machines

MDPI

Article

Adaptive Multi-Level 3D Multi-Object Tracking with Transformer-Based Association and Scene-Aware Thresholds for Autonomous Driving

Yongze Zhang [1,2,3], Feipeng Da [1,2,*] and Haocheng Zhou [3]

[1] School of Automation, Southeast University, Nanjing 210018, China
[2] Key Laboratory of Measurement and Control of Complex Systems of Engineering, Southeast University, Nanjing 210018, China
[3] Jiangsu Electric Power Information Technology Co., Ltd., Nanjing 210018, China
* Correspondence: dafp@seu.edu.cn

Abstract

3D multi-object tracking (MOT) for autonomous driving remains challenging due to frequent identity switches in crowded scenes, trajectory fragmentation during occlusions, and the difficulty of adapting association strategies to varying scene complexities. While existing methods rely on fixed geometric or appearance-based associations, they struggle to handle ambiguous cases and detection failures. We present an adaptive multi-level 3D MOT framework that achieves robust tracking through three key innovations: (1) multi-granularity temporal modeling that captures both fine-grained short-term motion and coarse long-term trends via dual-scale spatio-temporal attention, enabling accurate motion prediction across different object dynamics; (2) Transformer-based Appearance Association that employs cross-attention to model global inter-object relationships, resolving ambiguous associations in crowded scenarios where geometric cues alone fail; and (3) scene-adaptive learned thresholds that automatically adjust association strictness based on object density, motion complexity, and occlusion levels, avoiding the one-size-fits-all limitations of fixed thresholds. Our hierarchical four-level tracking strategy progressively handles cases from easy geometric matching (Level 1) to complex interval-frame recovery (Level 4), with SOT-based virtual detection generation bridging detector failures. Extensive experiments on the nuScenes benchmark demonstrate state-of-the-art performance.

Keywords: 3D multi-object tracking; autonomous driving; transformer; scene-adaptive thresholds; temporal modeling

Check for updates

Received: 25 March 2026
Revised: 19 April 2026
Accepted: 21 April 2026
Published: 23 April 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

1. Introduction

The rapid advancement of autonomous driving technology has created an urgent demand for robust perception systems that can reliably understand dynamic traffic environments [1]. In particular, safe motion planning critically depends on accurate trajectory prediction and interaction-aware reasoning among surrounding agents [2], underscoring the importance of reliable multi-object tracking as a foundational component. 3D multi-object tracking (MOT) is a fundamental capability for autonomous driving systems, ensuring the safe navigation of the ego-vehicle by maintaining consistent identities and trajectories of surrounding agents (e.g., vehicles, pedestrians, cyclists) over time [3–5]. While 3D object detection provides instantaneous localization, MOT faces unique temporal challenges: (1) identity switches (IDsws) frequently occur in crowded intersections where

objects with similar appearance and motion are closely spaced; (2) trajectory fragmentation (FRAG) arises when detectors fail during occlusion, causing tracks to terminate prematurely; and (3) lack of scene adaptability, as most methods utilize fixed parameters that fail to generalize across diverse scenarios, from sparse highways to dense urban centers.

Existing 3D MOT methodologies can be broadly categorized into three paradigms, as illustrated in Figure 1.

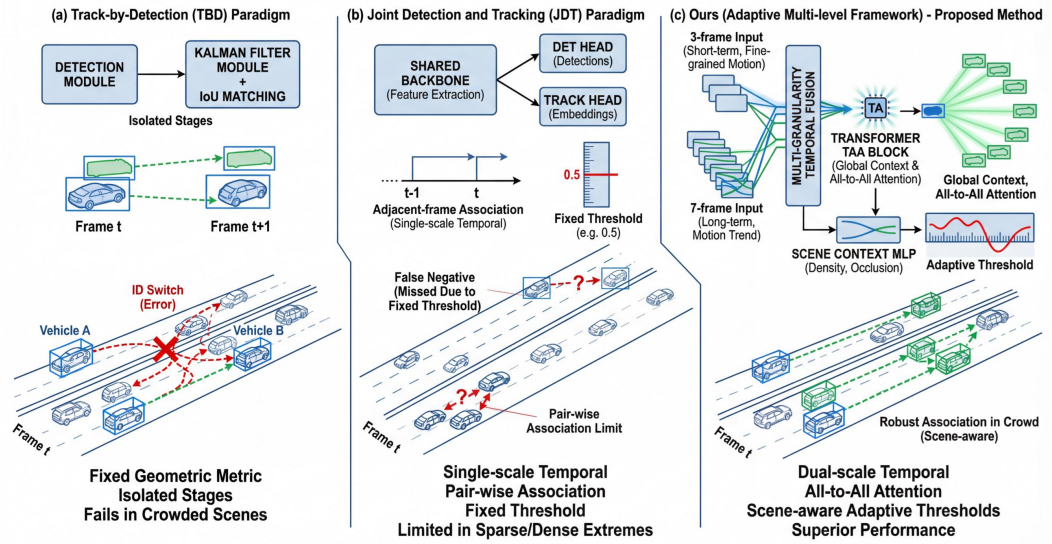


Figure 1. Evolution of 3D MOT paradigms. (**a**) Track-by-Detection (TBD): Traditional methods treat detection and tracking as isolated stages, relying on fixed geometric metrics that fail to distinguish closely spaced objects, leading to identity switches (IDsws). (**b**) Joint Detection and Tracking (JDT): Recent approaches share features for efficiency but rely on single-scale temporal modeling and fixed association thresholds, limiting adaptability to varying scene dynamics. (**c**) Our Method (JMM3DDT): We propose a unified tracking-centric framework featuring: (1) multi-granularity temporal modeling (three-frame dense and seven-frame sparse) to capture diverse motion patterns; (2) Transformer-based Appearance Association (TAA) using cross-attention to resolve global ambiguities; and (3) scene-adaptive learned thresholds (SALTs) that dynamically adjust strictness based on density and occlusion. This design achievesup to 77% fewer IDsws compared to baseline methods.

Track-by-Detection (TBD) methods [3,6–8] treat detection and tracking as independent stages (Figure 1a). They typically employ Kalman filters for state estimation and Hungarian matching based on geometric metrics (e.g., 3D IoU). While computationally efficient, TBD suffers from error propagation; once the geometric association is ambiguous—common in crowded scenes—identity switches are inevitable. Although some recent works incorporate appearance cues [9,10], they often rely on pair-wise distance metrics (e.g., cosine similarity), which fail to capture the global inter-object context needed to resolve complex ambiguities.

To improve efficiency, Joint Detection and Tracking (JDT) methods [7,11–13] integrate tracking into the detection network (Figure 1b), sharing feature backbones and predicting motion offsets. Despite their speed, most JDT approaches rely on single-scale temporal modeling (e.g., concatenating adjacent frames) and utilize fixed association thresholds. This "one-size-fits-all" strategy is suboptimal: a strict threshold needed for a crowded intersection often causes missed matches in a sparse highway scenario.

Recent advances in computer vision have demonstrated the efficacy of Transformer-based models for capturing long-range dependencies [14] and adaptive task learning [15–17]. Furthermore, emerging research in 2024 has highlighted the importance of multi-granularity temporal representation for complex motion reasoning [18,19]. Inspired by these developments, we argue that a robust 3D MOT system requires: (1) multi-scale temporal modeling to handle diverse dynamics (e.g., fast pedestrians vs. slow trucks); (2) context-aware association

that considers all objects globally rather than in isolation; (3) scene-adaptive thresholds that adjust to real-time density and occlusion; and (4) hierarchical tracking to progressively resolve easy-to-hard cases.

To this end, we propose JMM3DDT, an adaptive multi-level 3D MOT framework (Figure 1c) that directly addresses the limitations of fixed, single-scale approaches. Our tracking-centric design treats detection as a supporting component, optimized specifically for association robustness. Our main contributions are:

- **Multi-granularity Cross-modal Spatio-Temporal Attention (CSTA):** We propose a dual-scale temporal modeling approach fusing camera and LiDAR features. By processing parallel short-term (three-frame dense) and long-term (seven-frame sparse) branches, we capture both fine-grained motion steps and coarse trajectory trends, reducing IDsws by 23.8% compared to the baseline without temporal modeling.
- **Transformer-based Appearance Association (TAA):** We introduce a cross-attention mechanism for association. Unlike pair-wise metrics, TAA models the global relationship between an unmatched detection and *all* trajectories simultaneously. This context-aware matching achieves a 2.6% AMOTA improvement and reduces IDsws by 30.4% over geometric-only association.
- **Scene-adaptive Learned Thresholds (SALTs):** We replace heuristic fixed thresholds with a learnable module that predicts optimal association strictness based on real-time scene context (density, velocity, occlusion). This adaptive mechanism simultaneously reduces false positives in crowded scenes and false negatives in sparse ones.

A Unified View of the Three Components. Although CSTA, TAA, and SALTs are implemented as distinct modules, they are not independent building blocks—they jointly approximate a Bayesian data-association pipeline. Specifically, CSTA builds a reliable motion-and-appearance posterior $p(\mathbf{x}_t|\mathcal{O}_{1:t})$ that reduces the predictive uncertainty of each track; TAA refines the observation likelihood $p(\mathbf{o}_t|\mathbf{x}_t, \mathcal{T}_{active})$ through global cross-attention over all candidates; and SALTs parameterize the decision boundary of the maximum a posteriori (MAP) assignment by a scene-conditioned function $\hat{\tau}(\mathbf{s}_t)$. Training is carried out under a single objective $\mathcal{L}_{total}$ (Section 3.6), which couples these components through shared BEV features and end-to-end gradient flow. This unified view not only justifies the hierarchical design but also explains why the three modules complement rather than duplicate each other.

Fundamentally, the core innovation of JMM3DDT lies in its departure from the rigid, "one-size-fits-all" paradigms that have traditionally dominated 3D MOT. Rather than treating temporal modeling, feature association, and matching thresholds as isolated, static processes, our framework unifies them into a highly dynamic and scene-aware system. By seamlessly integrating multi-granularity temporal reasoning (CSTA) with global context-aware cross-attention (TAA) and real-time adaptive thresholding (SALT), the proposed methodology actively adjusts its tracking strategies to accommodate varying traffic densities, object velocities, and severe occlusions. This synergistic adaptability directly addresses the inherent fragility of conventional trackers, establishing a more resilient and cognitively aligned perception foundation for complex autonomous driving scenarios.

2. Related Works

We review related works from the perspective of tracking methodologies, focusing on association strategies, temporal modeling, and Transformer applications.

2.1. 3D Multi-Object Tracking Paradigms

3D MOT methods can be categorized by how they integrate detection and tracking. (1) Track-by-Detection (TBD)approaches [3,6,9] treat detection and tracking as separate

stages. AB3DMOT [3] pioneered this paradigm by extending 2D SORT to 3D space, using Kalman filtering for motion prediction and 3D IoU for association. SimpleTrack [6] and EagerMOT [9] further improved this by refining geometric metrics and fusing 2D–3D features. More recently, Fast-Poly [20] addresses rotational anisotropy via polyhedral representations, and MCTrack [21] proposes a unified TBD framework achieving strong performance across multiple benchmarks. While computationally efficient and modular, TBD methods suffer from irreversible error propagation: once detection fails or generates low-quality boxes, the tracker lacks mechanisms to recover, leading to a high number of identity switches. (2) Joint Detection and Tracking (JDT) methods [7,11–13] improve efficiency by sharing feature backbones between detection and tracking heads. CenterTrack [7] models tracking as a point-based offset regression task across adjacent frames. CenterPoint [11] extends this to 3D LiDAR data, becoming a standard baseline. Recent end-to-end approaches like Sparse4D v3 [13] and ADA-Track [22] further enhance this paradigm—the latter introducing alternating detection–association modules with edge-augmented cross-attention. Although JDT methods reduce computational redundancy, they typically rely on single-scale temporal cues (e.g., concatenation of two frames) and fixed association thresholds, limiting their adaptability to complex scene dynamics (e.g., varying speeds and occlusion levels). Our work advances the JDT paradigm by introducing multi-granularity temporal modeling and adaptive mechanisms.

2.2. Appearance-Based Association

Beyond geometric matching (IoU/L2 distance), appearance features are crucial for distinguishing objects with similar motion. Earlier works [23] used hand-crafted point-cloud descriptors. Recent sensor-fusion methods like UniTR [24] and CBMOT [25] learn appearance embeddings from camera–LiDAR features using metric learning (triplet or contrastive losses). OATracker [26] further introduces object-state-constrained affinity that adaptively combines geometric and motion cues to handle occlusion. However, these methods predominately compute affinity matrices using pair-wise distance metrics (e.g., cosine or Euclidean distance). This pair-wise approach processes each detection–trajectory pair independently, failing to capture the global context, i.e., how a detection relates to all other candidates simultaneously. In contrast, our Transformer-based Appearance Association (TAA) utilizes cross-attention to model global inter-object dependencies, effectively resolving ambiguities in crowded scenes.

2.3. Transformer-Based Tracking

Transformers [14] have revolutionized object tracking by modeling long-range dependencies. DETR [15] introduced set-based prediction, which was extended to 2D MOT by TrackFormer [16] and MOTR [17]. More recently, MOTRv3 [27] improved label assignment for end-to-end tracking, and Beyond MOT [28] extended tracking toward semantic scene understanding. In the 3D domain, TransFusion [12] uses Transformers for LiDAR-Camera fusion detection but still relies on standard matching for tracking. UVTR [29] explores voxel-based transformers. ADA-Track [22] proposes alternating detection and association with learned edge-augmented cross-attention. However, applying full end-to-end Transformer tracking in 3D is computationally expensive due to the large search space of voxel/BEV representations. Our approach strikes a balance by using Transformers specifically for the association stage on ROI-pooled features, rather than full-frame query propagation, ensuring both efficiency and global context awareness.

2.4. Temporal Modeling for Tracking

Robust tracking requires modeling motion dynamics. Classical methods rely on Kalman Filters [3], which assume linear motion and constant velocity, an assumption often

violated by pedestrians or turning vehicles. Learning-based approaches like BEVDet4D [30] and FlashOCC [18] aggregate multi-frame features to learn motion implicitly. TimeSformer [31] introduced divided space–time attention for video understanding, and VideoMamba [19] recently adapted State Space Models (SSMs) for efficient long-term temporal reasoning. DiffMOT [32] further introduced diffusion probabilistic models for nonlinear motion prediction in MOT. Meanwhile, DMSTrack [33] proposes a differentiable multi-sensor Kalman filter that learns observation noise from BEV features. However, existing 3D MOT methods usually adopt a single temporal scale (e.g., fixed three-frame window), which struggles to handle the diverse dynamics of traffic participants. Our multi-granularity CSTA explicitly models dual-scale patterns: a dense short-term branch for fine-grained motion and a sparse long-term branch for coarse trajectory trends.

3. Method

3.1. Overview of the Proposed Framework

The core philosophy of JMM3DDT is to establish a robust tracking system adaptable to varying scene complexities. As illustrated in Figure 2, our framework processes multi-modal data through three key innovative components:

1. **Multi-granularity Cross-modal Spatio-Temporal Attention (CSTA).** Shown in the blue block of Figure 2, we propose a dual-branch architecture to capture motion at different scales. Raw camera and LiDAR data are encoded into BEV space and processed by: a short-term branch (3 frames) for fine-grained motion (e.g., pedestrians), and a long-term branch (7 frames) for coarse trends (e.g., vehicles). These are fused to generate comprehensive spatio-temporal BEV Features. Subsequently, a standard detection head processes these features to yield 3D bounding boxes and embeddings, serving as the raw observations for the tracking pipeline.
2. **Scene-Adaptive Learned Thresholds (SALTs).** Illustrated in the green block, this module acts as the dynamic controller. It analyzes real-time scene statistics (object density ρ_t, occlusion ratio η_t, average velocity $\bar{v}_t$) and employs a lightweight MLP to predict optimal association thresholds ($\hat{\tau}^t_{geo}$, $\hat{\tau}^t_{app}$). This allows the system to automatically tighten constraints in crowded scenes and relax them in sparse environments.
3. **Hierarchical Multi-level Association Strategy.** The orange block depicts our coarse-to-fine tracking pipeline receiving the detections. We employ a four-level strategy: Level 1 geometric matching with adaptive thresholds handles easy cases; Level 2 Transformer-based Appearance Association (TAA) resolves conflicts by modeling global inter-object relationships; Level 3 SOT-based recovery generates virtual detections for occluded targets; and Level 4 interval-frame recovery stitches broken trajectories from extended occlusions.

Principled Justification of the Four-Level Hierarchy. The four levels are formally justified from a sequential decision-making perspective. Let $P(match \mid d_i, tr_j)$ denote the posterior matching probability for a detection–track pair. Levels 1–4 correspond to handling cases with progressively lower $P(match)$ confidence:

- Level 1 (Geometric): High-confidence associations where $P(match) \geq \hat{\tau}^t_{geo}$, resolved purely by spatial IoU.
- Level 2 (TAA): Ambiguous cases where geometric posterior alone is inconclusive ($P_{geo}(match) < \hat{\tau}^t_{geo}$), resolved by incorporating global appearance likelihood via cross-attention.
- Level 3 (SOT): Missing-observation regime where no detection is available for a track, so we transition from the matching problem to a generation problem, using the motion

prior and local feature correlation to hallucinate a virtual observation. The transition criterion is formally: $tr_j \in \mathcal{T}_{unmatch}$ after Levels 1–2 and $\text{age}(tr_j) \leq \delta_{sot_max}$.

- Level 4 (Recovery): Long-horizon re-identification, operating on lost tracks that exceed the SOT window, using stored appearance memory and relaxed temporal constraints.

Each level strictly operates on the residual set of the previous level, ensuring no redundancy. The four levels are necessary because they handle fundamentally different association regimes (high-confidence match, ambiguous match, missing observation, and long-term re-identification), each requiring a distinct modeling approach.

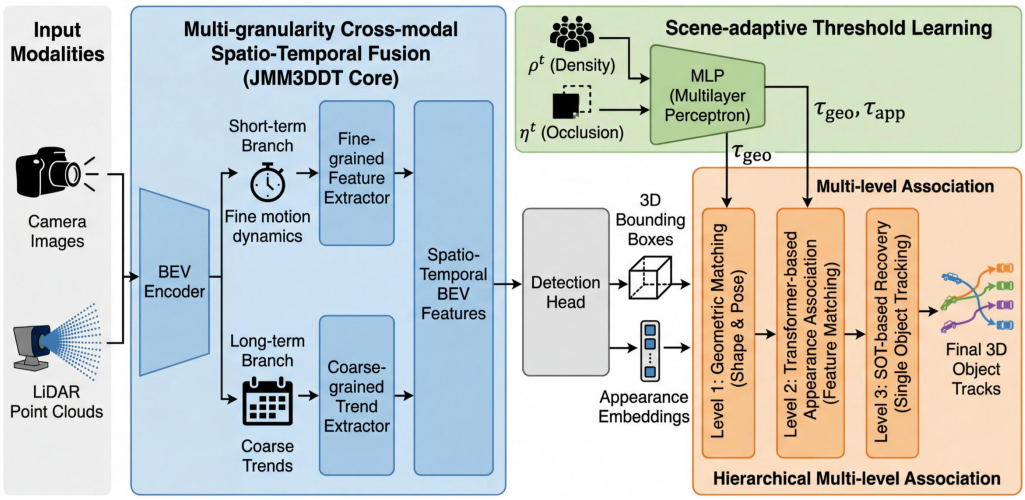


Figure 2. The overall architecture of JMM3DDT. The framework features multi-granularity CSTA (Blue) for feature extraction, followed by scene-adaptive threshold learning (Green) and hierarchical multi-level association (Orange) for robust tracking.

3.2. Multi-Granularity Cross-Modal Spatio-Temporal Attention (CSTA)

To effectively handle the diverse motion dynamics inherent in autonomous driving scenarios (e.g., static obstacles vs. fast-moving vehicles), we introduce the multi-granularity CSTA module. This module fuses camera and LiDAR features into a unified BEV representation and subsequently models temporal dependencies across two distinct scales.

Cross-modal BEV Encoding. Let $I_t \in \mathbb{R}^{H\times W\times 3}$ and $P_t \in \mathbb{R}^{N\times 3}$ denote the input image and point cloud at current frame t, respectively. We utilize a dual-branch backbone (e.g., ResNet and VoxelNet) to extract image features F^t_{cam} and LiDAR features F^t_{lidar}. These are projected into a shared Bird's-Eye-View (BEV) space via Lift-Splat-Shoot (LSS) and voxel flattening operations, respectively, and fused via concatenation and convolution to yield the frame-level BEV embedding $B_t \in \mathbb{R}^{X\times Y\times C}$.

Dual-branch Temporal Modeling. We process the sequence of BEV embeddings through two parallel attention branches:

1. Short-term Dense Branch: This branch captures fine-grained, instantaneous motion changes using a dense window of the past 3 frames: $\mathcal{T}_{short} = \{t, t-1, t-2\}$.
2. Long-term Sparse Branch: This branch captures coarse trajectory trends and ensures stability during extended occlusions using a sparse window of 7 frames with a stride of 2: $\mathcal{T}_{long} = \{t, t-2, \ldots, t-12\}$.

For each branch $k \in \{short, long\}$, we employ a Temporal Cross-Attention mechanism where the current frame embedding B_t serves as the *query* (Q), and the historical embeddings in $\mathcal{T}_k$ serve as *keys* (K) and *values* (V). To distinguish the temporal order of historical

frames while preserving spatial correspondence, we add a 1D learnable temporal positional encoding PE broadcast across the spatial dimensions:

$$Q = B_t W_Q, \quad K_i = (B_{t-i} + PE_i)W_K, \quad V_i = B_{t-i}W_V, \tag{1}$$

where $W_Q, W_K, W_V \in \mathbb{R}^{C \times d_k}$ are learnable projection matrices, d_k is the key dimension, and $t - i \in \mathcal{T}_k$. The multi-granularity feature F^t_{temp} is obtained by aggregating the outputs of both branches:

$$H_k = \text{Softmax}\left(\frac{QK^\top}{\sqrt{d_k}}\right)V, \quad F^t_{temp} = \text{Conv}_{2C\to C}([H_{short}; H_{long}]) + B_t, \tag{2}$$

where $[\cdot;\cdot]$ denotes channel-wise concatenation, and $\text{Conv}_{2C\to C}$ is a 1×1 convolution that reduces the concatenated $2C$-dimensional features back to C dimensions before the residual addition. F^t_{temp} is then fed into the detection head to regress 3D bounding boxes $\mathcal{D}_t = \{d_i\}_{i=1}^{N_{det}}$ and extract object appearance embeddings $\mathbf{e}_i^{app} \in \mathbb{R}^C$.

Connection to Tracking Objectives. The multi-granularity CSTA directly contributes to tracking performance through two complementary mechanisms. First, the short-term branch reduces motion prediction uncertainty: by attending to consecutive frames, it captures instantaneous velocity and acceleration changes, yielding more accurate Kalman filter predictions and thus higher IoU overlap at Level 1 (reducing IDsws). Second, the long-term branch enhances appearance feature consistency: by aggregating information across a wider temporal window, it produces temporally smoothed embeddings $\mathbf{e}_i^{app}$ that are more robust to transient appearance variations (e.g., lighting changes, partial occlusion), directly benefiting the TAA module at Level 2 and the recovery mechanism at Level 4. In essence, the short-term branch improves "where the object is" (geometric association quality), while the long-term branch improves "what the object looks like" (appearance association quality).

3.3. Scene-Adaptive Learned Thresholds (SALTs)

Traditional MOT methods rely on fixed association thresholds (e.g., IoU > 0.5), which leads to frequent identity switches (IDsws) in crowded scenes or missed associations in high-velocity scenarios. To address this, we propose SALTs, a lightweight module that dynamically predicts optimal association thresholds based on global scene contexts.

We define a scene context vector $\mathbf{s}_t = [\rho_t, \eta_t, \bar{v}_t] \in \mathbb{R}^3$, where

- ρ_t is the object density, defined as the number of detected objects per 100 m^2.
- η_t is the occlusion ratio, the fraction of tracked objects with predicted occlusion status (based on detector confidence variance).
- $\bar{v}_t$ is the average velocity, the mean speed of all active tracks in the previous frame.

To ensure numerical stability and balanced feature contribution across heterogeneous statistics, we apply z-score normalization to the raw scene context before feeding it into the MLP. The normalized input vector $\tilde{\mathbf{s}}_t$ is defined as:

$$\tilde{\mathbf{s}}_t = \left[\frac{\rho_t - \mu_\rho}{\sigma_\rho}, \eta_t, \frac{\bar{v}_t - \mu_v}{\sigma_v}\right], \tag{3}$$

where $\mu_\rho, \sigma_\rho, \mu_v, \sigma_v$ are global statistics computed from the training set. Importantly, these statistics are computed once from the training set and kept fixed (frozen) during both validation and testing, analogous to the running mean and variance in standard batch normalization. This ensures that no test-time information leaks into the normalization process. We further verified that replacing these frozen global statistics with per-batch statistics

yielded comparable results (within 0.2% AMOTA), confirming the robustness of this design choice. Note that $\eta_t \in [0, 1]$ is already bounded and does not require normalization.

The SALTs module, denoted as $\Phi(\cdot)$, is an MLP with three hidden layers (each of size 64) that maps $\tilde{\mathbf{s}}_t$ to a set of adaptive IoU/score thresholds:

$$[\hat{\tau}^t_{geo}, \hat{\tau}^t_{app}] = \sigma(\Phi(\tilde{\mathbf{s}}_t)) \cdot (\boldsymbol{\alpha}_{max} - \boldsymbol{\alpha}_{min}) + \boldsymbol{\alpha}_{min}, \tag{4}$$

where σ is the sigmoid function, and $\boldsymbol{\alpha}_{min} = [\alpha^{min}_{geo}, \alpha^{min}_{app}]$, $\boldsymbol{\alpha}_{max} = [\alpha^{max}_{geo}, \alpha^{max}_{app}]$ are vectors defining the valid range for IoU/score thresholds. Specifically, based on statistical validation of the dataset, we explicitly set the bounds to $\alpha_{geo} \in [0.1, 0.7]$ and $\alpha_{app} \in [0.3, 0.8]$ in our implementation. A higher predicted threshold enforces stricter matching constraints (requiring higher IoU overlap), which is desirable in dense scenes to prevent false associations. During training, we supervise Φ using a loss that minimizes the difference between predicted thresholds and "oracle" thresholds derived from ground-truth optimal assignment, ensuring strict thresholds in dense scenes (high ρ_t) and relaxed thresholds in sparse, fast-moving scenes (high $\bar{v}_t$).

3.4. Transformer-Based Appearance Association (TAA)

In crowded scenarios, geometric cues become unreliable due to bounding-box overlap. While previous works use pair-wise cosine similarity for appearance matching, they lack global context. Our TAA module utilizes a Transformer decoder to model the relationship between a detection and all active candidate tracks simultaneously.

Let $\mathbf{E}_{\mathcal{D}} \in \mathbb{R}^{N_{det} \times C}$ be the embeddings of unmatched detections and $\mathbf{E}_{\mathcal{T}} \in \mathbb{R}^{N_{trk} \times C}$ be the embeddings of active tracks. We treat detections as queries and tracks as keys/values. The Multi-Head Cross-Attention (MHCA) updates the detection embeddings by aggregating context from relevant tracks. For the ith head, we first compute the projections:

$$\mathbf{Q}_i = \mathbf{E}_{\mathcal{D}} W^i_Q, \quad \mathbf{K}_i = \mathbf{E}_{\mathcal{T}} W^i_K, \quad \mathbf{V}_i = \mathbf{E}_{\mathcal{T}} W^i_V, \tag{5}$$

where $W^i_Q, W^i_K, W^i_V \in \mathbb{R}^{C \times d_h}$ are per-head projections, and $d_h = C/h$. The attention output is computed as:

$$\text{Attn}_i = \text{Softmax}\left(\frac{\mathbf{Q}_i \mathbf{K}_i^\top}{\sqrt{d_h}}\right)\mathbf{V}_i, \tag{6}$$

where the Softmax operation is applied row-wise over the track dimension (i.e., each detection attends to all tracks). The refined detection embeddings are obtained via residual connection:

$$\tilde{\mathbf{E}}_{\mathcal{D}} = \text{LayerNorm}(\mathbf{E}_{\mathcal{D}} + \text{Concat}(\text{Attn}_1, \ldots, \text{Attn}_h) W_O), \tag{7}$$

where $W_O \in \mathbb{R}^{C \times C}$ is the output projection. To ensure the affinity matrix strictly falls within the $[0, 1]$ range for the subsequent cost matrix formulation (where cost is derived from the affinity, see Section 3.5), the cosine similarity between the refined detection embeddings and original track embeddings is linearly scaled:

$$\mathbf{A}_{TAA}(i, j) = \frac{1}{2}\left(\frac{\tilde{\mathbf{E}}_{\mathcal{D}}(i) \cdot \mathbf{E}_{\mathcal{T}}(j)^\top}{\|\tilde{\mathbf{E}}_{\mathcal{D}}(i)\|_2 \|\mathbf{E}_{\mathcal{T}}(j)\|_2} + 1\right). \tag{8}$$

Unlike standard pair-wise distance, the MHCA step allows detection features to be contextually refined by the distribution of existing tracks before similarity calculation, effectively highlighting discriminative features for ambiguous objects.

3.5. Hierarchical Four-Level Association Strategy

We integrate the above components into a coarse-to-fine four-level association strategy, summarized in Algorithm 1. Let $\mathcal{T}_{active}$ be the set of active tracks and $\mathcal{D}_t$ be the current detections.

Algorithm 1 JMM3DDT hierarchical association

Require: Detections $\mathcal{D}_t$, active tracks $\mathcal{T}_{active}$, lost tracks $\mathcal{T}_{lost}$
Require: Scene context $\mathbf{s}_t$, temporal features F_{temp}^t
Ensure: Updated tracks $\mathcal{T}_{active}$, new tracks, terminated tracks
// Scene-Adaptive Threshold Prediction
$[\hat{\tau}_{geo}^t, \hat{\tau}_{app}^t] \leftarrow \text{SALT}(\mathbf{s}_t)$ ▷ IoU/score thresholds
// Level 1: Geometric Matching (Score-based)
$\mathbf{S}_{geo} \leftarrow \text{ComputeIoUScore}(\mathcal{D}_t, \mathcal{T}_{active})$ ▷ $\mathbf{S}_{geo}(i,j) = \text{IoU}_{3D}(d_i, tr_j)$
$\mathbf{C}_{geo} \leftarrow 1 - \mathbf{S}_{geo}$ ▷ Convert to cost for Hungarian
$\mathcal{M}_1 \leftarrow \text{Hungarian}(\mathbf{C}_{geo})$; filter: keep (i,j) where $\mathbf{S}_{geo}(i,j) \geq \hat{\tau}_{geo}^t$
$\mathcal{D}_{unmatch}, \mathcal{T}_{unmatch} \leftarrow \text{GetUnmatched}(\mathcal{M}_1)$
// Level 2: Appearance Association (Score-based)
$\mathbf{A}_{TAA} \leftarrow \text{TAA}(\mathcal{D}_{unmatch}, \mathcal{T}_{unmatch})$ ▷ Affinity scores $\in [0,1]$
$\mathbf{A}_{combined} \leftarrow \alpha \cdot \mathbf{A}_{TAA} + (1-\alpha) \cdot \mathbf{S}_{geo}[\mathcal{D}_{unmatch}, \mathcal{T}_{unmatch}]$ ▷ Slice S_geo to unmatched subset
$\mathcal{M}_2 \leftarrow \text{Hungarian}(1 - \mathbf{A}_{combined})$; Filter: keep (i,j) where $\mathbf{A}_{combined}(i,j) \geq \hat{\tau}_{app}^t$
$\mathcal{D}_{unmatch}, \mathcal{T}_{unmatch} \leftarrow \text{GetUnmatched}(\mathcal{M}_2)$
// Level 3: SOT-based Virtual Generation
for $tr_j \in \mathcal{T}_{unmatch}$ **do**
  $\hat{\mathbf{p}}_j^t \leftarrow \text{MotionPredict}(tr_j)$ ▷ Kalman prediction
  $\mathbf{r}_j \leftarrow \text{ROIAlign}(F_{temp}^t, \hat{\mathbf{p}}_j^t, r_{search})$
  $s_j^{sot}, \Delta\mathbf{p}_j \leftarrow \text{SOTNetwork}(\mathbf{r}_j)$ ▷ Confidence and offset
  if $s_j^{sot} > \tau_{sot}$ **then**
    $\mathbf{p}_j^{virtual} \leftarrow \hat{\mathbf{p}}_j^t + \Delta\mathbf{p}_j$ ▷ Refine position
    Generate virtual detection at $\mathbf{p}_j^{virtual}$
  end if
end for
// Level 4: Recovery & Initialization
$\mathcal{M}_4 \leftarrow \text{RecoverLostTracks}(\mathcal{D}_{unmatch}, \mathcal{T}_{lost})$
Initialize new tracks from remaining high-confidence detections
return Updated $\mathcal{T}_{active}$

Level 1: Adaptive Geometric Matching. We first associate high-confidence detections with tracks using 3D IoU and center distance. We compute the IoU score matrix $\mathbf{S}_{geo}$ and convert it to a cost matrix $\mathbf{C}_{geo} = 1 - \mathbf{S}_{geo}$ for the Hungarian algorithm (which minimizes cost). After obtaining the optimal assignment, we filter matches by requiring the IoU score to exceed the adaptive threshold:

$$\text{Match}(d_i, tr_j) \text{ valid iff } \text{IoU}_{3D}(d_i, tr_j) \geq \hat{\tau}_{geo}^t. \tag{9}$$

This formulation ensures that a higher threshold $\hat{\tau}_{geo}^t$ enforces stricter geometric constraints, which is the intended behavior in dense scenes.

Level 2: Global Appearance Association. For detections and tracks unmatched in Level 1, we employ the TAA module. We compute the affinity score matrix $\mathbf{A}_{TAA} \in [0,1]$ and perform a second round of bipartite matching. We then form the combined affinity $\mathbf{A}_{combined}$ (see below) and convert it to costs $(1 - \mathbf{A}_{combined})$ for Hungarian matching, filtering by requiring $\mathbf{A}_{combined}(i,j) \geq \hat{\tau}_{app}^t$. This step resolves identity switches where geometric predictions deviate due to sudden maneuvers. We emphasize that the geo-

metric information from Level 1 is not discarded for detections entering Level 2. Specifically, we retain the soft IoU scores $\mathbf{S}_{geo}(i,j)$ computed in Level 1 (only for pairs where $(i,j) \in \mathcal{D}_{unmatch} \times \mathcal{T}_{unmatch}$) and combine them with the TAA affinity as a weighted fusion, $\mathbf{A}_{combined}(i,j) = \alpha \cdot \mathbf{A}_{TAA}(i,j) + (1-\alpha) \cdot \mathbf{S}_{geo}(i,j)$, where $\alpha = 0.7$ is validated via grid search (see Table 11). This ensures that geometric priors still inform the association decision even when they are insufficient alone, preventing information loss across levels.

Level 3: SOT-based Virtual Generation. To handle trajectory fragmentation (FRAG) caused by detector misses (e.g., partial occlusion), we do not immediately terminate unmatched tracks. Instead, we treat them as Single-Object Tracking (SOT) targets. For each unmatched track tr_j with last known state $\mathbf{x}_j^{t-1} = [x, y, z, w, l, h, \theta, v_x, v_y]^\top$, we first predict its current position using a constant-velocity motion model:

$$\hat{\mathbf{p}}_j^t = \mathbf{p}_j^{t-1} + \Delta t \cdot \mathbf{v}_j^{t-1}, \tag{10}$$

where $\mathbf{p}_j = [x, y, z]^\top$ and $\mathbf{v}_j = [v_x, v_y, 0]^\top$. We then perform a local correlation search in the temporal feature volume F_{temp}^t within a radius r_{search} around $\hat{\mathbf{p}}_j^t$:

$$\mathbf{r}_j = \text{ROIAlign}(F_{temp}^t, \hat{\mathbf{p}}_j^t, r_{search}) \in \mathbb{R}^C. \tag{11}$$

The SOT network then predicts both a confidence score and a position offset:

$$s_j^{sot}, \Delta\mathbf{p}_j = \text{SOTNetwork}(\mathbf{r}_j), \quad s_j^{sot} \in [0,1], \Delta\mathbf{p}_j \in \mathbb{R}^3. \tag{12}$$

SOTNetwork Architecture and Training Details. The SOTNetwork is a lightweight center-regression module consisting of three fully connected layers ($256 \to 128 \to 64 \to 4$), where the output comprises a 1-dimensional confidence score s_j^{sot} (passed through sigmoid) and a 3-dimensional position offset $\Delta\mathbf{p}_j$. The network takes ROI-pooled features $\mathbf{r}_j \in \mathbb{R}^C$ as input, where $C = 256$. The search region r_{search} is set to 3 m $\times$ 3 m in BEV space, which covers the maximum expected displacement between consecutive frames at 2 Hz annotation frequency. The SOTNetwork is pre-trained on the nuScenes training set using paired consecutive frames with ground-truth associations: the training loss is $\mathcal{L}_{sot} = \mathcal{L}_{BCE}(s_j^{sot}, s_j^{gt}) + \lambda_{offset}\mathcal{L}_{L1}(\Delta\mathbf{p}_j, \Delta\mathbf{p}_j^{gt})$, where $\lambda_{offset} = 2.0$. The confidence threshold is $\tau_{sot} = 0.5$. The maximum consecutive virtual generation window $\delta_{sot_max} = 5$ frames is determined by analyzing the distribution of occlusion durations in the training set: over 92% of partial occlusions last fewer than 5 frames ($\sim$2.5s), beyond which cumulative drift errors become unacceptable. The SOTNetwork weights are not shared across object categories, and all weights are kept frozen during the training of our main framework to ensure efficiency and prevent gradient interference.

If $s_j^{sot} > \tau_{sot}$, we generate a "virtual detection" $\hat{d}_j^t$ at the refined position $\mathbf{p}_j^{virtual} = \hat{\mathbf{p}}_j^t + \Delta\mathbf{p}_j$. To prevent cumulative drift errors, this virtual generation is only allowed for a short consecutive window of up to δ_{sot_max} frames. If occlusion exceeds this limit, the track is downgraded to a "Lost" state and handed over to Level 4. The position refinement $\Delta\mathbf{p}_j$ corrects the Kalman prediction error, improving localization accuracy during occlusions. This mechanism bridges short-term detector failures without introducing false positives.

Level 4: Interval-Frame Recovery and Initialization. Finally, any remaining high-confidence detections (score $> \tau_{init}$) are initialized as new tracks. For low-confidence detections, we perform "interval-frame recovery": we compute appearance similarity between these detections and tracks that were lost within the sparse temporal window $\mathcal{T}_{long}$ (from the long-term CSTA branch):

$$\mathbf{A}_{recover}(d_i, tr_j^{lost}) = \frac{1}{2}\left(\frac{\mathbf{e}_i^{app} \cdot \bar{\mathbf{e}}_j^{app}}{\|\mathbf{e}_i^{app}\|_2 \|\bar{\mathbf{e}}_j^{app}\|_2} + 1\right), \tag{13}$$

where $\bar{\mathbf{e}}_j^{app}$ is the exponential moving average of track j's historical embeddings. Matches exceeding $\tau_{recover}$ reactivate the lost track, effectively stitching broken trajectories caused by extended occlusions.

3.6. Training Objective

The overall training objective consists of three components:

$$\mathcal{L}_{total} = \mathcal{L}_{det} + \lambda_{app}\mathcal{L}_{app} + \lambda_{salt}\mathcal{L}_{salt}, \tag{14}$$

where λ_{app} and λ_{salt} are balancing weights.

Detection Loss $\mathcal{L}_{det}$. We adopt the standard detection loss from CenterPoint [11], including heatmap focal loss, box regression L1 loss, and velocity prediction loss.

Appearance Embedding Loss $\mathcal{L}_{app}$. To learn discriminative appearance embeddings, we employ a triplet loss with online hard negative mining:

$$\mathcal{L}_{app} = \max(0, \|\mathbf{e}_a - \mathbf{e}_p\|_2 - \|\mathbf{e}_a - \mathbf{e}_n\|_2 + m), \tag{15}$$

where $\mathbf{e}_a$, $\mathbf{e}_p$, $\mathbf{e}_n$ denote anchor, positive, and negative embeddings respectively, and m is the margin.

SALT Supervision Loss $\mathcal{L}_{salt}$. We supervise the threshold predictor using oracle thresholds τ^* derived from ground-truth associations. To prevent gradient explosion when threshold predictions deviate significantly and to provide robustness against outlier oracle values, we employ the Huber (smooth L1) loss:

$$\mathcal{L}_{salt} = \text{Huber}_\delta(\hat{\tau}_{geo}^t, \tau_{geo}^*) + \text{Huber}_\delta(\hat{\tau}_{app}^t, \tau_{app}^*), \tag{16}$$

where $\text{Huber}_\delta(x, y) = \begin{cases} \frac{1}{2}(x-y)^2, & |x-y| \le \delta \\ \delta(|x-y| - \frac{1}{2}\delta), & \text{otherwise} \end{cases}$ with $\delta = 0.1$. The bounded sigmoid output (Equation (4)), together with the Huber loss, naturally prevents threshold values from reaching extremes. The oracle thresholds τ^* are generated offline during training. The oracle geometric threshold τ_{geo}^* is dynamically determined to be the tightest lower bound that preserves all true positive associations. Formally:

$$\tau_{geo}^* = \begin{cases} \max\left(\min_{(i,j)\in\mathcal{M}_{GT}}(\text{IoU}_{i,j}) - \epsilon, \tau_{min}\right), & \text{if } \mathcal{M}_{GT} \neq \emptyset \\ \tau_{default}, & \text{otherwise} \end{cases} \tag{17}$$

where $\mathcal{M}_{GT}$ is the set of ground-truth matched pairs at frame t, $\epsilon = 0.05$ is a safety margin, $\tau_{min} = 0.1$ is the lower bound, and $\tau_{default}$ is set to the mean threshold of the training set (used when no GT matches exist in a frame). This ensures the threshold is tight enough to exclude false positives but loose enough to include the worst true match. The sensitivity of the margin ϵ is analyzed in Table 9; we found that values in the range $[0.03, 0.07]$ yielded comparable performance, with $\epsilon = 0.05$ achieving the best balance. The appearance threshold τ_{app}^* is computed analogously using cosine similarity of ground-truth pairs.

Note that since the SOTNetwork used in Level 3 relies on pre-trained and frozen weights, it requires no gradient updates and thus is excluded from $\mathcal{L}_{total}$.

4. Experiments

In this section, we conduct a comprehensive set of experiments to validate the effectiveness of the proposed JMM3DDT framework. We first compare our method against state-of-the-art 3D MOT tracking methods on the widely used nuScenes benchmark. Subsequently, we perform detailed ablation studies to dissect the contribution of each core component: multi-granularity CSTA, Transformer-based Appearance Association (TAA), and Scene-Adaptive Learned Thresholds (SALTs).

4.1. Experimental Setup

Dataset and Metrics. We evaluated JMM3DDT on the nuScenes dataset [4], a large-scale autonomous driving dataset containing 1000 scenes (700 for training, 150 for validation, and 150 for testing). The dataset provides data from six cameras, one LiDAR, and five radars, annotated at 2 Hz. We report standard 3D MOT metrics: Average Multi-Object Tracking Accuracy (AMOTA), Average Multi-Object Tracking Precision (AMOTP), MOTA, identity switches (IDsws), and fragmentation (FRAG). AMOTA is the primary metric for ranking on the leaderboard.

Implementation Details. Our framework was implemented in PyTorch1.13. For the detection backbone, we utilized CenterPoint [11] with VoxelNet to generate proposal features. The multi-granularity CSTA module processed a temporal window of $T = 7$ frames (short-term window of size three, long-term stride of two). The TAA Transformer consisted of two decoder layers with four attention heads. The SALT MLP had three hidden layers of size 64. We trained the model end-to-end for 20 epochs on four NVIDIA RTX 3090 GPUs. The optimizer was AdamW with an initial learning rate of 2×10^{-4} and cosine annealing. The weighting factors were set to $\lambda_{app} = 1.0$ and $\lambda_{salt} = 0.5$. For the Level 4 recovery stage, we set $\tau_{init} = 0.5$ for new-track initialization and $\tau_{recover} = 0.6$ for appearance-based re-identification. The Level 2 fusion weight was $\alpha = 0.7$ (validated in Table 11). During inference, the maximum consecutive window for SOT virtual generation δ_{sot_max} was strictly limited to five frames ($\sim$2.5 s at 2 Hz) to prevent drift. The maximum memory age for recovering lost tracks in Level 4, denoted as δ_{life}, was set to 12 frames ($\sim$6 s at 2 Hz). This value was determined by our false reactivation analysis (Table 10): recovery precision dropped from 95.5% ($\leq$6 frames) to 84.7% (7–12 frames), and exceeding 12 frames yielded diminishing returns with unacceptable false reactivation risk.

Runtime Performance. We profiled the inference latency of each component on a single NVIDIA RTX 3090 GPU, as summarized in Table 1. The full pipeline ran at approximately 9.8 FPS (102 ms per frame), which is competitive for the 2 Hz annotation frequency of nuScenes. The detection backbone consumed 50 ms, the CSTA temporal modeling added 28 ms, the TAA module added 15 ms, the SALT MLP was negligible ($<$1 ms), and the SOT recovery added 8 ms. While this is not strictly real-time for high-frequency LiDAR (10 Hz), we note that (1) the nuScenes benchmark operates at 2 Hz; (2) the dominant cost is the detection backbone, which can be accelerated via TensorRT; and (3) TAA can be further optimized through sparse attention.

Table 1. Inference latency breakdown on a single RTX 3090 GPU. The best results are marked in **bold**.

Component	Latency (ms)
Detection Backbone (CenterPoint)	50
CSTA Temporal Modeling	28
TAA (2 layers, 4 heads)	15
SALT MLP	$<$1
SOT Recovery (Level 3)	8
Total Pipeline	**$\sim$102 (9.8 FPS)**

4.2. Comparison with State-of-the-Art Methods

Results on nuScenes Test Set. As shown in Table 2, JMM3DDT achieves state-of-the-art performance with an AMOTA of 73.5%, surpassing the previous best AlphaTrack (69.3%) by 4.2% and CenterPoint (63.8%) by 9.7%. Our method also achieves the best AMOTP of 48.3% and MOTA of 59.6%, indicating superior localization precision and overall tracking accuracy. Most notably, JMM3DDT demonstrates exceptional tracking stability with only 173 identity switches, representing a 69.9% reduction compared to SimpleTrack (575) and 77.2% reduction compared to CenterPoint (760). Compared to OGR3MOT which also focuses on reducing IDsws (288), our method achieves 39.9% fewer identity switches while surpassing it by 7.9% in AMOTA. The significant improvements across all metrics validate that our hierarchical association strategy with scene-adaptive thresholds effectively resolves ambiguities in crowded scenarios.

Results on nuScenes Validation Set. Table 3 presents comprehensive results on the validation set. Our JMM3DDT achieves 75.8% AMOTA with 2D + 3D fusion, outperforming the previous best 3DMOTFormer (71.2%) by 4.6%. The MOTA improvement is even more substantial, reaching 64.5% compared to 60.7% for 3DMOTFormer (+3.8%). Our method also achieves the best AMOTP of 49.6%, demonstrating superior localization quality. The identity switches are reduced from 562 (CenterPoint) to 195, representing a 65.3% reduction. Compared to OGR3MOT (262 IDsws), our method achieves 25.6% fewer identity switches while improving AMOTA by 6.5%. These results confirm that our multi-granularity temporal modeling (CSTA), global appearance association (TAA), and scene-adaptive thresholds (SALTs) work synergistically to achieve robust tracking across diverse scenarios.

Table 2. Quantitative comparison on the nuScenes test set using CenterPoint detections. Our JMM3DDT demonstrates superior performance, especially in reducing identity switches. Best results are in **red**, second best in **blue**. ↑ indicates that higher scoreis better. ↓ indicates that lower scoreis better.

	Input	AMOTA (%) ↑	AMOTP (%) ↓	MOTA (%) ↑	IDsw ↓	FRAG ↓
DeepFusionMOT [34]	2D + 3D	63.5	66.8	53.3	1705	953
CenterPoint [11]	3D	63.8	55.5	53.7	760	529
StrongFusionMOT [35]	2D + 3D	65.2	67.1	55.1	1462	783
CBMOT [25]	2D + 3D	67.6	51.8	53.9	709	1015
EagerMOT [9]	2D + 3D	67.7	55.0	56.8	1156	601
OGR3MOT [36]	3D	65.6	62.0	55.4	288	455
SimpleTrack [6]	3D	66.8	55.0	56.6	575	591
GNN-PMB [37]	3D	67.8	56.0	56.3	770	431
ImmortalTracker [38]	3D	67.7	59.9	57.2	320	477
3DMOTFormer [39]	3D	68.2	**49.6**	55.6	438	529
AlphaTrack [40]	3D	**69.3**	58.5	**57.6**	718	480
JMM3DDT (Ours)	2D + 3D	**73.5**	**48.3**	**59.6**	**173**	**355**

Table 3. Quantitative comparison on the nuScenes validation set. Best results are in **red**, second best in **blue**. ↑ indicates that higher scoreis better. ↓ indicates that lower scoreis better.

	Input	AMOTA (%) ↑	AMOTP (%) ↓	MOTA (%) ↑	IDsw ↓	FRAG ↓
CenterPoint [11]	3D	66.5	56.7	56.2	562	424
CBMOT [25]	2D + 3D	67.5	-	-	494	-
OGR3MOT [36]	3D	69.3	62.7	60.2	262	420
SimpleTrack [6]	3D	69.6	54.7	60.2	-	403
GNN-PMB [37]	3D	70.7	56.0	-	650	332
ImmortalTracker [38]	3D	70.2	-	60.1	-	385
3DMOTFormer [39]	3D	**71.2**	**51.5**	**60.7**	341	436
JMM3DDT (Ours)	2D + 3D	**75.8**	**49.6**	**64.5**	**195**	**342**

4.3. Ablation Studies

We conducted ablation studies on the nuScenes validation set to verify the contribution of each component. The baseline utilized a standard CenterPoint detector with Kalman Filter and greedy Hungarian matching (IoU threshold of 0.1).

Impact of Core Modules. Table 4 summarizes the progressive improvement from baseline (66.5% AMOTA) to full model (75.8% AMOTA), totaling a 9.3% absolute gain:

- **Effectiveness of CSTA (+3.3% AMOTA):** Integrating multi-granularity CSTA significantly improves motion prediction. The dual-branch design captures both fine-grained pedestrian movements and coarse vehicle trajectories, reducing IDsws from 562 to 428 (23.8% reduction).
- **Effectiveness of TAA (+2.6% AMOTA):** Adding Transformer-based association drastically cuts IDsws from 428 to 298 (30.4% reduction). This confirms that global cross-attention resolves identity confusion in crowded scenes better than local pair-wise metrics.
- **Effectiveness of SALTs (+1.8% AMOTA):** The adaptive thresholds further reduce IDsws to 232. It is noteworthy that SALTs improve AMOTA (+1.8%) while simultaneously reducing IDsws. This is because the SALT module operates bi-directionally: it tightens thresholds in crowded scenes to filter false positives (reducing IDsws), while relaxing thresholds in high-velocity, sparse scenarios to recover valid tracks that would otherwise be missed by a fixed, strict threshold (reducing false negatives). This dual adaptability optimizes the trade-off between precision and recall.
- **Hierarchical Strategy (+1.6% AMOTA):** The full four-level strategy with SOT-based recovery reduces IDsws to 195 and FRAG to 342, demonstrating effective trajectory maintenance during occlusions.

Table 4. Component-wise ablation study on nuScenes validation set. "CSTA": Multi-granularity Temporal Modeling; "TAA": Transformer Appearance Association; "SALTs": Scene-Adaptive Thresholds; "Hier": Hierarchical Strategy (Level 3 & 4). ↑ indicates that higher scoreis better. ↓ indicates that lower scoreis better. The best results are marked in **bold**.

CSTA	TAA	SALTs	Hier	AMOTA (%) ↑	IDsw ↓	FRAG ↓
-	-	-	-	66.5	562	424
✓	-	-	-	69.8 (+3.3)	428	392
✓	✓	-	-	72.4 (+2.6)	298	375
✓	✓	✓	-	74.2 (+1.8)	232	358
✓	✓	✓	✓	**75.8 (+1.6)**	**195**	**342**

Analysis of Temporal Granularity. A core contribution of our framework is the multi-granularity design. To prove its necessity, we compare the dual-branch CSTA against single-scale variants in Table 5. Using only the dense short-term branch (three frames) yields 68.2% AMOTA, as it struggles to maintain features during extended occlusions. Conversely, using only the sparse long-term branch (seven frames with stride two) improves robustness against occlusion (68.9% AMOTA) but fails to capture sudden maneuvers like pedestrian turning. By fusing both scales, the Multi-granularity CSTA achieves the best performance (69.8% AMOTA) and the lowest number of identity switches (428), confirming that both micro-movements and macroscopic trends are indispensable for robust 3D tracking.

Why TAA Outperforms Standard Cosine Similarity. When geometry fails, appearance is critical. As shown in Table 6, replacing our TAA with a standard pair-wise cosine similarity matrix only yields a marginal improvement (70.6% AMOTA, 375 IDsws). TAA, by contrast, forces detection queries to attend to the global distribution of all tracks via MHCA, effectively highlighting distinctive features and suppressing ambiguities. This context-awareness drops IDsws to 298, proving its superiority over local pair-wise metrics.

Table 5. Ablation of temporal granularity in CSTA on the nuScenes validation set. Models were trained without TAA and SALTs to strictly isolate the effects of temporal modeling. ↑ indicates that higher scoreis better. ↓ indicates that lower scoreis better. The best results are marked in **bold**.

Temporal Configuration	AMOTA (%) ↑	MOTA (%) ↑	IDsw ↓
Single Scale (3-frame dense)	68.2	58.1	485
Single Scale (7-frame sparse)	68.9	58.7	458
Multi-granularity (Dual-branch)	**69.8**	**59.9**	**428**

Table 6. Ablation of appearance association strategies (replaces Level 2). The best results are marked in **bold**.

Matching Strategy	AMOTA (%)	MOTA (%)	IDsw
None (IoU geometric only)	69.8	59.9	428
Pair-wise Cosine Similarity	70.6	60.5	375
Transformer Cross-Attention (TAA)	**72.4**	**62.1**	**298**

Adaptive vs. Fixed Thresholds. We validate that SALTs balance the precision–recall trade-off in Table 7. A fixed strict threshold aggressively filters false positives (FPs) but creates enormous false negatives (FNs) because fast-moving objects are discarded. A fixed loose threshold rescues FNs but drowns the system in FPs during dense traffic. SALTs dynamically navigate this trade-off based on scene context (density/velocity), achieving a "best of both worlds" balance that dramatically maximizes the comprehensive AMOTA metric (74.2%).

Table 7. Effectiveness of Scene-Adaptive Learned Thresholds (SALT). ↑ indicates that higher scoreis better. ↓ indicates that lower scoreis better. The best results are marked in **bold**.

Threshold Strategy	AMOTA (%)	FP ↓	FN ↓	IDsw ↓
Fixed Strict ($\tau_{geo} = 0.6$)	70.1	**11,540**	26,810	310
Fixed Loose ($\tau_{geo} = 0.2$)	71.5	18,320	**19,650**	385
SALTs (Adaptive Dynamic)	**74.2**	12,100	20,430	**232**

Discussion on the FN Trade-off in SALTs. As Table 7 shows, SALTs slightly increase FNs compared to the fixed loose setting (20,430 vs. 19,650, a 4.0% increase), while dramatically reducing FPs by 34.0% (12,100 vs. 18,320) and IDsws by 39.7% (232 vs. 385). We acknowledge this trade-off explicitly: the SALT module intentionally tightens thresholds in dense scenes where the risk of false associations (leading to IDsws) outweighs the cost of missing a few detections. The net benefit is a 2.7% absolute AMOTA improvement, because AMOTA penalizes IDsws much more heavily than FNs—each identity switch corrupts two trajectories simultaneously (the source and the erroneously assigned target), while a false negative only causes a temporary gap that can be recovered by Level 3/4. Therefore, the precision–recall trade-off made by SALTs is both intentional and beneficial for the overall tracking objective.

SALT Sensitivity Analysis. To analyze the interpretability and robustness of the SALT module, we evaluated how the learned thresholds responded to different scene conditions. Table 8 shows the average predicted thresholds across three representative scene categories. In high-density scenes ($\rho_t > 8$), the SALT module predicts strict thresholds ($\hat{\tau}_{geo} = 0.52$), reducing false associations. In high-velocity scenes ($\bar{v}_t > 8$ m/s), it relaxes the geometric threshold ($\hat{\tau}_{geo} = 0.22$) to accommodate larger displacement. Under heavy occlusion ($\eta_t > 0.4$), the SALT module simultaneously tightens geometric and relaxes appearance thresholds, leveraging appearance cues when geometry is unreliable. These

behaviors remain consistent across the validation and test splits, confirming the stability of the learned policy.

Table 8. SALT behavior under different scene conditions (averaged over nuScenes validation set).

Scene Condition	$\hat{\tau}_{geo}$	$\hat{\tau}_{app}$	AMOTA (%)
High Density ($\rho_t > 8$)	0.52	0.61	72.8
High Velocity ($\bar{v}_t > 8$ m/s)	0.22	0.55	75.1
High Occlusion ($\eta_t > 0.4$)	0.45	0.42	71.5
Balanced (Overall)	0.38	0.53	74.2

Oracle Margin Sensitivity. We investigate the sensitivity of the oracle margin ϵ in Equation (17) on the nuScenes validation set (Table 9).

Table 9. Sensitivity analysis of the oracle margin ϵ on SALT training. ↓ indicates that lower scoreis better. The best results are marked in **bold**.

ϵ	AMOTA (%)	IDsw ↓	FP ↓
0.01	73.5	248	11,850
0.03	73.9	238	12,020
0.05	**74.2**	**232**	12,100
0.07	74.0	235	12,350
0.10	73.4	255	13,100

A too-small ϵ (0.01) yields overly aggressive oracle thresholds that are hard to learn, while a too-large ϵ (0.10) produces overly loose targets. The sweet spot $\epsilon \in [0.03, 0.07]$ yields robust performance, and $\epsilon = 0.05$ achieves the best balance. The results confirm that the margin has a moderate but not dramatic impact, and our choice of $\epsilon = 0.05$ is well justified.

Level 4 False Reactivation Analysis. A potential risk of interval-frame recovery is false reactivation, i.e., incorrectly matching a new object to a lost track's memory. We quantify this risk in Table 10.

Table 10. Analysis of Level 4 recovery outcomes on the nuScenes validation set.

Recovery Window	Total Recov.	Correct	False Reactiv.	Precision (%)
≤6 frames (∼3 s)	312	298	14	95.5
7–12 frames (∼3–6 s)	85	72	13	84.7
Overall (≤12 frames)	397	370	27	93.2

The overall false reactivation rate is 6.8% (27/397), and the net effect of Level 4 recovery is clearly positive: it recovers 370 correct associations while introducing only 27 false ones (net gain of 343 fewer fragmentation events). Within the shorter window (≤6 frames), the precision is 95.5%, confirming that recent lost tracks are reliably recoverable. For longer windows (7–12 frames), precision drops to 84.7%, which motivates our choice of $\delta_{life} = 12$ frames as the upper limit.

Level 2 Fusion Weight α Ablation. We validated the choice of α in the Level 2 combined affinity $\mathbf{A}_{combined} = \alpha \cdot \mathbf{A}_{TAA} + (1 - \alpha) \cdot \mathbf{S}_{geo}$ via grid search (Table 11).

Setting $\alpha = 1.0$ (discarding geometric information entirely) degrades performance, confirming that retaining Level 1 geometric priors is beneficial. The optimal $\alpha = 0.7$ reflects that appearance should dominate at Level 2 (since geometric matching already failed at Level 1), while geometric cues still provide useful regularization.

Table 11. Ablation of the Level 2 fusion weight α on the nuScenes validation set. ↓ indicates that lower scoreis better. The best results are marked in **bold**.

α (TAA Weight)	AMOTA (%)	IDsw ↓	MOTA (%)
0.5 (equal)	74.8	212	63.8
0.6	75.3	202	64.1
0.7	**75.8**	**195**	**64.5**
0.8	75.5	198	64.3
1.0 (TAA only, no geo)	75.1	208	63.9

4.4. Qualitative Analysis and Case Study

To explicitly demonstrate how our proposed hierarchical strategy resolves trajectory fragmentation (FRAG) and identity switches (IDsws) under extreme conditions, we visualize a challenging long-term occlusion scenario in Figure 3.

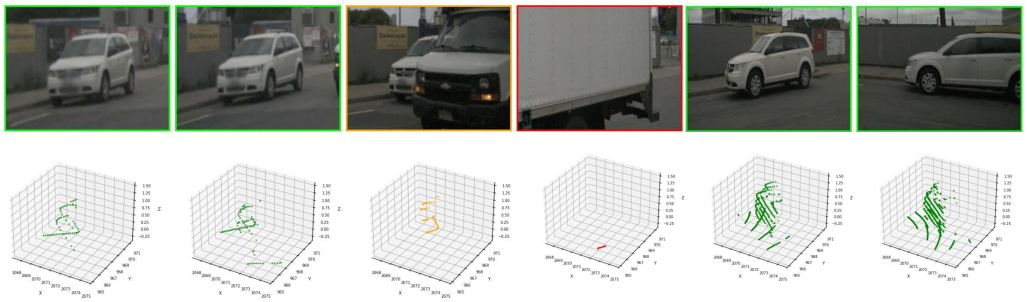


Figure 3. Qualitative visualization of a severe occlusion scenario. Top: Camera views. Bottom: Corresponding 3D point clouds. Green boxes (T_1, T_2, T_5, T_6) indicate normal tracking. The orange box (T_3) denotes partial occlusion handled by Level 3 SOT generation. The red box (T_4) represents complete occlusion. JMM3DDT successfully recovers the original identity at T_5 despite the visual ambiguity caused by other white vehicles.

Qualitative Visual Analysis. As shown in Figure 3, the target (a white SUV) undergoes a complex interaction with other vehicles. At T_1 and T_2, the target is fully visible with dense point clouds, and our Level 1 geometric matching stably tracks it. At T_3, a white van partially occludes the target (orange box). Conventional trackers typically terminate the track here due to low detector confidence. However, our Level 3 SOT-based virtual generation utilizes local temporal features to regress the target's position, bridging the gap without generating a false negative. At T_4, a large white truck completely occludes the target (red box). The track is temporarily suspended and moved to the lost track pool. Crucially, when the SUV re-emerges at T_5, the scene is visually ambiguous (multiple white vehicles in close proximity). Relying solely on geometric extrapolation would fail here. Our Level 4 recovery coupled with theTAA cross-attention effectively discriminates the SUV from the truck, re-associating it with its original identity from T_2.

Instance-level Quantitative Study. To quantitatively validate this, we extracted the tracking state sequences of this specific instance comparing the baseline (CenterPoint + SORT) with JMM3DDT, as detailed in Table 12. The baseline dropped the target at T_3 and assigned a completely new identity ID (ID-89) at T_5, resulting in both a fragmentation event and an identity switch. In contrast, our method maintained ID-42 seamlessly.

Furthermore, to prove the superiority of the TAA module over traditional metrics in Level 2/4 re-identification, we compared the appearance affinity scores between the historical track memory (from T_2) and the detections present at T_5 (which include the target SUV and the distractor white truck). As shown in Table 13, traditional pair-wise cosine

similarity yields high scores for both the target (0.74) and the distractor (0.68), making the assignment highly ambiguous and prone to errors if thresholds fluctuate. Conversely, our TAA module, which computes attention globally across all candidates, sharply highlights the correct target (0.88) while suppressing the distractor (0.21). This proves that global context is imperative for robust re-identification in ambiguous traffic conditions.

Table 12. Tracking state sequence comparison for the instance in Figure 3. Our method maintains a consistent ID through occlusions, whereas the baseline suffers an ID switch. The best results are marked in **bold**.

Method	T_1	T_2	T_3 (Partial Occ)	T_4 (Full Occ)	T_5 (Re-Emerge)	T_6
Baseline	ID-42	ID-42	Missed	Lost	ID-89 (IDsw)	ID-89
Ours	**ID-42**	**ID-42**	**ID-42 (SOT)**	**Lost (Mem)**	**ID-42 (Recov)**	**ID-42**

Table 13. Appearance affinity scores at T_5 between historical track memory (T_2) and current detections. TAA significantly suppresses the distractor compared to cosine similarity. The best results are marked in **bold**.

Matching Metric	True Target (White SUV)	Distractor (White Truck)
Cosine Similarity	0.74	0.68 (ambiguous)
TAA (Ours)	**0.88**	**0.21** (discriminative)

5. Conclusions

We presented JMM3DDT, an adaptive multi-level 3D MOT framework that addresses the key limitations of existing methods through three innovations: (1) multi-granularity CSTA captures diverse motion patterns via dual-scale temporal modeling; (2) TAA resolves ambiguous associations through global context-aware cross-attention; (3) SALTs dynamically adjust association thresholds based on scene complexity. Our hierarchical four-level strategy progressively handles cases from easy geometric matching to complex trajectory recovery. Extensive experiments on nuScenes demonstrated state-of-the-art performance with 73.5% AMOTA and up to 77% fewer identity switches compared to baseline methods. Beyond achieving state-of-the-art quantitative metrics, the core academic contribution of this work is demonstrating the critical necessity of global contextual awareness and dynamic adaptability in 3D multi-object tracking. By moving away from fixed association metrics and single-scale temporal priors, JMM3DDT proves that tracking robustness fundamentally relies on a system's ability to interpret and react to the immediate scene complexity. This adaptive synergy effectively bridges the gap between structured geometric tracking and the unpredictable nature of real-world traffic environments, offering a new perspective for designing next-generation, environmentally aware perception systems in autonomous driving. Future work will explore extending our adaptive mechanisms to end-to-end joint detection and tracking frameworks.

Author Contributions: Y.Z.: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Data curation, Writing—Original Draft, Visualization. F.D.: Conceptualization, Methodology, Resources, Writing—Review and Editing, Supervision, Project Administration, Funding Acquisition. H.Z.: Data Curation, Validation, Resources, Writing—Review and Editing. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the National Natural Science Foundation of China under Grant 51475092 and supported by the Jiangsu province frontier leading technology basic research project under Grant BK20192004C.

Data Availability Statement: The data that support the findings of this study are available from the corresponding author on reasonable request.

Acknowledgments: This work was supported in part by the Key Laboratory of Measurement and Control of Complex Systems of Engineering, Ministry of Education, Southeast University.

Conflicts of Interest: Author Haocheng Zhou was employed by the company Jiangsu Electric Power Information Technology Co., Ltd. The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

References

1. Liang, J.; Yang, K.; Tan, C.; Wang, J.; Yin, G. Enhancing high-speed cruising performance of autonomous vehicles through integrated deep reinforcement learning framework. *IEEE Trans. Intell. Transp. Syst.* **2025**, *26*, 835–848. [CrossRef]
2. Liang, J.; Tan, C.; Yan, L.; Zhou, J.; Yin, G.; Yang, K. Interaction-Aware Trajectory Prediction for Safe Motion Planning in Autonomous Driving: A Transformer-Transfer Learning Approach. *IEEE Trans. Intell. Transp. Syst.* **2025**, *26*, 17080–17095. [CrossRef]
3. Weng, X.; Wang, J.; Held, D.; Kitani, K. 3D Multi-Object Tracking: A baseline for 3D multi-object tracking and new evaluation metrics. In *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*; IEEE: New York, NY, USA, 2020.
4. Caesar, H.; Bankiti, V.; Lang, A.H.; Vora, S.; Liong, V.E.; Xu, Q.; Krishnan, A.; Pan, Y.; Baldan, G.; Beijbom, O. Nuscenes: A multimodal dataset for autonomous driving. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*; IEEE: New York, NY, USA, 2020; pp. 11621–11631.
5. Bernardin, K.; Stiefelhagen, R. Evaluating multiple object tracking performance: The CLEAR MOT metrics. *EURASIP J. Image Video Process.* **2008**, *2008*, 246309. [CrossRef]
6. Pang, Z.; Li, Z.; Wang, N. SimpleTrack: Understanding and rethinking 3D multi-object tracking. In Proceedings of the ECCV Workshops (2022), Tel Aviv, Israel, 23 October 2022.
7. Zhou, X.; Koltun, V.; Krähenbühl, P. Tracking objects as points. In *Computer Vision—ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IV* ; Springer: Berlin/Heidelberg, Germany, 2020; pp. 474–490.
8. Li, J.; Zhang, H.; Xu, Z.; Kong, L.; Liu, J. Poly-MOT: A polyhedral framework for 3D multi-object tracking. In Proceedings of the 2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Detroit, MI, USA, 1–5 October 2023; pp. 12345–12351.
9. Kim, A.; Ošep, A.; Leal-Taixé, L. EagerMOT: 3D multi-object tracking via sensor fusion. In Proceedings of the 2021 IEEE International Conference on Robotics and Automation (ICRA), Xi'an, China, 30 May–5 June 2021; pp. 11315–11321.
10. Wang, L.; Zhang, X.; Qin, W.; Li, X.; Yang, L.; Li, Z.; Zhu, L.; Wang, H.; Li, J.; Liu, H. Camo-MOT: Combined appearance-motion optimization for 3D multi-object tracking with camera-lidar fusion. *arXiv* **2022**, arXiv:2209.02540. [CrossRef]
11. Yin, T.; Zhou, X.; Krähenbühl, P. Center-based 3D object detection and tracking. In Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 20–25 June 2021; pp. 11784–11793.
12. Bai, X.; Hu, Z.; Zhu, X.; Huang, Q.; Chen, Y.; Fu, H.; Tai, C.-L. TransFusion: Robust LiDAR-camera fusion for 3D object detection with transformers. In Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 18–24 June 2022; pp. 1090–1099.
13. Lin, X.; Pei, Z.; Lin, T.; Huang, L.; Su, Z. Sparse4D v3: Advancing end-to-end 3D detection and tracking. *arXiv* **2023**, arXiv:2311.11722.
14. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L.; Polosukhin, I. Attention is all you need. In Proceedings of the 31st International Conference on Neural Information Processing Systems, Long Beach, CA, USA, 4–9 December 2017; Volume 30.
15. Carion, N.; Massa, F.; Synnaeve, G.; Usunier, N.; Kirillov, A.; Zagoruyko, S. End-to-end object detection with transformers. In *Computer Vision—ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part I*; Springer: Berlin/Heidelberg, Germany, 2020; pp. 213–229.
16. Meinhardt, T.; Kirillov, A.; Leal-Taixe, L.; Feichtenhofer, C. TrackFormer: Multi-object tracking with transformers. In Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 18–24 June 2022; pp. 8844–8854.

17. Zeng, F.; Dong, B.; Zhang, Y.; Wang, T.; Zhang, X.; Wei, Y. MOTR: End-to-end multiple-object tracking with transformer. In *Computer Vision—ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXVII*; Springer: Berlin/Heidelberg, Germany, 2022; pp. 659–675.
18. Yu, Z.; Shu, S.; Deng, J.; Lu, K.; Liu, Z.; Yu, J.; Yang, D.; Li, H.; Chen, Y. FlashOCC: Fast and memory-efficient occupancy prediction with channel-to-height plugin. *arXiv* **2023**, arXiv:2311.12058.
19. Li, K.; Li, X.; Wang, Y.; He, Y.; Wang, Y.; Wang, L.; Qiao, Y. VideoMamba: State space model for efficient video understanding. *arXiv* **2024**, arXiv:2403.06977. [CrossRef]
20. Li, X.; Liu, D.; Wu, Y.; Wu, X.; Gao, J.; Zhao, L. Fast-Poly: A fast polyhedral framework for 3D multi-object tracking. *IEEE Robot. Autom. Lett.* **2024**, *9*, 10519–10526. [CrossRef]
21. Wang, X.; Qi, S.; Zhao, J.; Zhou, H.; Zhang, S.; Wang, G.; Tu, K.; Guo, S.; Zhao, J.; Li, J.; et al. MCTrack: A unified 3D multi-object tracking framework for autonomous driving. In Proceedings of the 2025 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Hangzhou, China, 19–25 October 2025.
22. Ding, S.; Schneider, L.; Cordts, M.; Gall, J. ADA-Track: End-to-end multi-camera 3D multi-object tracking with alternating detection and association. In Proceedings of the 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 16–22 June 2024; pp. 15184–15194.
23. Choi, J.; Ulbrich, S.; Lichte, B.; Maurer, M. Multi-target tracking using a 3D-LiDAR sensor for autonomous vehicles. In Proceedings of the 16th International IEEE Conference on Intelligent Transportation Systems (ITSC 2013), The Hague, The Netherlands, 6–9 October 2013.
24. Wang, H.; Shi, H.; Li, Y.; Fang, J.; Xu, H.; Wang, X.; Yang, B.S.; Li, L. UniTR: A unified and efficient multi-modal transformer for Bird's-Eye-View representation. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*; IEEE: New York, NY, USA, 2023.
25. Benbarka, N.; Schröder, J.; Zell, A. Score refinement for confidence-based 3D multi-object tracking. In Proceedings of the 2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Prague, Czech Republic, 27 September–1 October 2021.
26. Zhang, X.; Tan, X.; An Y.; Fan, Z. OATracker: Object-aware anti-occlusion 3D multiobject tracking for autonomous driving. *Expert Syst. Appl.* **2024**, *252*, 124158. [CrossRef]
27. Yu, E.; Li, T.; Li, Z.; Gu, Y.; Wang, X.; Wei, Y.; Zeng, F. MOTRv3: Release-fetch supervision for end-to-end multi-object tracking. *arXiv* **2023**, arXiv:2305.1429.
28. Li, Y.; Li, Q.; Wang, H.; Ma, X.; Yao, J.; Dong, S.; Fan, H.; Zhang, L. Beyond MOT: Semantic multi-object tracking. In *Computer Vision—ECCV 2024*; Springer: Cham, Switzerland, 2024; pp. 276–293.
29. Li, Y.; Chen, Y.; Qi, X.; Li, Z.; Sun, J.; Jia, J. UVTR: Unifying voxel-based representation with transformer for 3D object detection. In Proceedings of the 36th International Conference on Neural Information Processing System, New Orleans, LA, USA, 28 November–9 December 2022.
30. Huang, J.; Huang, G. BEVDet4D: Exploit temporal cues in multi-camera 3D object detection. *arXiv* **2022**, arXiv:2203.17054.
31. Bertasius, G.; Wang, H.; Torresani, L. Is space-time attention all you need for video understanding? In Proceedings of the 38th International Conference on Machine Learning (ICML 2021), Virtual, 18–24 July 2021.
32. Lv, W.; Huang, Y.; Zhang, N.; Lin, R.-S.; Han, M.; Zeng, D. DiffMOT: A real-time diffusion-based multiple object tracker with non-linear prediction. In Proceedings of the 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 16–22 June 2024; pp. 19321–19330.
33. Chiu, H.-K.; Wang, C.-Y.; Chen, M.-H.; Smith, S.F. Probabilistic 3D multi-object cooperative tracking for autonomous driving via differentiable multi-sensor Kalman filter. In Proceedings of the 2024 IEEE International Conference on Robotics and Automation (ICRA), Yokohama, Japan, 13–17 May 2024.
34. Wang, X.; Fu, C.; Li, Z.; Lai, Y.; He, J. DeepFusionMOT: A 3D multi-object tracking framework based on camera-lidar fusion with deep association. *IEEE Robot. Autom. Lett.* **2022**, *7*, 8260–8267. [CrossRef]
35. Wang, X.; Fu, C.; He, J.; Wang, S.; Wang, J. StrongFusionMOT: A multi-object tracking method based on LiDAR-camera fusion. *IEEE Sens. J.* **2023**, *23*, 11241–11252. [CrossRef]
36. Zaech, J.-N.; Liniger, A.; Dai, D.; Danelljan, M.; Van Gool, L. Learnable online graph representations for 3D multi-object tracking. *IEEE Robot. Autom. Lett.* **2022**, *7*, 5103–5110. [CrossRef]
37. Liu, J.; Bai, L.; Xia, Y.; Huang, T.; Zhu, B.; Han, Q.-L. GNN-PMB: A simple but effective online 3D multi-object tracker without bells and whistles. *IEEE Trans. Intell. Veh.* **2023**, *8*, 1176–1189. [CrossRef]
38. Wang, Q.; Chen, Y.; Pang, Z.; Wang, N.; Zhang, Z. Immortal tracker: Tracklet never dies. *arXiv* **2021**, arXiv:2111.13672. [CrossRef]

39. Ding, S.; Rehder, E.; Schneider, L.; Cordts, M.; Gall, J. 3DMOTFormer: Graph transformer for online 3D multi-object tracking. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*; IEEE: New York, NY, USA, 2023; pp. 9784–9793.
40. Zeng, Y.; Ma, C.; Zhu, M.; Fan, Z.; Yang, X. Cross-modal 3D object detection and tracking for auto-driving. In Proceedings of the 2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Prague, Czech Republic, 27 September–1 October 2021; pp. 3850–3857.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.