IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2026, Vienna, Austria. Cite as RA-L paper.

IMH-MOT: Interactive Multi-Hierarchical Image and Point Cloud Fusion for Multi-Object Tracking

Wenyuan Qin[1,2], Zhiyan Zhou[1], Luo Jiong[1], Chengwei Pan[1], Hao Xu[1], Xiwang Dong[1], *Senior Member, IEEE* and Danwei Wang[3], *Life Fellow, IEEE*

***Abstract*— Multi-object tracking (MOT) plays a critical role in applications such as autonomous driving and surveillance. Camera-based approaches offer rich texture features for object association, while LiDAR-based methods provide accurate geometric information for spatial reasoning. Although each modality addresses different challenges, their intrinsic discrepancies hinder effective cross-modal fusion and unified representation learning. To overcome these limitations, we propose IMH-MOT, an interactive multi-hierarchical MOT framework comprising three key modules. The Multi-modality Alignment Module (MMAM) enhances spatial representations by sampling and clustering instance-level point clouds. From different modalities are motion cues integrated by the Multi-modality Motion Estimation Module (MMEM) to build a unified motion model. To mitigate the impact of occlusion on single-frame appearance features, the Long-term Appearance Module (LAM) captures temporal appearance consistency by constructing a long-term appearance embedding. Guided by modality-aware cues from MMAM, MMEM generates reliable spatial representations, while LAM encodes robust long-term appearance features. These components are jointly integrated through a Multi-hierarchical Data Association (MHDA) strategy, enabling stable and accurate tracking. Extensive experiments on the KITTI MOT benchmark demonstrate the effectiveness of our framework, achieving 80.90% HOTA, 89.73% MOTA, and 470 IDSW, outperforming state-of-the-art methods in both standard and challenging scenarios.**

I. INTRODUCTION

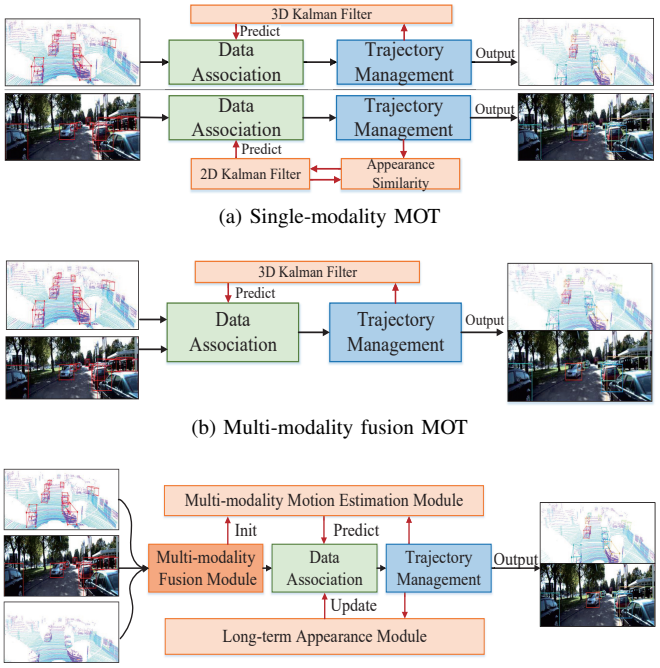
Multi-object tracking (MOT) aims to maintain consistent identification for each target across the sensor data stream and build trajectories for newly detected targets. Accurate and continuous MOT algorithms are essential for ensuring the safety of autonomous driving systems [1], [2] and mobile robots [3], [4]. Robotic systems equipped with Light Detection and Ranging (LiDAR) and cameras utilize MOT methods to collect critical information from the surrounding environment and assess the status of targets to reduce potential risks [1], [2].

This work was supported by the National Key R&D Plan of China under Grant No. 2024YFB4708300, the National Natural Science Foundation of China under Grants U2241217, 62473027, 62473029, 62403038, 62203032, and 62388101, and the Beijing Natural Science Foundation under Grants JQ23019 and 4232046. (Corresponding author: Hao Xu))

[1]The authors are with the Science and Technology on Aircraft Control Laboratory, Beihang University, Beijing, 100191(e-mail: wyqin@buaa.edu.cn; zhouzhiyan@buaa.edu.cn; zy2203509@buaa.edu.cn; pancw@buaa.edu.cn; xuhao3e8@buaa.edu.cn; xwdong@buaa.edu.cn)

[2]Zhongguancun Laboratory, Beijing, 100094, China

[3]Danwei Wang is with the School of Electrical and Electronic Engineering,Nanyang Technological University, Singapore 639798 (e-mail: edwwang@ntu.edu.sg).



(a) Single-modality MOT

(b) Multi-modality fusion MOT

(c) IMH-MOT

Fig. 1. Different MOT methods based on the TBD paradigm. **(a)** Early methods directly use single 3D or 2D detections to estimate and update the motion state. **(b)** While recent methods integrate 2D detections into the data association stage of 3D MOT, they often fail to consider that the lack of simultaneous observations typically results from insufficient point cloud density. **(c)** IMH-MOT aims to leverage multi-modal synergy to enhance the tracking process. First, MMAM processes the detections and clusters the original point cloud to recover the spatial information of 2D-only observations. Then, MMEM predicts target motion across modalities. Finally, LAM aggregates historical trajectory information to enhance appearance modeling and performs association with current detections.

Existing MOT methods primarily utilize camera [3], [5] and LiDAR [1], [4] as key sensors, as illustrated in Fig. 1a. Cameras provide rich texture details by capturing color and light intensity information about targets, and their cost-effectiveness has made them a popular choice in robotics [5]. However, the performance of camera can degrade significantly in overexposed or low-light environments, resulting in a substantial decline in detection and tracking accuracy. Additionally, due to the lack of spatial information, occlusion frequently leads to tracking failures. In contrast, LiDAR is less influenced by ambient lighting conditions and consistently captures spatial information, ensuring stable performance regardless of the time of day, which has driven the development of LiDAR-based method [2]. However, due to the inherent sparsity of LiDAR data, these methods rely on extensive point cloud

accumulation to ensure reliable performance and often struggle with irregular object motion.

Despite considerable advancements in single-modality approaches [2], [5], their fundamental limitations persist under complex scenarios. Under occlusion conditions or irregular motion patterns, these methods suffer significant performance degradation. Numerous multi-modality fusion methods [1], [6]–[8] have emerged with representative architectures illustrated in Fig. 1b. AdaptiveTrack [8] achieves robust tracking by employing multi-feature fusion with adaptive weighting to compute cross-modal similarity and dynamically select the association range. FusionTrack [7] leverages 2D detection results as prior information to filter out false positives in 3D detection, thereby enhancing overall tracking performance. MCCA-MOT [9] proposes an end-to-end framework that incorporates an adaptive diffusion fusion module for point cloud features, effectively increasing point cloud density and improving tracking stability. However, these methods still face challenges when objects are located at long distances. In such cases, objects may be visible in the 2D image plane but fail to appear in the 3D point cloud due to sparsity or occlusion, leading to incomplete spatial information. Moreover, current approaches often rely on rigid priors or fixed modality fusion strategies, which may degrade performance under varying observation conditions. A more flexible and context-aware fusion mechanism is therefore needed to improve robustness in diverse scenarios.

To address the aforementioned challenges, this paper presents a multi-hierarchical MOT framework for robust object tracking by fusing point cloud and image data, as illustrated in Fig. 1c. At each time step, the LiDAR and camera independently capture point clouds and images, which are processed by 3D and 2D detectors to generate detection results. MMAM evaluates cross-modal similarity to identify non-overlapping view targets. For these targets, an instance-level clustering algorithm is applied to recover their spatial information from sparse point clouds. In parallel, 3D detections are projected onto the 2D image plane, enabling LAM to extract texture features for each target. During the association stage, MMEM performs joint motion estimation across modalities, while LAM leverages long-term appearance cues to enhance association robustness. Finally, MHDA fuses spatial and appearance information to achieve stable and reliable multi-object tracking.

In summary, the contributions of this study are as follows:

- To address the limitation of overlapping-view between images and point cloud, a Multi-modality Alignment Module (MMAM) is introduced. By exploiting inter-modal geometric consistency through physical symmetry constraints, our module effectively suppresses outlier point cloud measurements.
- To address the unreliability of appearance features in single-frame observations, we propose a Long-Term Appearance Module (LAM) with temporal encoding. By leveraging a Long-Term Transformer, LAM encodes texture features along the temporal dimension, significantly enhancing trajectory appearance representation. Additionally, we introduce a Multi-Hierarchical Data Association (MHDA) strategy to integrate state associations across different modalities.
- Extensive experiments demonstrate that the proposed IMH-MOT achieves state-of-the-art performance on the KITTI benchmark. Additionally, ablation studies validate the effectiveness of both the MMAM and LAM.

II. RELATED WORK

A. Camera-based Multi-Object Tracking

Many camera-based methods have achieved significant progress, primarily driven by advances in object detection. A common approach is the tracking-by-detection (TBD) framework, which treats detection and association as separate steps. SORT [10] employs a Kalman filter to predict bounding box motion and uses intersection-over-union (IoU) for association. Building on this, SelectMOT [11] adopts a two-stage matching scheme and improves accuracy through quality-aware selection, mitigating the misalignment issues of confidence-based methods. To enable joint optimization of detection and tracking, end-to-end architectures have also been explored. ADA-Track [5] introduces a learnable association module based on edge-enhanced cross-attention, effectively fusing appearance and geometric features to enhance tracking performance. However, these methods still struggle with spatial reasoning. In particular, under occlusion, degraded texture information often leads to reduced accuracy.

B. LiDAR-based Multi-Object Tracking

Recent advancements in point cloud representation and 3D object detection have significantly improved multi-object tracking (MOT) performance [2], [12]. PolarMOT [12] formulates detections as graph nodes and employs a graph neural network to compute trajectory similarity, enabling robust tracking. HybridTrack [2] enhances accuracy by introducing a learnable Kalman filter that directly estimates residuals and Kalman gain from data, addressing the limited generalizability of hand-crafted models. However, these methods often struggle with irregular object motion, which degrades tracking stability. Furthermore, despite recent progress, they remain limited by point cloud sparsity. In sparse environments, detectors may lack sufficient contextual cues, hindering accurate and consistent tracking.

C. Fusion-based methods

EagerMOT [1] achieves high performance by independently modeling the motion of 2D and 3D detection bounding boxes, but it continues to struggle with occluded targets. To address this limitation, CAMO-MOT [13] introduces an occlusion prediction network and adaptively selects optimal texture features during the association process to maintain stable tracking. Building on this, DeepFusion [14] improves robustness by dynamically

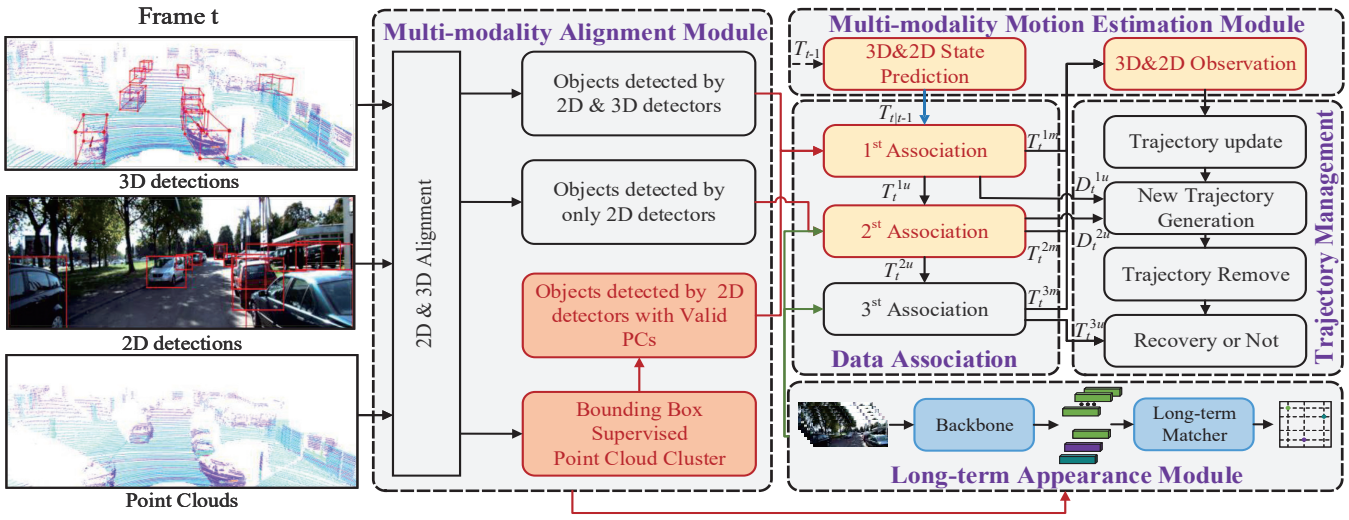


Fig. 2. **Architecture of the Proposed IMH-MOT Framework.** First, MMAM assigns cross-modal information to jointly observed targets. For targets visible only in the image modality, a 2D detection-guided clustering approach is applied to the raw point cloud to recover spatial information. For the remaining detections, only their 2D state information is retained. Subsequently, MMME predicts motion states based on the available modality, while LAM encodes long-term appearance features. Finally, MHDA performs multi-stage data association, and the trajectory management module handles the update, creation, and termination of trajectories based on the association outcomes.

selecting modality-specific cues to compute similarity under varying scenarios. Distinct from previous approaches, YONTD [15] unifies detection and tracking within a single framework to simplify the association process. In addition, it incorporates a confidence fusion module that jointly suppresses redundant trajectories and detections, thereby enhancing tracking accuracy. However, these methods often overlook the overlapping-view problem between modalities. When 2D and 3D observations are not temporally aligned, detections are processed independently without leveraging complementary information from the raw data. To mitigate this issue, IMH-MOT classifies 2D and 3D detections to identify targets with inconsistent observations and applies a target-guided point selection strategy to extract sparse point clouds, thereby compensating for missing spatial information. Furthermore, to address the limitations of fusion schemes that rely solely on single-frame appearance features, a LAM is introduced to enable trajectory modeling over extended temporal windows. Finally, a MHDA mechanism integrates both modules to achieve complementary fusion of spatial and texture information across modalities.

III. IMH-MOT

This section presents an overview of the proposed MOT framework IMH-MOT as shown in Fig. 2. MMAM leverages 2D detection priors to guide point cloud sampling and clustering, thereby enriching spatial information in non-co-view scenarios. MMME then fuses motion cues from both modalities for accurate motion estimation. Finally, LAM addresses the limitations of single-frame appearance association. It encodes temporal and aggregates appearance features over time.

A. Multi-modality Alignment Module (MMAM)

MMAM utilizes 3D and 2D detectors to generate two detection sets, $\mathbf{D}^{3d} = \{D_1^{3d}, D_2^{3d}, \ldots, D_t^{3d}\}$ and $\mathbf{D}^{2d} = \{D_1^{2d}, D_2^{2d}, \ldots, D_t^{2d}\}$, where the subscript t represents the frame number, and F is the total frame number of a video sequence. At frame t, each detection in D_t^{3d} is defined as $\{x, y, z, w, l, h, \theta\}$, where (x, y, z) denotes the object center, (w, l, h) the width, length, and height, and θ the heading angle of the 3D bounding box. Similarly, each detection in D_t^{2d} is represented as $\{u, v, w_p, h_p\}$, where (u, v) is the 2D box center and (w_p, h_p) are its width and height.

Existing methods [1], [14] typically consider either the texture features of detection results or the spatial information of the point cloud but fail to account for situations where the 3D detector may fail to detect the target due to an insufficient number of point clouds. 2D detector typically have a longer viewing distance, enabling them to effectively observe distant targets. However, it is often affected by occlusion due to the lack of spatial information. The original point cloud is sampled and clustered based on 2D detections to compensate for the lack of spatial information in the image plane. As described in [1], IoU is employed to establish the correspondence between 3D and 2D detections within the image plane at moment t. The similarity matrix can be written as $U_{3D}^{2D} = \text{IoU}(\pi(D_t^{3d}), D_t^{2d})$, where $\pi(\cdot)$ denotes the projection operation that maps a 3D bounding box from the world coordinate system onto the image plane, incorporating both intrinsic and extrinsic camera parameters. Based on the similarity matrix U_{3D}^{2D}, the Hungarian algorithm is used to obtain the corresponding one-to-one matching. According to the matching threshold θ_{fusion}, both D_t^{3d} and D_t^{2d} are divided into two non-overlapping sets: $D_t^{3d} = {}^{both}D_t^{3d} \cup {}^{only}D_t^{3d}$ and $D_t^{2d} = {}^{both}D_t^{2d} \cup {}^{only}D_t^{2d}$.

Based on ${}^{only}D_t^{2d}$, the point cloud P is projected onto the image plane, and points within each 2D bounding

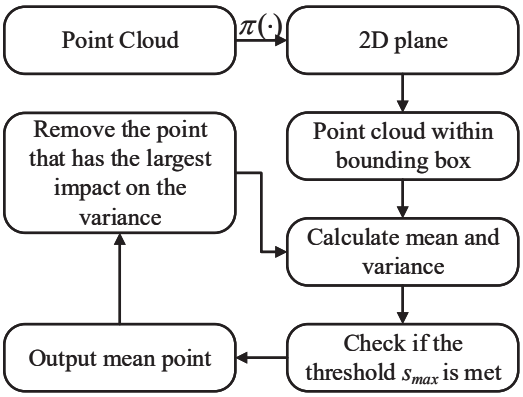


Fig. 3. Clustering Flowchart

box are extracted. To reduce noise and irregularities, the point closest to the box center is selected as the reference origin, and all points are transformed to this local coordinate frame. The center of mass is computed, and the mean squared deviation (MSD), defined as the average squared Euclidean distance to the center, is used to quantify spatial dispersion. If the MSD exceeds a predefined threshold s_{max}, the farthest point is iteratively removed and the MSD recalculated until the condition is satisfied. The final center is regarded as the estimated 3D position of the target. Accordingly, ${}^{only}D_t^{2d}$ is refined as ${}^{only}D_t^{2d} = {}^{point}D_t^{2d} \cup {}^{pixel}D_t^{2d}$, where ${}^{point}D_t^{2d}$ includes 2D detections with point cloud refinement, and ${}^{pixel}D_t^{2d}$ includes detections based solely on image bounding boxes. An illustration is provided in Fig. 3.

B. Multi-modality Motion Estimation Module (MMEM)

The motion state of the target is a critical clue of MOT, which has been extensively studied [2], [10], [11]. However, existing methods often focus on motion representation within a single modality, overlooking the correlation between multiple modalities. To address this limitation, MMME is proposed to select the most suitable motion strategy from different modalities based on the current state of the target.

Based on the motion states $\mathbf{X}_t = \{X_t^1, X_t^2, \ldots, X_t^M\}$, a motion model is constructed, where M denotes the number of trajectories. Each state $X_t^j = \{{}^{3D}X_t^j, {}^{2D}X_t^j\}$ incorporates both 2D and 3D information. Specifically, ${}^{3D}Xt^j = [x, y, z, \dot{x}, \dot{y}, \dot{z}, \ddot{x}, \ddot{y}, \ddot{z}]$ and ${}^{2D}Xt^j = [u, v, w, a, \dot{u}, \dot{v}, \dot{w}, \dot{a}]$, where (x, y, z), $(\dot{x}, \dot{y}, \dot{z})$, and $(\ddot{x}, \ddot{y}, \ddot{z})$ represent the object's 3D position, velocity, and acceleration, respectively. (u, v, w, a) denote the pixel coordinates of the bounding box center, its width, and aspect ratio, while $(\dot{u}, \dot{v}, \dot{w}, \dot{a})$ represent their corresponding rates of change. A 3D Kalman filter [4] is employed to estimate motion in 3D space for Dt^{3D} and ${}^{\text{point}}Dt^{2D}$, while a 2D Kalman filter [10] is used for ${}^{\text{pixel}}D_t^{2D}$ to estimate motion within the image plane.

After obtaining the predicted state, the core task in MOT is to find an observed state Z_t^i from the set of observations $\mathbf{Z}_t = \{Z_t^1, Z_t^2, \ldots, Z_t^M\}$ that meets the association criteria, and associate it with the predicted state $\hat{X}_t^j$, where M is the number of detections. If a match is found, the target state X_t^j is updated accordingly. For the 3D modality, the prediction

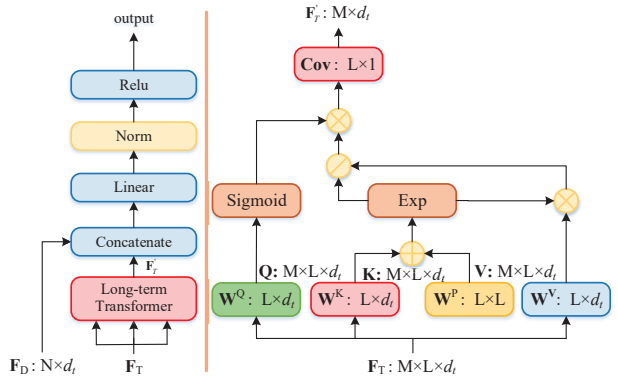


Fig. 4. Illustration of Long-term Matcher. Left: The structure of the Long-term Matcher module. Right: The computation graph of the Long-term Transformer.

and update steps are formulated as follows:

$$ {}^{3D}\hat{X}_t^j = {}^{3D}A\,{}^{3D}X_{t-1}^j, \quad (1) $$

$$ {}^{3D}\hat{S}_t^j = {}^{3D}A\,{}^{3D}S_{t-1}^j\,{}^{3D}A^T + Q, \quad (2) $$

$$ {}^{3D}X_t^j = {}^{3D}\hat{X}_t^j + K_t^{3D}({}^{3D}Z_t^i - {}^{3D}H\,{}^{3D}\hat{X}_t^j), \quad (3) $$

where $Q \in \mathbb{R}^{9\times9}$ is the covariance matrix of the state function, ${}^{3D}\mathrm{A} = \begin{bmatrix} E_{3\times3} & \delta E_{3\times3} & \frac{1}{2}\delta^2 E_{3\times3} \\ O_{3\times3} & E_{3\times3} & \delta E_{3\times3} \\ O_{3\times3} & O_{3\times3} & E_{3\times3} \end{bmatrix}$, ${}^{3D}\mathrm{H} = [\ E_{3\times3} \quad O_{3\times3} \quad O_{3\times3}\ ]$, $S_{t-1} \in \mathbb{R}^{9\times9}$ represents the error covariance at discrete time $t-1$, and $\hat{S}_t \in \mathbb{R}^{9\times9}$ denotes the predicted error covariance at discrete time t. K_t^{3D} is the 3D Kalman gain. Similar to the 3D motion state, ${}^{pixel}D_t^{2d}$ is updated using the following formula:

$$ {}^{2D}X_t^j = {}^{2D}\hat{X}_t^j + K_t^{2D}({}^{2D}Z_t^i - {}^{2D}H\,{}^{2D}\hat{X}_t^j), \quad (4) $$

where ${}^{2D}\mathrm{A} = \begin{bmatrix} E_{4\times4} & E_{4\times4} \\ O_{4\times4} & E_{4\times4} \end{bmatrix}$ and ${}^{2D}\mathrm{H} = [\ E_{4\times4} \quad O_{4\times4}\ ]$. By updating trajectory states with both 2D and 3D detections, the object's motion is jointly modeled in both spatial domains. In the subsequent association stage, the two modalities are processed independently to fully leverage their respective advantages.

C. Long-term Appearance Module (LAM)

Although the spatial relationship is established, it is susceptible to disruption by the irregular motion of the target, which is effectively addressed through the image texture. However, existing methods [5], [13] depend on the texture features from a single moment as the appearance representation of the target, often overlooking the fact that historical data is a crucial attribute of the trajectory.

To address this problem, LAM is proposed, and its structure is shown in Fig. 4. Unlike the previous method [3], [8], which uses the re-id branch as the identity representation of the target. LAM draws on [16] and directly outputs the similarity matrix $\mathbf{S_a}$ of detection and trajectory. First, image data spanning L consecutive frames along the temporal dimension is used as input. The trajectory and its associated appearance features are then modeled using a long-term matcher. To fuse the historical trajectory information, a long-term transformer is employed. Inspired by [17], the

temporal dimension is encoded with a learnable positional parameter $\mathbf{W}^P \in \mathbb{R}^{L\times L}$. The appearance features at multiple moments are then taken as input and multiplied by the key, query, and value, along with the corresponding projection matrices $\mathbf{W}^Q, \mathbf{W}^K, \mathbf{W}^V \in \mathbb{R}^{d_t\times d_t}$. Specifically, given the input trajectory feature $\mathbf{F}_T \in \mathbb{R}^{M\times L\times d_t}$, the output feature $\mathbf{F}_\mathbf{T}^{'} \in \mathbb{R}^{M\times d_t}$ is the fusion of position encoding and appearance features across multiple moments:

$$\mathbf{K}_{t'} = \frac{\exp((\mathbf{F}_T\mathbf{W}^K)_{t'} + \mathbf{W}^P_{t,t'})}{\sum_{t'=1}^{L}\exp((\mathbf{F}_t\mathbf{W}^K)_{t'} + \mathbf{W}^P_{t,t'})}, \quad (5)$$

where $\alpha'_t = \sigma_Q(\mathbf{Q}_t) \odot \mathbf{K}_{t'}, \mathbf{F}_\mathbf{T}^{'} = \sum_{t'=1}^{T}\alpha_{t'}\mathbf{V}_{t'}$, t' represents the data of the corresponding temporal dimension. Then, the trajectory features across multiple time steps are encoded into an embedding $\mathbf{F}'_T$, which is fused with the appearance features of the detections $\mathbf{F}_D \in \mathbb{R}^{N\times d_t}$ to produce $\mathbf{S}_a$, where M and N denote the number of trajectories and detections, respectively.

The loss function for appearance similarity across multiple moments consists of two components. The first one is the association formed using the data from a single frame, and the second component is the association loss derived from long-term appearance fusion.

$$\begin{aligned}
\mathcal{L}_{\text{single}} &= \frac{1}{L}\sum_{t=1}^{L}\sum_{i=1}^{N_{max}+1}\sum_{j=1}^{N_{max}+1} A^*[i,j]\log(\mathbf{S}^*_\mathbf{a}[i,j]),\\
\mathcal{L}_{\text{fusion}} &= \sum_{i=1}^{N_{max}+1}\sum_{j=1}^{N_{max}+1} A^*[i,j]\log(\mathbf{S}_\mathbf{a}[i,j]), \quad (6)\\
\mathcal{L}_{\text{joint}} &= \mathcal{L}_{\text{single}} + \mathcal{L}_{\text{fusion}},
\end{aligned}$$

where A^* represents the ground truth matching matrix. For the element $\{i,j\} \in A^*$, If the trajectory i and detection j correspond to the same object, the entry is set to 1; otherwise, it is set to 0. $\mathbf{S_a}^*$ denotes the similarity matrix computed from a single frame. To enhance the robustness of the algorithm, the data are expanded to N_{max} dimensions during training, with random adjustments applied to object positions to prevent the network from converging to local minima. Subsequently, both row-wise and column-wise Softmax operations are applied to $\mathbf{S_a}^*$ to normalize the similarity scores into association probabilities, thereby enabling joint optimization over both matched and unmatched pairs.

D. Multi-hierarchical Data Association (MHDA)

Through the above state modeling, the corresponding features offer essential clues for association. However, existing methods [2], [11] focus only on scenarios where either a single modality or both modalities are present, while neglecting associations when the modalities are asynchronous. To address this issue, this paper divides the association into three stages to enhance the information complementarity between the modalities.

First stage data association. Since spatial information can effectively mitigate the impact of occlusion, the first stage of association prioritizes matching detected instances with existing tracks. Specifically, the sets D_t^{3d} and ${}^{point}D_t^{2d}$ are used to compute association similarity based on 3D euclidean distance. At this stage, a higher confidence threshold is applied to jointly observed targets, which are then incorporated into the similarity matching process:

$$\mathbf{S^{3D}_m}[i,j] = \frac{\|{}^{3D}Z_t^i - {}^{3D}\hat{X}_t^j\|}{\alpha_{fusion}}, \quad (7)$$

where $\alpha_{fusion} = \frac{\alpha_{3D}+\alpha_{2D}}{2}$, α_{3D} and α_{2D} represent the detection confidences of 3D and 2D, respectively, each of which is normalized to a range between 0 and 1. The successfully associated trajectory set $\mathbf{T}_t^m$ is updated according to part of subsection III-B, while the remaining set $\mathbf{T}_t^u$ proceeds to the second stage of association.

Second stage data association. In this stage, ${}^{pixel}D_t^{2d}$ is merged with the remaining unmatched observations from the previous step and associated with $\mathbf{T}_t^u$ using the $\mathbf{S}_a$. However, due to threshold constraints, certain instances may lie close in the image plane but fail to meet the similarity criterion. To address this, 2D motion estimation is additionally employed to refine the similarity calculation.

Third stage data association. This stage primarily handles the reappearance of previously lost tracks and the

Algorithm 1 IMH-MOT algorithm

Input: Video sequence as ordered list $I = \{i_0, i_1, \cdots, i_{T-1}\}$ of images and $L = \{l_0, l_1, \cdots, l_{T-1}\}$ of Point Clouds.
Output: Set of object trajectories $\mathcal{T} = \{T_1, T_2, \cdots, T_k\}$ with $T_k = \{\mathbf{S}^k_{t_1}, \mathbf{S}^k_{t_2}, \cdots, \mathbf{S}^k_{t_N} \mid 0 \le t_1, \cdots, t_N \le T-1$ as a list of ordered object bounding boxes $\mathbf{S}^k_t = (D_t^{3d}, D_t^{2d}, \mathbf{F}_T)$.

$\mathcal{T}, \mathcal{T}_{\text{active}}, \mathcal{T}_{\text{remove}} \leftarrow \emptyset$;
for each $i_t, l_t \in I, L$ **do**
  MMEM.predict();
  $D_t^{3d}, D_t^{2d} \leftarrow detector.detections(i_t, l_t)$;
  $D_t^{3d}, {}^{point}D_t^{2d}, {}^{pixel}D_t^{2d} \leftarrow$MMAM$(D_t^{3d}, D_t^{2d})$;
  First stage:
  $\mathbf{S^{3D}_m} \leftarrow$Euclidean_distance$(D_t^{3d}, {}^{point}D_t^{2d}, \mathcal{T}_{\text{active}})$;
  $m_1, ut_1, ud_1 \leftarrow$Assignment$(S_m^{3D})$;
  Second stage:
  $\mathbf{S^*_a} \leftarrow$LAM$(ud_1, ut_1)$;
  $m_2, ut_2, ud_2 \leftarrow$Assignment$(\mathbf{S^*_a})$;
  $\mathbf{S^{2D}_m} \leftarrow$IOUs$(ud_2, ut_2)$;
  $m_3, ut_3, ud_3 \leftarrow$Assignment$(\mathbf{S^{2D}_m})$;
  Third stage:
  $\mathbf{S_a} \leftarrow$LAM$(ud_3, \mathcal{T}_{\text{remove}})$;
  $m_4, ut_4, ud_4 \leftarrow$Assignment$(\mathbf{S^*_a})$;
  for each $T_t \in \mathcal{T}_{\text{active}}$; **do**
    if $T_t.age - T_t.match\_age > \lambda_{new}$ **then**
      $\mathcal{T}_{\text{active}} \leftarrow \mathcal{T}_{\text{active}} - \{T_t\}$;
    end if
    if $T_t.lost\_num > \lambda_{remove}$ **then**
      $\mathcal{T}_{\text{active}} \leftarrow \mathcal{T}_{\text{active}} - \{T_t\}$;
      $\mathcal{T}_{\text{remove}} \leftarrow \mathcal{T}_{\text{remove}} + \{T_t\}$;
    end if
  end for
  for each $ud \in ud_4$; **do**
    $\mathcal{T}_{\text{active}} \leftarrow \mathcal{T}_{\text{active}} + ud$;
  end for
  Update$(\mathcal{T}_{\text{active}})$
end for
$\mathcal{T} \leftarrow \mathcal{T} + \mathcal{T}_{active}$;

TABLE I

IMH-MOT AND YONTD WITH SAME DETECTOR

Detector	HOTA↑	MOTA↑	MOTP↑	IDSW↓	FN↓	FP↓
PointRCNN [18]	54.25	48.34	87.17	16	4627	776
PVRCNN [19]	75.60	76.13	87.68	18	566	1757
Casa [20]	76.36	80.53	89.08	6	1411	493
PointRCNN [18]	73.53	80.79	86.78	618	654	612
PVRCNN [19]	76.03	83.14	88.71	480	550	624
Casa [20]	77.87	84.24	89.21	381	404	761

The results for YONTD are depicted in purple, while the results for IMH-MOT are shown in green

TABLE II

COMPARATIVE SUMMARY OF CAPABILITIES ACROSS METHODS

Method	No-overlapping-view	Occlusion	Sparse points	Long-term
mmMOT [22]	✗	✗	✗	✗
EagerMOT [1]	✗	✗	✗	✗
DeepFusion [14]	✗	✓	✓	✗
YONTD [15]	✗	✓	✓	✓
IMH-MOT†(ours)	✓	✓	✓	✓

initialization of new ones. Since unassociated tracks may be lost due to irregular motion or occlusion, appearance cues are crucial for re-identification. If re-association fails, the track is discarded. To reduce false positives, a new track must be successfully associated across multiple frames before being confirmed as valid.

By integrating the aforementioned modules, we implement an online multi-modal MOT framework, termed IMH-MOT. The complete workflow is summarized in Algorithm 1.

IV. EXPERIMENTAL RESULTS

A. Settings

1) Datasets: The KITTI benchmark [21] is widely recognized for evaluating autonomous driving and related technologies. All experiments in this paper are conducted using the KITTI MOT dataset, which includes 21 training sequences and 29 test sequences. Since the official dataset does not provide a training/validation split, follow the approach in [4] and select sequences 01, 06, 08, 10, 12, 13, 14, 15, 16, 18, and 19 as the validation set, with the remaining sequences used for training.

2) Evaluation Metrics: The performance of the proposed method is evaluated on the KITTI benchmark using standard MOT metrics from CLEAR MOT [23], including Multi-Object Tracking Accuracy (MOTA), Multi-Object Tracking Precision (MOTP), ID Switches (IDSW), False Positives (FP), False Negatives (FN). In addition to these metrics, High-Order Tracking Accuracy (HOTA), Detection Accuracy (DetA), and Association Accuracy (AssA) [24] are also adopted, providing a more comprehensive understanding of the tracking performance compared to traditional metrics.

B. Implementation Details

The model proposed in Section III-C is trained on 6 RTX 3090 GPUs using the ADAM optimizer with a learning rate of 1.25×10^{-4}. All code implementations are carried out using PyTorch 2.0.0.

1) 3D object detection: For evaluation on the KITTI test set, the 3D detection results generated by Point-GNN [25] and Virconv [26]. In the ablation study, the module using the results from Point-GNN, which is employed by our baseline method EagerMOT [1].

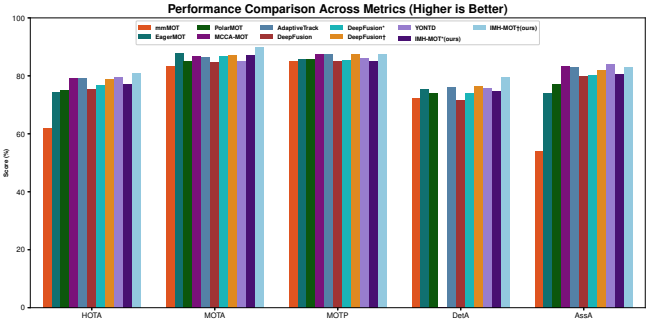
2) 2D object detection: For evaluation and ablation experiments on the KITTI test and validation sets, the 2D detection results generated by RRC [27], which is employed by our baseline method EagerMOT [1].

TABLE III

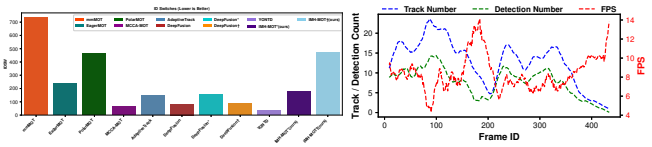
COMPARISON WITH THE STATE OF ART METHODS ON KITTI BENCHMARK

Method	Input	HOTA↑	MOTA↑	MOTP↑	IDSW↓	DetA↑	AssA↑
mmMOT [6]	C/LD	62.05	83.23	85.03	733	72.29	54.02
EagerMOT* [1]	C/LD	74.39	87.82	85.69	239	75.27	74.16
PolarMOT [12]	LD	75.16	85.08	85.63	462	73.94	76.95
MCCA-MOT [9]	C/LD	79.31	86.71	87.51	66	/	83.49
AdaptiveTrack [8]	C/LD	79.23	86.60	87.52	145	76.07	83.13
DeepFusion [14]	C/LD	75.46	84.63	85.02	84	71.54	80.05
DeepFusion* [14]	C/LD	76.67	86.88	85.34	154	73.90	80.19
DeepFusion† [14]	C/LD	78.91	87.20	**87.56**	90	76.44	82.10
YONTD [15]	C/LD	79.52	88.06	86.24	**37**	75.83	**84.01**
IMH-MOT*(ours)	C/LD	77.16	87.19	85.12	179	74.55	80.50
IMH-MOT†(ours)	C/LD	**80.90**	**89.73**	87.32	470	**79.42**	83.10

The green row represents the Baseline method. "*" and "†" denote the use of the same detection. "/" indicates the corresponding result is not available. "C" and "LD" represent camera and LiDAR, respectively



(a) Algorithm performance improves as the metrics increases



(b) IDSW

(c) FPS under varying numbers of targets

Fig. 5. Tracking Metrics and FPS

C. Results

The proposed method is compared with other state-of-the-art approaches [1], [9], [12], [14], [15], [22] on the KITTI test set, utilizing the evaluation metrics outlined in Section IV-A.2, a comparison of the capabilities of various algorithms is shown in Table II.

As shown in Table III, the algorithm proposed in this study achieves the highest HOTA, a comprehensive metric for evaluating current MOT algorithms. When evaluated against the baseline [1], the proposed method achieves +2.77% in HOTA, a decrease of 60 in IDSW, and +6.34% in AssA using the same detector. In comparison with the first proposed multimodal MOT method [22], the proposed method demonstrates +18.85% in HOTA, +6.5% in MOTA, and a reduction of -554 in IDSW. Relative to the TBD method [14], the proposed method exhibits gains of +5.44% in HOTA, +5.1% in MOTA, and +3.05% in AssA. Lastly,

(a) Inconsistent Observation Switching

(b) Occlusion

Fig. 6. Visualization of some special cases.

TABLE IV

TRACKING PERFORMANCE UNDER DIFFERENT MEMORY LENGTHS L

L	HOTA↑	MOTA↑	MOTP↑	IDSW↓	DetA↑	AssA↑
1	76.78	91.01	90.70	282	84.78	69.65
3	**78.63**	**91.34**	90.76	**256**	**85.44**	**72.47**
6	75.84	90.52	90.77	320	84.73	67.99
8	76.52	90.86	90.77	292	84.63	69.29
10	77.09	91.16	**90.77**	274	84.77	70.21

when compared with the SOTA joint detection and tracking method [15], the method shows an increase of +1.38% in HOTA, +1.67% in MOTA, and +8.62% in MT. A visual comparison is presented in Fig. 5 to more clearly demonstrate the effectiveness of the proposed algorithm. To mitigate the impact of detection on the algorithm, the same 3D detectors are employed to evaluate YONTD, and the validation set is partitioned based on YONTD. The results are presented in Table I.

To evaluate the runtime performance of the proposed algorithm under varying target densities, the relationship between FPS and the number of detections and trajectories is illustrated in Fig. 5c.

Visualization results are presented in Fig. 6 to illustrate the effectiveness of the proposed algorithm. At large viewing distances, LiDAR-based MOT fails to track the target, as depicted in Fig. 6a. However, the proposed algorithm ensures stable tracking and provides early warnings, irrespective of whether the viewing distance increases or decreases. Fig. 6b illustrates that camera-based MOT is particularly vulnerable to occlusion. Upon occlusion, the target's state changes, causing tracking failure. In contrast, the proposed method continues to track the target stably. These results substantiate the effectiveness of the proposed algorithm.

D. Ablation Study

To evaluate the effectiveness of the proposed module, an ablation studies is conducted on the KITTI [21] validation set.

1) Module ablation: The validation set is used to evaluate the performance of the algorithms in each module. When only the 2D detection and 2D motion model were included, the algorithm achieved 66.99% HOTA and 76.70% MOTA. With the inclusion of appearance features, HOTA increased by 7.85%, MOTA increased by 7.09%, and IDSW decreased by 583. Incorporating the mean of point cloud clustering as spatial information further improved performance, with HOTA rising by 2.53%, MOTA by 9.46%, and IDSW decreasing by 782, compared to the method using only 2D information. Building on this, the introduction of the point cloud sampling filtering algorithm proposed in this paper led to a further enhancement of HOTA by 5.09%, MOTA by 2.56%, and a reduction in IDSW by 200. Additionally, with the further incorporation of appearance features, HOTA increased by 11.64%, MOTA by 14.64%, and IDSW decreased by 1,185, compared to the initial state. The inclusion of 3D detection further improved the results. Upon separately adding various modules, compared to the baseline with only 2D and 3D motion information, HOTA improved by 1.42%, MOTA by 1.33%, and IDSW decreased by 38. These comparisons strongly support the effectiveness of the algorithm proposed in this paper.

2) Effect of memory length L: A comparison is conducted to assess the effectiveness of the proposed LAM across different memory lengths L. Notably, when 3D detection

TABLE V

ABLATION STUDY OF IMH-MOT ON THE KITTI VALIDATION SET

3D	2D	Points	Filter	AP	HOTA↑	MOTA↑	MOTP↑	IDSW↓	DetA↑	AssA↑
✗	✓	✗	✗	✗	66.99	76.70	**90.82**	1441	**85.46**	52.70
✗	✓	✗	✗	✓	74.84	83.79	90.81	858	85.62	65.54
✗	✓	✓	✗	✗	69.52	86.16	90.81	659	85.04	56.96
✗	✓	✓	✓	✗	74.61	88.72	85.33	459	85.33	65.41
✗	✓	✓	✓	✓	78.63	91.34	90.76	256	85.44	72.47
✓	✗	✗	✗	✗	78.46	86.99	87.65	63	77.12	80.09
✓	✗	✗	✗	✓	79.21	87.33	87.65	**35**	77.12	**81.61**
✓	✓	✗	✗	✗	78.69	87.67	87.65	249	79.28	78.37
✓	✓	✗	✗	✓	79.10	87.73	87.65	244	79.28	79.19
✓	✓	✓	✗	✗	79.20	87.84	87.65	235	79.28	78.56
✓	✓	✓	✓	✗	79.54	**89.12**	87.65	220	79.28	79.18
✓	✓	✓	✓	✓	**80.11**	89.00	87.65	211	79.28	80.52

"3D" and "2D" denote the use of only motion estimation models, "Points" indicates using the mean point within the 2D bounding box as the spatial center, "Filter" corresponds to the proposed MMAM, and "AP" represents the LAM

TABLE VI

IMH-MOT WITH DIFFERENT CLUSTERING THRESHOLD s_{MAX}

Thresh	HOTA↑	MOTA↑	MOTP↑	IDSW↓	FN↓	FP↓
0.5m	**78.63**	91.34	90.76	256	**409**	61
1.5m	78.36	91.41	90.77	249	411	60
2.5m	78.20	91.72	90.77	225	410	59
3.0m	78.51	**91.83**	**90.77**	**217**	411	**57**

is included, the algorithm's performance shows minimal variation with changes in the time dimension. Therefore, this section focuses on 2D evaluation to more clearly highlight the algorithm's performance. As shown in Table IV, the performance of the algorithm with memory lengths ranging from 1 to 10 is tested. Using appearance features at a single time point, the algorithm achieved a HOTA of 76.8%, a MOTA of 91.01%, and an IDSW of 282. As the memory length increases, the algorithm's performance gradually improves. The highest HOTA of 78.63% was achieved with a memory length of 3. Compared to the single-frame case, this resulted in a 1.85% improvement in HOTA and 0.31% increase in MOTA, and a reduction of 26 in IDSW.

3) Effect of clustering threshold s_{max}: When 3D detection is introduced, fluctuations in the index remain minimal. In order to more clearly illustrate the impact of different clustering thresholds $s_{\max}$, this section concentrates on 2D detections, with the reported value reflecting the physical-space distance. As shown in Table VI, with a clustering threshold of 0.5m, the algorithm achieves the highest HOTA of 78.63%. When the clustering threshold is set to 3m, the algorithm attains the highest MOTA of 91.83% and the lowest IDSW of 217.

V. CONCLUSIONS

In this paper, an effective and robust interactive multi-hierarchical multi-object tracking framework, IMH-MOT, was proposed to address the challenges of multi-modal MOT. A novel multi-modality alignment module was designed to enable precise instance point cloud sampling and provide target spatial information. To ensure stable motion association across modalities, a multi-modality motion estimation module was proposed to facilitate accurate motion prediction. Additionally, a long-term appearance module was introduced to achieve accurate embedding association by modeling long-term trajectory appearance features, thereby mitigating association errors caused by occlusion or irregular motion. Extensive experiments on the KITTI tracking benchmark demonstrated that the proposed method significantly outperformed state-of-the-art techniques in tracking accuracy.

REFERENCES

[1] A. Kim, A. Ošep, and L. Leal-Taixé, "Eagermot: 3d multi-object tracking via sensor fusion," in *2021 IEEE Int. Conf. Robot. Autom. (ICRA)*, 2021, pp. 11315–11321.

[2] L. D. Bella, Y. Lyu, B. Cornelis, and A. Munteanu, "Hybridtrack: A hybrid approach for robust multi-object tracking," *IEEE Rob. Autom. Lett.*, vol. 10, no. 7, pp. 7238–7245, 2025.

[3] M. Y. Lee, C. D. W. Lee, J. Li, and M. H. Ang, "Dino-mot: 3d multi-object tracking with visual foundation model for pedestrian re-identification using visual memory mechanism," *IEEE Rob. Autom. Lett.*, vol. 10, no. 2, pp. 1202–1208, 2025.

[4] 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 Trans. Intell. Transp. Syst.*, vol. 23, no. 6, pp. 5668–5677, 2022.

[5] S. Ding, L. Schneider, M. Cordts, and J. Gall, "Ada-track: End-to-end multi-camera 3d multi-object tracking with alternating detection and association," in *2024 IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2024, pp. 15184–15194.

[6] W. Zhang, H. Zhou, S. Sun, Z. Wang, J. Shi, and C. C. Loy, "Robust multi-modality multi-object tracking," in *2019 IEEE Int. Conf. Comput. Vis. (ICCV)*, 2019, pp. 2365–2374.

[7] W. Zeng, J. Fan, X. Tian, H. Chu, and B. Gao, "Fusiontrack: An online 3d multi-object tracking framework based on camera-lidar fusion," in *2024 IEEE/RSJ Int. Conf. Intell. Rob. Syst. (IROS)*, 2024, pp. 4920–4925.

[8] M. Liu, M. Wu, W. Wang, P. Liu, K. Chang, M. Li, and C. Piao, "Adaptive searching range-based data association for multi-object tracking with multi-information fusion," *IEEE Trans. Intell. Transp. Syst.*, pp. 1–13, 2025.

[9] H. Li, H. Liu, Z. Du, Z. Chen, and Y. Tao, "Mcca-mot: Multimodal collaboration-guided cascade association network for 3d multi-object tracking," *IEEE Trans. Intell. Transp. Syst.*, vol. 26, no. 1, pp. 974–989, 2025.

[10] A. Bewley, Z. Ge, L. Ott, F. Ramos, and B. Upcroft, "Simple online and realtime tracking," in *2016 IEEE Int. Conf. Image Process (ICIP)*, 2016, pp. 3464–3468.

[11] H. Li, Z. Wang, W. Kong, and X. Zhang, "Selectmot: Improving data association in multiple object tracking via quality-aware bounding box selection," *IEEE Sensors J.*, pp. 1–1, 2025.

[12] A. Kim, G. Brasó, A. Ošep, and L. Leal-Taixé, "Polarmot: How far can geometric relations take us in 3d multi-object tracking?" in *2022 Eur. Conf. Comput. Vis. (ECCV)*, 2022, pp. 41–58.

[13] L. Wang, X. Zhang, W. Qin, X. Li, J. Gao, L. Yang, Z. Li, J. Li, L. Zhu, H. Wang, and H. Liu, "Camo-mot: Combined appearance-motion optimization for 3d multi-object tracking with camera-lidar fusion," *IEEE Trans. Intell. Transp. Syst.*, vol. 24, no. 11, pp. 11981–11996, 2023.

[14] 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 Rob. Autom. Lett.*, vol. 7, no. 3, pp. 8260–8267, 2022.

[15] X. Wang, C. Fu, J. He, M. Huang, T. Meng, S. Zhang, H. Zhou, Z. Xu, and C. Zhang, "A multi-modal fusion-based 3d multi-object tracking framework with joint detection," *IEEE Rob. Autom. Lett.*, vol. 10, no. 1, pp. 532–539, 2025.

[16] S. Sun, N. Akhtar, H. Song, A. Mian, and M. Shah, "Deep affinity network for multiple object tracking," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 1, pp. 104–119, 2021.

[17] S. Zhai, W. Talbott, N. Srivastava, C. Huang, H. Goh, R. Zhang, and J. Susskind, "An attention free transformer," 2021. [Online]. Available: https://arxiv.org/abs/2105.14103

[18] S. Shi, X. Wang, and H. Li, "Pointrcnn: 3d object proposal generation and detection from point cloud," in *2019 IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2019, pp. 770–779.

[19] S. Shi, C. Guo, L. Jiang, Z. Wang, J. Shi, X. Wang, and H. Li, "Pv-rcnn: Point-voxel feature set abstraction for 3d object detection," in *2020 IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2020, pp. 10526–10535.

[20] H. Wu, J. Deng, C. Wen, X. Li, C. Wang, and J. Li, "Casa: A cascade attention network for 3-d object detection from lidar point clouds," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–11, 2022.

[21] A. Geiger, P. Lenz, and R. Urtasun, "Are we ready for autonomous driving? the kitti vision benchmark suite," in *2012 IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2012, pp. 3354–3361.

[22] W. Zhang, H. Zhou, S. Sun, Z. Wang, J. Shi, and C. C. Loy, "Robust multi-modality multi-object tracking," in *2019 IEEE Int. Conf. Comput. Vis. (ICCV)*, 2019, pp. 2365–2374.

[23] B. Keni and S. Rainer, "Evaluating multiple object tracking performance: The clear mot metrics," *EURASIP J. on Image Video Process.*, vol. 2008, no. 1, pp. 246309–246319, 2008.

[24] J. Luiten, A. Osep, P. Dendorfer, P. Torr, A. Geiger, L. Leal-Taixe, and B. Leibe, "Hota: A higher order metric for evaluating multi-object tracking," *Int. J. Comput. Vis.*, vol. 129, no. 2, pp. 548–578, 2021.

[25] W. Shi and R. Rajkumar, "Point-gnn: Graph neural network for 3d object detection in a point cloud," in *2020 IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2020, pp. 1708–1716.

[26] H. Wu, C. Wen, S. Shi, X. Li, and C. Wang, "Virtual sparse convolution for multimodal 3d object detection," in *2023 IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2023, pp. 21653–21662.

[27] J. Ren, X. Chen, J. Liu, W. Sun, J. Pang, Q. Yan, Y.-W. Tai, and L. Xu, "Accurate single stage detector using recurrent rolling convolution," in *2017 IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2017, pp. 752–760.