state-of-the-art (SOTA) performance across multiple tracking metrics, including HOTA and sAMOTA, while maintaining real-time performance at 90 FPS. Specifically, our method improves sAMOTA tracking accuracy by 13% over the best existing methods and achieves a HOTA score of 50.24%, surpassing all compared methods.**

I. INTRODUCTION

Multiobject tracking, as a core technology for environmental perception, is widely applied in fields such as autonomous driving, intelligent transportation, surveillance systems, and robotics [1]. In autonomous driving systems, the primary task of MOT is to track multiple dynamic objects in the environment in real time and continuously update their position and state data to support critical functions such as path planning, collision warning, and decision-making control. By tracking surrounding vehicles, pedestrians, and obstacles in real time, the MOT system helps autonomous vehicles avoid potential collisions and make safe decisions, thus enhancing overall safety and robustness [2], [3].

Currently, the main methods of MOT include tracking-by-detection [4], [5], joint detection and tracking [6], [7], joint detection and embedding [8], [9], and tracking-by-attention [10]. Among these methods, tracking-by-detection methods are widely applied because of their simplicity and efficiency [4], [5]. The basic process of tracking-by-detection involves target detection and cross-frame data association to form continuous target trajectories. Although deep learning has improved the accuracy of target detection in recent years, existing methods still face challenges in complex scenarios, particularly in cases of target occlusion, target disappearance, interference from similar targets, and environmental changes. These issues lead to data association errors, as well as difficulties in target prediction.

One of the fundamental causes of these challenges is the limitations of single-sensor systems. While visual sensors provide rich texture and color information aiding object classification and recognition, they lack depth information [11]. LiDAR (light detection and ranging), on the other hand, captures high-precision three-dimensional spatial data, but its sparse and unordered point clouds lack semantic information, leading to poor object recognition and classification performance [12]. To overcome these limitations, this paper introduces multisensor fusion by combining target information from 2D images with point cloud features, thereby leveraging the strengths of different sensors to enhance 3D object detection and tracking performance.

The existing LiDAR-camera fusion methods can be categorized into three types: data-level, feature-level, and decision-level fusion [1]–[3], [5], [7], [9], [13], [14]. Data-level fusion enhances point clouds with image semantics but is constrained by calibration errors and adaptability to dynamic scenes [13], [14]. Feature-level fusion improves robustness through multi-modal feature collaboration, yet faces challenges such as alignment accuracy and computational complexity [1], [5], [7], [9]. Decision-level fusion independently generates detection proposals and performs cross-scale fusion, complementing the semantic information of images with the spatial localization advantage of LiDAR, thereby improving both accuracy and robustness [2], [3].

This paper proposes a method based on decision-level fusion, aimed at high-dynamic autonomous driving scenarios, to efficiently integrate LiDAR and camera data. The effectiveness of this method relies on the precise alignment of camera and LiDAR features [13]. Particularly when the target is not facing the camera directly, traditional methods often encounter misalignment due to inconsistent viewpoints, leading to shifts in the target's side view and its projected position in the camera's field of view. This results in "width expansion" errors. To address this issue, this paper introduces a center-plane adaptive fusion method to eliminate the width expansion errors caused by variations in the target pose.

This work was supported by the Hunan Provincial Natural Science Foundation of China under Grant 2025JJ60392, by the Research Foundation of the Education Bureau of Hunan Province under Grant 24B0406, by the Hunan Provincial Teaching Reform Research Project of Higher Education Institutions under Grant 202502000731, and by the Hunan Provincial "New Engineering, New Medicine, New Agriculture, and New Liberal Arts" Research and Practice Project under Grant 202503000064.

Chao Jiang, Chao Wang, and Liang Nie are with the School of Electrical Engineering, University of South China, Hengyang, China. Mingyue Zhang is with the University of South China, Hengyang, China. Yuting Zhou is with the University of South China, Hengyang, China, and also with the Hunan University of Science and Engineering, Yongzhou, China.

Corresponding author: Chao Jiang (e-mail: jc2009@mail.ustc.edu.cn).

In addition, vehicle motion in dynamic environments, especially rotational motion, can affect the relative position and motion state of targets, thereby impacting data association accuracy. Existing vehicle motion estimation methods [15], [16] can alleviate this problem, but their application in complex scenarios involving highly dynamic rotations remains limited. This paper further integrates IMU and GPS data and proposes a vehicle pose estimation and compensation method to mitigate the interference of vehicle motion in target tracking. Since data association depends on the relative pose, this method maintains strong adaptability even without GPS signals by relying on relative pose estimation.

To achieve robust target tracking, in addition to higher accuracy requirements for fusion alignment and motion compensation strategies, the tracking system must also maintain efficient and accurate data association in dynamic traffic environments with high target density and frequent occlusions [17]. However, as the number of targets and interframe motion uncertainty increase, traditional data association methods based on intersection over union (IoU) struggle to meet the dual requirements of robustness and real-time performance, particularly in scenarios involving target rotation or fast movement, where matching errors or ID switches are prone to occur [5], [18], [19]. While association methods based on complex cost functions, such as appearance similarity, can improve matching robustness, they increase the computational load and degrade real-time performance [8]. In light of these challenges, this paper designs an efficient 3D MOT method that combines adaptive fusion strategies, precise motion compensation, and multilevel data association mechanisms, improving accuracy while maintaining computational efficiency and robustness in complex traffic scenarios. The main contributions of this paper are as follows:

- An improved LiDAR and visual data fusion strategy is proposed based on a center-plane adaptive fusion method, which overcomes the limitations of single-sensor systems in depth information and target classification and eliminates the width expansion errors caused by target pose variations in traditional fusion methods.
- A motion compensation mechanism based on an IMU and a GPS is introduced to estimate the vehicle's pose and displacement accurately, thereby mitigating the effects of vehicle motion on target tracking.
- A cost function based on the RGDIoU is proposed, which combines multilevel spatial indexing and fine-grained retrieval to increase the efficiency and accuracy of data association.

The remainder of this paper is organized as follows. Section 2 reviews related work in multisensor fusion–based multi-object tracking. Section 3 presents the proposed method. Section 4 reports the experimental results. Section 5 concludes the paper and discusses future work.

II. Background

A. Multisensor Fusion

Multisensor fusion technology can compensate for the limitations of individual sensors and improve the performance of multiobject tracking (MOT) systems. Current mainstream multimodal fusion methods can be categorized into data-level, feature-level, and decision-level fusion methods. Data-level fusion methods [14], [15] align raw sensor data to generate multimodal inputs, making full use of the original information. However, in practical applications, these methods face challenges such as large data volumes and high computational resource consumption, which make it difficult for them to meet real-time requirements. Feature-level fusion [1], [5], [7], [9] combines features from multiple modalities through joint encoding within a network to increase the robustness of perception systems. However, this approach has high computational complexity and requires high-quality and consistent data annotations. Moreover, it still faces limitations in adapting to dynamic scenarios and heterogeneous sensors. Decision-level fusion [2], [3] integrates information after independently performing detection or tracking tasks, offering high flexibility and fault tolerance. Although these methods improve the performance of multisensor fusion, achieving more efficient fusion in complex environments and addressing challenges related to real-time performance and robustness remain pressing issues. The multisensor fusion framework proposed in this paper offers advantages in addressing these problems, as it enhances robustness by eliminating errors caused by target pose variations through a center-plane adaptive fusion method.

B. Motion Compensation

In dynamic environments, vehicle motion causes global background shifts, which in turn interfere with target prediction and data association in multiobject tracking. To address this, existing methods mainly use three mechanisms for compensation: first, Kalman filtering combined with appearance features to enhance occlusion robustness [15]; second, explicit incorporation of IMU or visual odometry information to optimize state prediction [16]; and third, coordinate alignment on the basis of global motion modeling from the camera [20]. Although these methods have improved tracking continuity and matching accuracy to some extent, most still face challenges in real-time performance and adaptability in highly dynamic or resource-constrained scenarios. This paper proposes a lightweight vehicle pose estimation and compensation method that combines an IMU and a GPS, which ensures real-time performance while maintaining tracking robustness in highly dynamic scenarios.

C. Prediction and Association

Detection-based MOT methods are widely used in autonomous driving. These methods associate predicted trajectories with detection results via trajectory prediction and cost functions. Common trajectory prediction methods include history-based motion feature prediction using Long Short-Term Memory (LSTM) networks [8] and prediction methods based on kinematics and filtering models [4], [21]. LSTMs offer higher accuracy but are computationally expensive, whereas kinematics and filtering models improve real-time performance without training.

In data association, traditional methods such as nearest neighbor (NN) [17], intersection over union (IoU) [22], and joint probabilistic data association (JPDA) [23] are computationally efficient and suitable for low-complexity scenarios. However, in high-density or complex occlusion environments, they are prone to misassociation and ID switching. Deep learning-based methods, such as graph neural networks (GNNs) [24], improve association accuracy by combining appearance, trajectory, and spatiotemporal information. However, their high computational complexity limits their widespread use in real-time applications. To address this issue, the Hungarian algorithm [4], [5], [25], as an efficient association method, provides excellent performance in real-time scenarios, but its effectiveness depends on the design of the association cost function.

D. Association Cost

The design of association cost functions is of paramount importance in MOT [26]. Owing to its simple structure, ease of implementation, and scale invariance, the commonly used IoU-based cost function is widely applied in MOT methods [4]. However, when the target boxes do not overlap, the IoU fails to reflect the relative distance or spatial relationship between the targets, limiting its application in complex scenarios. To address this, researchers have proposed variants such as the P-IoU [18], SDIoU [5], and distance IoU (DIoU) [19], which optimize the matching process by introducing the minimum enclosing rectangle or center point distance. However, in scenarios with complex occlusions or target intersections, association errors or failure in the fusion of detection results may still occur. Therefore, designing a cost function that comprehensively considers both target geometry and motion information [17] while ensuring high accuracy and controlling computational complexity remains a key challenge in current MOT research. The cost function based on the RGDIoU proposed in this paper optimizes computational efficiency while improving accuracy, thus effectively enhancing the stability of data associations.

III. Proposed Method

This paper proposes an online 3D MOT framework for autonomous driving, which integrates adaptive fusion technology, motion compensation, and multilevel data association methods to improve detection and tracking accuracy and robustness in complex traffic scenarios. The overall process of the framework is illustrated in Figure 1.

The framework consists of complementary modules. First, LiDAR and camera data are preprocessed for multisensor fusion. The center-plane adaptive fusion method combines these two types of sensor data, increasing the detection accuracy. A motion compensation mechanism based on an IMU and a GPS reduces the impact of vehicle motion on target localization. An RGDIoU cost function, combined with multilevel spatial indexing, improves data association. Finally, a coupled state filter optimizes target state estimation, enhancing tracking stability.

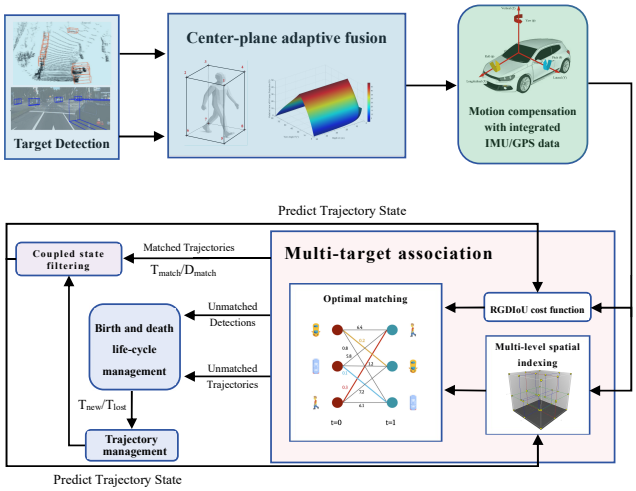

Fig. 1: Proposed system pipeline.

A. Center-Plane Adaptive Fusion

Target Detection: For each frame of the LiDAR point cloud data, 3D object detection is performed via the voxel convolutional network proposed by J. Deng et al. [12]. This method produces a 3D detection set $D_t^L = \{(x_k, y_k, z_k, h_k, w_k, l_k, \theta_k), c_k^L\}_{k=1}^{N_t}$, where (x_k, y_k, z_k) represents the target's center coordinates, (h_k, w_k, l_k) are the target's dimension parameters, θ_k is the yaw angle, and c_k^L is the target's class label. On the corresponding visual image at the same timestamp, 2D object detection is performed via the YOLOv8 model [27], resulting in the 2D detection set $D_t^V = \{b_s^V, c_s^V\}_{s=1}^{M_t}$, where b_s^V is the 2D bounding box of the target, and c_s^V is the class label. To enhance the robustness and generalization ability of the YOLOv8 model, data augmentation strategies such as color normalization, random cropping, and brightness perturbation are introduced during the training process.

Center-Plane Adaptive Fusion: First, the 3D detection box $D_t^P\{k\}$ in the LiDAR coordinate system $\{L\}$ is converted into a homogeneous coordinate vector of the eight vertices $P_L^{(i)} = (x_l^{(i)}, y_l^{(i)}, z_l^{(i)}, 1)^\top$, where i denotes the vertex index, $i = 1, \ldots, 8$. Using the LiDAR-to-camera transformation matrix $T_{CL} \in \mathbb{R}^{4\times4}$, which includes both the rotation and translation components, and the camera calibration matrix R_{rect}, the vertex coordinates in the camera coordinate system are computed as:

$$P_C^{(i)} = R_{\text{rect}} \cdot T_{CL} \cdot P_L^{(i)} = (x_c^{(i)}, y_c^{(i)}, z_c^{(i)}, 1)^\top \quad (1)$$

Next, a composite projection matrix $M = P_2 R_{\text{rect}} T_{CL} \in \mathbb{R}^{3\times4}$ is constructed via the camera intrinsic matrix $P_2 \in \mathbb{R}^{3\times4}$, and the standard pixel coordinates $\{(u_i, v_i)\}_{i=1}^8$ for each vertex $P_L^{(i)}$ are calculated as:

$$(u_i, v_i) = \left(\left[MP_L^{(i)}\right]_1 / \left[MP_L^{(i)}\right]_3, \left[MP_L^{(i)}\right]_2 / \left[MP_L^{(i)}\right]_3\right) \quad (2)$$

where $\left[MP_L^{(i)}\right]_j$ denotes the j-th component of the vector $MP_L^{(i)}$. By performing extremum calculations, the axis-aligned 2D detection box $b_k^L = (u_{\min}, v_{\min}, u_{\max}, v_{\max})$ is

obtained, where $u_{\min} = \min_i u_i$, $v_{\min} = \min_i v_i$, $u_{\max} = \max_i u_i$, and $v_{\max} = \max_i v_i$. The horizontal width of the detection box is given by:

$$\hat{w}_{8\text{pt}} = \left| \frac{f\frac{w}{2}\cos\theta + f\frac{l}{2}\sin\theta}{Z_c - \frac{w}{2}\sin\theta + \frac{l}{2}\cos\theta} - \frac{-f\frac{w}{2}\cos\theta - f\frac{l}{2}\sin\theta}{Z_c + \frac{w}{2}\sin\theta - \frac{l}{2}\cos\theta} \right| \tag{3}$$

where f represents the camera focal length. When the target's orientation angle is not 90 degrees, $\hat{w}_{8\text{pt}}$ will be larger than the actual target width. To avoid diagonal pseudodistortion, this paper proposes an adaptive projection correction method based on the center-plane. This method selects two pairs of center points from the sides that best represent the target's width and automatically decides whether to use the 'left-right' or 'front-back' sides for correction on the basis of an adaptive surface selection criterion. This adjustment aims to correct the horizontal boundary of the projection box.

Specifically, the center points of the left and right sides, c_R and c_L, and the center points of the front and back sides, c_F and c_B, are defined as:

$$c_R = \frac{1}{4}\sum_{i\in\{1,3,5,7\}} P_C^{(i)}, \quad c_L = \frac{1}{4}\sum_{i\in\{2,4,6,8\}} P_C^{(i)}$$

$$c_F = \frac{1}{4}\sum_{i\in\{1,2,5,6\}} P_C^{(i)}, \quad c_B = \frac{1}{4}\sum_{i\in\{3,4,7,8\}} P_C^{(i)} \tag{4}$$

The adaptive selection of the two center points, c^+ and c^-, and the adaptive surface selection criterion are as follows:

$$(c^+, c^-) = \begin{cases} c_B, c_F, & \text{if } \theta > \tan^{-1}\left(\frac{w}{l}\right) \\ c_R, c_L, & \text{if } \theta \le \tan^{-1}\left(\frac{w}{l}\right) \end{cases} \tag{5}$$

The horizontal width is given by:

$$\hat{w}_{\text{CAA}} = \begin{cases} \dfrac{flZ_c\sin\theta}{Z_c^2 - l\cos\theta/2^2}, & \text{if } \theta > \tan^{-1}\left(\frac{w}{l}\right) \\ \dfrac{fwZ_c\cos\theta}{Z_c^2 - w\sin\theta/2^2}, & \text{if } \theta \le \tan^{-1}\left(\frac{w}{l}\right) \end{cases} \tag{6}$$

To prevent instability due to $\hat{w}_{8\text{pt}} \to 0$, a lower limit $s_{\min} = 0.75$ is introduced. The horizontal boundary is then corrected by the coefficient:

$$s_c = \max\left(\hat{w}_{\text{CAA}}/\hat{w}_{8\text{pt}},\, s_{\min}\right) \tag{7}$$

Finally, the corrected boundaries are as follows:

$$u_{\min}^{\text{CAA}} = (u_{\min} + u_{\max} - s_c(u_{\max} - u_{\min}))/2$$
$$u_{\max}^{\text{CAA}} = (u_{\min} + u_{\max} + s_c(u_{\max} - u_{\min}))/2 \tag{8}$$

2D MultiObject Association and Classification Correction: In the 2D object association phase, the center-plane adaptive projection box of the LiDAR is denoted as $\{b_k^L, c_k^L, Z_c^k, w_k, l_k\}_{k=1}^N$, and the YOLOv8 detection box is denoted as $\{b_s^V, c_s^V\}_{s=1}^M$. For each pair (k, s), the L_1 distance between the centers of the two boxes is calculated as $\Delta_{ks} = \|m(b_k^L) - m(b_s^V)\|_1$, where $m(\cdot)$ represents the pixel center coordinates of the detection boxes. A candidate pair is added to the matching list only if $\Delta_{ks} \le \beta\frac{w_k + l_k}{2Z_{ck}}$, where $\beta = 0.15$. This depth-normalized threshold is automatically relaxed for closer distances and tightened for longer distances.

For all candidate pairs, a composite score is defined as:

$$S_{ks} = w_{\text{IoU}} \cdot \text{IoU}(b_k^L, b_s^V) + w_{\text{cls}} \cdot \mathbb{1}_{c_k^L = c_s^V} \tag{9}$$

where $\text{IoU}(\cdot)$ is the intersection over union function, and where $\mathbb{1}$ is the Iverson bracket function. The Hungarian algorithm is used to minimize the cost $-S_{ks}$ to obtain a global optimal one-to-one match, discarding pairs with scores below a threshold τ_S.

For category correction, a weighted posterior method is used, specifically:

$$c_j^L = \arg\max_{c\in C}\left[0.2P_L(c|b_j^L) + 0.8P_V(c|b_s^V)\right] \tag{10}$$

where P_L and P_V represent the classification confidence of LiDAR and visual detection, respectively. For unmatched LiDAR boxes, if they are outside the camera field of view, they are retained; if inside the view, they are retained only if $Z_c^j \ge 30\,\text{m}$; otherwise, they are considered false detections and discarded.

B. Motion Compensation

This paper proposes a motion compensation method that combines an inertial measurement unit (IMU) and a global positioning system (GPS) for vehicle motion. By accurately estimating the vehicle's pose and displacement, the target position can be transformed from the vehicle coordinate system to the world coordinate system, thereby reducing the impact of vehicle motion on target position estimation and association. The specific method is shown in Algorithm 1.

Algorithm 1: Object Coordinate Transformation

Input: LiDAR object position $P_{lidar} = [x_{lidar}, y_{lidar}, z_{lidar}]$, vehicle GPS (lat_{car}, lon_{car}), attitude (ϕ, θ, ψ), reference GPS (lat_0, lon_0), Earth radius R

Output: Object position in world coordinates $P_{world_{xy}} = [x_{world}, y_{world}]$

1: **Step 1:** Convert GPS to world coordinates using Mercator projection
$x_{world} = R \cdot (lon_{car} - lon_0) \cdot \cos(lat_{car} \times \frac{\pi}{180})$,
$y_{world} = R \cdot \ln\left(\tan\left(\frac{90 + lat_{car}}{360} \cdot \pi\right)\right) \cdot \cos(lat_{car} \times \frac{\pi}{180})$

2: **Step 2:** Compute rotation matrix $R_{vehicle}$ for vehicle-to-world transformation

$$R_{vehicle} = R_z(\psi) \cdot R_y(\theta) \cdot R_x(\phi)$$

where $R_z(\psi)$, $R_y(\theta)$, and $R_x(\phi)$ are the rotation matrices for yaw, pitch, and roll angles, respectively.

3: **Step 3:** Transform LiDAR position to world coordinates

$$P_{world} = R_{vehicle} \cdot P_{lidar} + T_{vehicle}$$

where $T_{vehicle}$ is the vehicle's position in world coordinates.

In the tracking process of each frame, the target position is first transformed into the world coordinate system via Algorithm 1. Then, target prediction and data association are performed in this coordinate system. This transformation ensures that the prediction and association processes are unaffected by the vehicle's movement and rotation. Even

when the vehicle undergoes a large-angle rotation, the relative position of the target remains stable, thus improving the accuracy of target association and the robustness of the tracking system. Since the IMU provides high-frequency pose and acceleration data while the GPS provides low-frequency position information, it is necessary to perform linear interpolation on the GPS data to align it with the IMU timestamps. In practical applications, an error-state Kalman filter (ESKF) [28] is used to filter the IMU and GPS data, optimizing the data accuracy.

C. State Modeling and Filtering

To comprehensively characterize the three-dimensional dynamic properties of the target, a high-dimensional coupled state vector is adopted. This vector combines the target's 3D position, velocity, size, size change rate, heading angle, and rate of change. The state vector is defined as:

$$S(k) = [x, y, z, l, w, h, \theta, v_x, v_y, v_z, v_l, v_w, v_h, \omega]^T \quad (11)$$

where x, y, z are position coordinates, v_x, v_y, v_z are velocity components, l, w, h are target dimensions, v_l, v_w, v_h are size change rates, θ is the heading angle, and ω is the heading angle rate.

The system uses a constant velocity (CV) motion model with standard linear Gaussian Kalman filtering. The state transition equation is as follows:

$$S(k+1) = \Phi \cdot S(k) + \omega(k), \quad \Phi = \begin{bmatrix} I_{7\times7} & \Delta t I_{7\times7} \\ 0_{7\times7} & I_{7\times7} \end{bmatrix} \quad (12)$$

where $\omega(k)$ represents process noise and Φ is the state transition matrix.

The observation equation is as follows:

$$M(k) = H \cdot S(k) + v(k), \quad H = \begin{bmatrix} I_{7\times7} & 0_{7\times7} \end{bmatrix} \quad (13)$$

where $v(k)$ represents observation noise. The system follows standard Kalman filter recursive updates [4]. Coupled modeling improves continuity and adaptability, ensuring stable tracking in dynamic environments.

D. 3D RGDIoU-based MultiObject Association

In 3D MOT for autonomous driving, as the number of detected objects N and tracking trajectories M increase, the calculation of the association costs between trajectories and detections becomes a performance bottleneck. The computational complexity of traditional methods is $O(M \times N)$, which makes it difficult to meet real-time requirements when the number of objects and the computational load of the cost function are large. To address this, an efficient 3D multiobject association method is proposed, aiming to reduce computational complexity and improve the accuracy of associations.

Multilevel Spatial Association: To improve computational efficiency, a multilevel spatial association structure combining coarse-grained filtering and fine-grained retrieval is proposed to reduce the number of associations and computational load. First, on the basis of scene density and object distribution, the 3D space is dynamically divided into coarse-grained grids with a side length of L, and fine-grained tree structures are built in grids with a larger number of objects for retrieval. For each predicted trajectory position, the query radius r_{kd} is dynamically set on the basis of the maximum velocity v_{max}, the time step Δt, and a safety margin δ as:

$$r_{kd} = v_{max} \cdot \Delta t + \delta \quad (14)$$

The association cost $\mathrm{cost}_{i,j}$ is calculated only for the retrieved object pairs, thereby reducing the computational complexity.

3D RGDIoU-based Cost Function Construction: To measure the overlap between trajectories and detected objects accurately, we introduce an improved 3D RGDIoU. This method combines the object's rotational angle and geometric distance factors to provide a more precise overlap metric.

1) 3D IoU Calculation: First, the overlap region in the 2D plane, denoted as I_{2D}, is computed. Then, on the basis of the height of the overlap region, the 3D overlap region is obtained as $I_{3D} = I_{2D} \cdot \mathrm{overlap\_height}$. Next, the union of the volumes of the two boxes is calculated as:

$$U_{3D} = (w_a \cdot l_a \cdot h_a) + (w_b \cdot l_b \cdot h_b) - I_{3D} \quad (15)$$

where w_a, l_a, h_a and w_b, l_b, h_b represent the width, length, and height of boxes a and b, respectively. The basic 3D IoU value is as follows:

$$\mathrm{IoU}_{3D} = I_{3D}/U_{3D}, \quad \text{with} \quad U_{3D} > 10^{-6} \quad (16)$$

2) 3D RGDIoU Calculation: First, on the basis of the Bird's Eye View (BEV), the diameter of the minimum enclosing circle $c = \max_{i,j}\left(\sqrt{(x_i - x_j)^2 + (y_i - y_j)^2}\right)$ and the center point distance $d = \sqrt{(x_a - x_b)^2 + (y_a - y_b)^2}$ are calculated. Here, (x_i, y_i) and (x_j, y_j) are the coordinates of all the vertices of the boxes, and (x_a, y_a) and (x_b, y_b) are the center coordinates of boxes a and b, respectively. Then, the normalized distance is calculated as:

$$\mathrm{dis}_n = d/(c + 10^{-6}) \quad (17)$$

Next, the difference in rotation angles between the two boxes is calculated as $\theta = |\theta_a - \theta_b|$, and we take $\theta = \min(\theta, 2\pi - \theta)$ to ensure that the rotation angle difference remains within the range $[0, \pi]$. The normalized rotation angle difference is then computed as:

$$\mathrm{ang}_n = \theta/\pi \quad (18)$$

Finally, the 3D RGDIoU value is given by:

$$\mathrm{RGDIoU}_{i,j} = \mathrm{IoU}_{3D} - \alpha \cdot \mathrm{dis}_n - \beta \cdot \mathrm{ang}_n \quad (19)$$

where $\alpha = 0.7$ and $\beta = 0.5$ are tuning weights that control the influence of distance and rotation differences on the RGDIoU.

3) Cost Calculation: To minimize the cost function, the cost function is defined as:

$$\mathrm{cost}_{i,j} = 1 - \mathrm{RGDIoU}_{i,j} \quad (20)$$

The objective of this cost function is to minimize the association error. A smaller cost indicates a more accurate association result.

Compared with traditional IoU methods, 3D RGDIoU accounts for the rotational differences and geometric distances between targets, providing a more accurate overlap measurement when dealing with rotating and dynamic targets. By introducing normalized distance and angle ratios, 3D RGDIoU improves the accuracy of target association in autonomous driving scenarios, particularly in dynamic, rotating, and complex environments, offering significant advantages.

Optimization Matching: After determining the association cost between tracks and detected targets, we use the Hungarian algorithm to achieve optimal matching and minimize the total association cost. First, construct a cost matrix of size $m \times n$, where m is the number of tracks being tracked currently, and where n is the number of detected candidate targets. The matrix elements $\mathrm{cost}_{i,j}$ are computed as in equation (20). The cost matrix is defined as:

$$\mathrm{cost}_{matrix} = \begin{bmatrix} \mathrm{cost}_{1,1} & \cdots & \mathrm{cost}_{1,n} \\ \vdots & \ddots & \vdots \\ \mathrm{cost}_{m,1} & \cdots & \mathrm{cost}_{m,n} \end{bmatrix} \tag{21}$$

The cost matrix cost_{matrix} is then input into the Hungarian algorithm to solve the minimum cost bipartite graph matching problem. We define the binary variable $x_{i,j}$ to indicate whether track i is matched with detection j. The optimization objective is as follows:

$$\min \sum_{i,j} \mathrm{cost}_{i,j} \cdot x_{i,j} \tag{22}$$

IV. Experiment

In this section, we analyze the experimental results of the proposed method on the KITTI and nuScenes datasets to evaluate its performance in 2D and 3D MOT tasks and compare it with existing state-of-the-art methods. The experiments include quantitative analysis and ablation studies, aimed at validating the effectiveness, generalizability, and advantages of the proposed method. All the experiments were conducted on a lightweight computing platform equipped with an Intel Core i7-13700 processor, 32 GB of RAM, and an NVIDIA RTX 2050 GPU.

A. Experimental Setup

Datasets: The primary dataset used in this experiment is the KITTI MOT dataset [29], which provides images, 3D LiDAR point clouds, and GPS/IMU reference trajectories. Experiments were performed on both the validation and test sets to assess the proposed method comprehensively. Additionally, we performed a quantitative analysis on the nuScenes dataset [30]. Both the KITTI and nuScenes datasets provide ground truth labels for target detection, but the ground truth labels for the test sets are not directly accessible. To evaluate the test set, submissions are made to the official KITTI and nuScenes websites for comparative analysis. Since pedestrian tracking is more challenging than vehicle tracking is, our focus is primarily on the pedestrian subset, and the results are compared with those of other state-of-the-art (SOTA) methods.

Evaluation Metrics: In the KITTI dataset, the evaluation is based on the CLEAR MOT [31] and HOTA [32] standards, which include metrics such as the number of identity switches (IDSw), frame rate (FPS), number of fragments (Frag), multiobject tracking precision (MOTP), multiobject tracking accuracy (MOTA), AMOTP, AMOTA, sAMOTA, and the overall metric HOTA. These metrics provide a comprehensive evaluation of tracking accuracy and target association performance.

Baseline Methods: To demonstrate the effectiveness of the proposed method, we compare it with several recent state-of-the-art (SOTA) methods, including QDense (CVPR2021) [11], CTrack (ECCV2020) [6], Polar (ECCV2022) [26], AB3D (IROS 2020) [4], FNC2 (TIV 2024) [21], Eager (ICRA 2021) [21], SF (SJ 2023) [5], MPN (IJCV 2022) [24], Triplet (CVPR2022) [8], EAFF (SP 2024) [17], YONTD (arXiv 2023) [7], MM (TII 2024) [9], APP (IJCV2025) [33], AHMOT (IOTJ 2025) [2], and CR3DT (IROS 2024) [25].

Object Detectors: This study uses widely adopted object detectors for a fair comparison with SOTA methods. YOLOv8 [27] is used as the 2D detector in both the KITTI and nuScenes datasets, while Voxel R-CNN [12] is employed for 3D detection on KITTI, and Largekernel [34] is used for 3D detection on nuScenes.

B. Quantitative Experimental Results

KITTI 2D MOT: To comprehensively evaluate the performance of the proposed method in 2D MOT, we compare it with various state-of-the-art algorithms. Table I shows the performance of different methods on the KITTI 2D MOT test set, including key metrics such as MOTA, HOTA, DetPr, AssA, and FPS.

Table I clearly shows that the proposed method outperforms the other methods in terms of both the MOTA and HOTA metrics. Compared with MM, our method shows an improvement of 3.77 percentage points in MOTA and 0.96 percentage points in HOTA, indicating a advantage in accuracy and consistency for multiobject tracking. Although the detection precision (DetPr) is slightly lower, our method clearly has an advantage in terms of association accuracy (AssA) because of its efficient data association strategy, which outperforms most comparison methods. Despite a lower FPS than some other methods, the FPS is still sufficient to meet the requirements of most real-time applications, demonstrating that our method maintains reasonable computational efficiency while optimizing accuracy.

Fig. 2 compares the tracking performance of our method with AB3D under scenarios involving multiple object interactions and occlusions. Between frames 136 and 165, AB3D (red and purple dashed boxes) suffers from missed detections, identity switches, and tracking errors caused by ego-vehicle motion. In contrast, our method maintains accurate associations and continuous trajectories despite challenges

Fig. 2: Comparison of tracking performance on the KITTI dataset between the AB3D method (a) and our proposed method (b). Different colors correspond to distinct object identities. In sequence 18, AB3D demonstrates identity switches (a), while our method ensures continuous trajectory maintenance despite challenging conditions (b).

TABLE I: Performance Comparison of LiDAR or Image-based Methods on the KITTI 2D MOT Leaderboard

Method	MOTA↑	HOTA↑	DetPr↑	AssA↑	FPS↑
QDense [11]	55.55	41.12	70.39	38.10	14.3
CTrack [6]	53.84	40.35	66.83	36.93	**222**
Eager [21]	49.82	39.38	61.49	38.72	90.9
MPN [24]	46.23	45.26	58.30	47.28	50
Polar [26]	46.98	43.59	57.40	48.12	50
SF [5]	39.04	43.42	53.81	48.83	100
Triplet [8]	50.08	42.77	71.91	46.54	10
AB3D [4]	38.13	37.81	59.35	44.33	212
YONTD [7]	26.19	25.89	54.99	25.02	10
EAFF [17]	42.01	40.20	60.03	45.63	100
FNC2 [21]	56.05	46.55	59.38	46.68	100
MM [9]	56.19	49.28	**72.98**	**55.33**	74
APP [33]	55.45	42.73	67.27	41.15	25
Ours	**58.92**	**50.24**	63.33	50.71	90

such as ego-motion, dense target distributions, and severe occlusions, thereby significantly reducing mis-associations and identity switches and ensuring stable, robust tracking.

nuScenes 3D MOT: To validate the generalizability and robustness of the proposed method in more complex scenarios, we conducted experiments on the nuScenes dataset, with the results shown in Table II.

TABLE II: 3D MOT performance achieved on the nuScenes Test Sequences.

Method	AMOTA↑	AMOTP↑	Recall↑	IDSw↓
QDense [11]	0.312	0.485	0.448	2011
Triplet [8]	0.362	0.362	0.489	528
AB3D [4]	0.141	0.141	0.257	1088
CTrack [6]	0.142	**0.887**	0.431	2086
Eager [21]	0.744	0.414	0.797	574
Polar [26]	0.806	0.398	0.800	**171**
SF [5]	0.621	0.505	0.698	572
CR3DT [25]	0.339	0.409	0.430	864
AHMOT [2]	0.714	0.439	0.750	359
Ours	**0.808**	0.419	**0.861**	311

As shown in Table II, the proposed method outperforms other methods in terms of AMOTA and Recall, demonstrating stronger robustness and higher recall ability, enabling more effective detection and tracking of targets. The IDSw of 311 is lower than that of most comparison methods, indicating that the proposed method performs better in terms of target association and identity consistency, effectively reducing identity switches and ensuring continuous tracking of targets. These results show that the proposed method has strong generalizability and robustness in complex scenarios, maintaining efficient and stable tracking performance.

C. Ablation Study

To comprehensively evaluate the contribution of each module in the proposed system, we designed and conducted four sets of ablation experiments, analyzing the effectiveness of the RGDIoU-based cost function (R.), center-plane adaptive fusion (F.), and inertial navigation-based motion compensation (C.) modules. Table III summarizes the experimental results under different configurations.

TABLE III: Impact of Component Removal on Tracking Performance: Ablation Study Results

R.	F.	C.	sAMOTA↑	AMOTP↑	Recall↑	IDSw↓	Frag↓
✓	✓	✓	**86.55**	70.13	88.01	**173**	**365**
✓	✓	×	78.88	**70.35**	83.33	308	492
✓	×	✓	80.98	61.33	85.36	184	527
×	✓	✓	84.08	68.09	**90.16**	303	535
×	×	×	71.75	61.25	83.73	269	582

The results of the ablation study show that center-plane adaptive fusion, the RGDIoU-based cost function, and inertial navigation-based motion compensation modules all play significant roles in system performance. When all the modules are enabled, sAMOTA, Recall, and IDSw achieve the best performance, indicating that the synergistic effect of the modules effectively improves the accuracy and consistency of multiobject tracking. The removal of any module leads to a decrease in performance, particularly in accuracy and stability. For example, removing the fusion module results in a decrease in sAMOTA to 0.8098; removing the motion compensation module increases IDSw to 308; and removing

the RGDIoU-based cost function significantly increases Frag. Overall, the collaboration of these modules plays a critical role in achieving efficient and stable object tracking in complex dynamic environments.

V. Conclusion

In this paper, we propose a novel 3D multiobject tracking method that integrates center-plane adaptive multi-sensor fusion, motion compensation, and precise data association algorithms. The aim is to enhance the safety and reliability of autonomous driving systems in complex traffic environments. The experimental results show that the proposed method outperforms the compared state-of-the-art methods in terms of tracking accuracy and stability on both the KITTI and nuScenes datasets. Notably, substantial improvements were observed in the HOTA and sAMOTA metrics. Despite the excellent performance in most scenarios, there is still room for improvement in metrics such as MOTP. Future research will focus on further optimizing tracking accuracy, particularly in complex scenarios with fast-moving objects and occlusions. Additionally, research will explore more efficient sensor fusion and data association optimization strategies to further enhance the practical application potential of multiobject tracking systems in the field of autonomous driving.

References

[1] Z. Shao, H. Wang, Y. Cai, L. Chen, and Y. Li, "UA-Fusion: Uncertainty-Aware Multimodal Data Fusion Framework for 3-D Object Detection of Autonomous Vehicles," *IEEE Trans. Instrum. Meas.*, vol. 74, pp. 1–16, 2025, Art. no. 2514916.

[2] C. Jiang, M. Zhang, Y. Wang, and A. Zhang, "AHMOT: Adaptive Kalman Filtering and Hierarchical Data Association for 3-D Multiobject Tracking in IoT-Enabled Autonomous Vehicles," *IEEE Internet Things J.*, vol. 12, no. 12, pp. 21290–21303, Jun. 15, 2025.

[3] X. Ge, S. Zhu, Y. Gu, *et al.*, "An Anomaly Detection Method for Railway Track Using Semisupervised Learning and Vision-Lidar Decision Fusion," *IEEE Trans. Instrum. Meas.*, vol. 73, pp. 1–15, 2024, Art. no. 5023215.

[4] X. Weng, J. Wang, D. Held, and K. Kitani, "3D Multi-Object Tracking: A Baseline and New Evaluation Metrics," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, Las Vegas, NV, USA, 2020, pp. 10359–10366.

[5] X. Wang, C. Fu, J. He, S. Wang, and J. Wang, "StrongFusionMOT: A Multi-Object Tracking Method Based on LiDAR-Camera Fusion," *IEEE Sensors J.*, vol. 23, no. 11, pp. 11241–11252, Jun. 2023.

[6] X. Zhou, V. Koltun, and P. Krähenbühl, "Tracking Objects as Points," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2020, pp. 474–490.

[7] X. Wang, J. He, C. Fu, T. Meng, and M. Huang, "You Only Need Two Detectors to Achieve Multi-Modal 3D Multi-Object Tracking," *arXiv preprint arXiv:2304.08709*, 2023.

[8] N. Marinello, M. Proesmans, and L. Van Gool, "TripletTrack: 3D Object Tracking Using Triplet Embeddings and LSTM," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2022, pp. 4500–4510.

[9] L. Xu and Y. Huang, "Rethinking Joint Detection and Embedding for Multiobject Tracking in Multiscenario," *IEEE Trans. Ind. Informat.*, vol. 20, no. 6, pp. 8079–8088, Jun. 2024.

[10] B. Cheong, J. Zhou, and S. Waslander, "JDT3D: Addressing the Gaps in LiDAR-Based Tracking-by-Attention," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2024, pp. 161–177.

[11] J. Pang, C. Li, Y. Zhang, L. Li, and C. Shi, "Quasi-Dense Similarity Learning for Multiple Object Tracking," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2021, pp. 164–173.

[12] J. Deng, S. Shi, P. Li, W. Zhou, Y. Zhang, and H. Li, "Voxel R-CNN: Towards High Performance Voxel-Based 3D Object Detection," in *Proc. AAAI Conf. Artif. Intell.*, vol. 35, no. 2, pp. 1201–1209, 2021.

[13] Y. Xu, Z. Chen, C. Deng, S. Wang, and J. Wang, "LCDL: Toward Dynamic Localization for Autonomous Landing of Unmanned Aerial Vehicle Based on LiDAR–Camera Fusion," *IEEE Sensors J.*, vol. 24, no. 16, pp. 26407–26415, 2024.

[14] S. Li, K. Geng, G. Yin, *et al.*, "MVMM: Multiview Multimodal 3-D Object Detection for Autonomous Driving," *IEEE Trans. Ind. Informat.*, vol. 20, no. 1, pp. 845–853, 2023.

[15] N. Aharon, R. Orfaig, and B. Z. Bobrovsky, "BoT-SORT: Robust Associations Multi-Pedestrian Tracking," *arXiv preprint arXiv:2206.14651*, 2022.

[16] N. Mahdian, M. Jani, A. M. S. Enayati, *et al.*, "Ego-Motion Aware Target Prediction Module for Robust Multi-Object Tracking," *arXiv preprint arXiv:2404.03110*, 2024.

[17] 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 Process.*, vol. 218, 109367, 2024.

[18] X. Wu and J. Xu, "P-IoU: Accurate Motion Prediction Based Data Association for Multi-Object Tracking," in *Proc. Int. Conf. Neural Inf. Process. (ICONIP)*, Singapore, 2024, pp. 484–496.

[19] Z. Zheng, P. Wang, W. Liu, J. Li, R. Ye, and D. Ren, "Distance-IoU Loss: Faster and Better Learning for Bounding Box Regression," in *Proc. AAAI Conf. Artif. Intell.*, vol. 34, no. 7, pp. 12993–13000, 2020.

[20] K. Yi, K. Luo, X. Luo, X. Peng, D. Wang, and Y. Song, "UCMCTrack: Multi-Object Tracking With Uniform Camera Motion Compensation," in *Proc. AAAI Conf. Artif. Intell.*, vol. 38, no. 7, pp. 6702–6710, 2024.

[21] C. Jiang, Z. Wang, H. Liang, and Y. Wang, "A Novel Adaptive Noise Covariance Matrix Estimation and Filtering Method: Application to Multi-Object Tracking," *IEEE Trans. Intell. Vehicles*, vol. 9, no. 1, pp. 626–641, Jan. 2024.

[22] A. Kim, A. Osep, and L. Leal-Taixé, "EagerMOT: 3D Multi-Object Tracking via Sensor Fusion," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, Xi'an, China, 2021, pp. 11315–11321.

[23] S. He, H. S. Shin, and A. Tsourdos, "Trajectory Optimization for Multitarget Tracking Using Joint Probabilistic Data Association Filter," *J. Guid. Control Dyn.*, vol. 43, no. 1, pp. 170–178, 2020.

[24] G. Brasó, O. Cetintas, and L. Leal-Taixé, "Multi-Object Tracking and Segmentation Via Neural Message Passing," *Int. J. Comput. Vis.*, vol. 130, no. 12, pp. 3035–3053, Dec. 2022.

[25] N. Baumann, *et al.*, "CR3DT: Camera-RADAR Fusion for 3D Detection and Tracking," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, Abu Dhabi, United Arab Emirates, 2024, pp. 4926–4933.

[26] A. Kim, G. Brasó, A. Ošep, *et al.*, "PolarMOT: How Far Can Geometric Relations Take Us in 3D Multi-Object Tracking?," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Cham, Switzerland, 2022, pp. 41–58.

[27] G. Jocher, "YOLOv8 Release v8.1.0," *GitHub*, 2024. [Online]. Available: https://github.com/ultralytics/ultralytics/releases/tag/v8.1.0

[28] J. He, C. Sun, B. Zhang, *et al.*, "Adaptive Error-State Kalman Filter for Attitude Determination on a Moving Platform," *IEEE Trans. Instrum. Meas.*, vol. 70, pp. 1–10, 2021.

[29] A. Geiger, P. Lenz, and R. Urtasun, "Are We Ready for Autonomous Driving? The KITTI Vision Benchmark Suite," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2012, pp. 3354–3361.

[30] H. Caesar, V. Bankiti, A. H. Lang, *et al.*, "nuScenes: A Multimodal Dataset for Autonomous Driving," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2020, pp. 11621–11631.

[31] K. Bernardin and R. Stiefelhagen, "Evaluating Multiple Object Tracking Performance: The CLEAR MOT Metrics," *EURASIP J. Image Video Process.*, vol. 2008, pp. 1–10, 2008.

[32] J. Luiten, A. Osep, P. Dendorfer, P. Torr, A. Geiger, L. Leal-Taixé, and B. Leibe, "HOTA: A Higher Order Metric for Evaluating Multi-Object Tracking," *Int. J. Comput. Vis.*, vol. 129, pp. 548–578, 2021.

[33] T. Zhou, Q. Ye, W. Luo, *et al.*, "APPTracker+: Displacement Uncertainty for Occlusion Handling in Low-Frame-Rate Multiple Object Tracking," *Int. J. Comput. Vis.*, vol. 133, pp. 2044–2069, 2025.

[34] Y. Chen, J. Liu, X. Zhang, *et al.*, "Largekernel3D: Scaling Up Kernels in 3D Sparse CNNs," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2023, pp. 13488–13498.