Integrated Detection and Tracking Framework for 3D Multi-Object Tracking in Vehicle-Infrastructure Cooperation

Tao Hu, Ping Wang, Xinhong Wang
College of Electronic and Information Engineering
Tongji University, Shanghai, China

***Abstract*—Vehicle-infrastructure cooperative perception has emerged as a promising approach to enhance 3D multi-object tracking by leveraging complementary data from vehicle and infrastructure sensors. However, existing methods face significant challenges, including difficulty in handling occlusions, suboptimal identity association, and inefficiencies in trajectory management, limiting their performance in real-world scenarios. In this paper, we propose a novel vehicle-infrastructure cooperative 3D multi-object tracking framework that addresses these challenges through three key innovations. First, an integrated detection-tracking framework jointly optimizes object detection and tracking, enhancing temporal consistency and reducing errors caused by separately handling the two tasks. Second, the XIOU identity association metric leverages 3D spatial and geometric relationships, ensuring robust object matching even under occlusions. Third, a four-stage cascade matching (FSCM) strategy adaptively manages trajectories by leveraging detection and prediction confidences, enabling accurate tracking in complex environments. Evaluated on the V2X-Seq dataset, our method achieves a MOTA of 57.23 and a MOTP of 74.64, significantly reducing identity switches while ensuring low bandwidth consumption and reliable tracking, highlighting its effectiveness and suitability for real-world deployment.**

***Keywords*—Vehicle-infrastructure cooperative perception; 3D multi-object tracking; XIOU metric; four-stage cascade matching; integrated detection-tracking framework**

I. Introduction

In recent years, while single-vehicle perception for autonomous driving has achieved notable advancements, *vehicle-infrastructure cooperative perception* has emerged as a transformative alternative. By integrating data from infrastructure-side sensors, this approach overcomes the inherent limitations of vehicle-only systems, such as limited perception range, blind spots, and lower detection confidence. It extends perception capabilities, enhances detection accuracy, and improves system reliability, positioning itself as a critical research focus in modern autonomous driving.

A cornerstone of vehicle-infrastructure cooperative perception is *3D multi-object tracking (MOT)*, which ensures temporal consistency by tracking objects across consecutive frames. Accurate 3D MOT is pivotal for downstream tasks like trajectory prediction and collision avoidance, directly contributing to the safety and effectiveness of autonomous driving systems. The objective is to accurately localize objects in 3D space while consistently maintaining their identities over time.

However, despite advancements in cooperative perception, 3D MOT methods still encounter significant challenges:

1) Insufficient Interaction Between Detection and Tracking: Many existing methods treat detection and tracking as separate processes, often relying on the *Tracking-by-Detection* paradigm. This limits the mutual enhancement between detection and tracking, as detection results are primarily optimized for tracking while failing to incorporate the valuable priors that tracking can provide for detection—an essential aspect for temporal consistency in dynamic environments.
2) Insufficient Utilization of 3D Spatial Information: Current methods often adapt 2D identity association techniques to 3D scenarios without fully harnessing the spatial richness of 3D point clouds. This limitation is particularly pronounced in vehicle-infrastructure cooperative contexts, where precise spatial alignment and 3D pose estimation are key for robust identity association.
3) Challenges in Occlusion Handling: Occlusion is a common occurrence in real-world autonomous driving scenarios, yet existing methods struggle to maintain consistent object identities when visibility is compromised. Effective mechanisms to handle occlusion and manage identity switches during visibility transitions remain underdeveloped.

These limitations underscore the need for more integrated, adaptive, and robust approaches to 3D MOT in vehicle-infrastructure cooperative scenarios. To address these challenges, we propose a novel *vehicle-infrastructure cooperative 3D multi-object tracking algorithm* based on an integrated detection and tracking architecture. By leveraging LiDAR point cloud data from both vehicle and infrastructure sources, our method provides a more accurate and real-time solution for 3D object detection and tracking. Specifically, it introduces the following key innovations:

1) Integrated Detection and Tracking Framework: We propose a fully integrated architecture that processes detection and tracking in a unified manner. By employing position encoding and cross-attention mechanisms, the framework seamlessly combines appearance and motion cues. Additionally, a temporal prior enhancement module is introduced, allowing tracking results to inform the detection process, leveraging

temporal priors to improve detection accuracy.

2) 3D-Task Adaptive Identity Association: Our method takes full advantage of 3D spatial pose information from LiDAR point clouds. We introduce a new identity association metric that accounts for spatial overlap, positional similarity, and orientation alignment, enabling robust object matching over time in 3D space.
3) Occlusion-Adaptive Matching Algorithm: We present a four-stage cascade matching algorithm that dynamically adjusts the tracking process based on the confidence levels of detection and prediction results. This strategy ensures robust tracking during occlusion events, effectively maintaining object identities even when objects are partially or fully hidden from view.

The remainder of this paper is organized as follows: Section II reviews related work, discussing existing approaches to vehicle-infrastructure cooperative 3D perception. Section III describes the proposed integrated detection and tracking framework in detail. Section IV outlines the experimental setup and results, covering datasets, evaluation metrics, and comprehensive analyses. Section V concludes the paper with final remarks and future research directions.

II. Related Work

This section discusses recent developments in *vehicle-infrastructure cooperative perception* and *multi-object tracking (MOT)*, both critical for improving autonomous driving systems.

A. Vehicle-Infrastructure Cooperative Perception

Cooperative perception enhances the perception capabilities of individual vehicles by integrating data from infrastructure-based sensors. It can be categorized into three main types based on the shared data:

1) Data-Level Cooperation: In data-level cooperation, raw sensor data such as LiDAR point clouds are directly shared between vehicles and infrastructure [1], [14], [29]. This approach allows vehicles to expand their perception range significantly by receiving unprocessed sensor data from other sources. However, the substantial data transmission requirements place a significant burden on network bandwidth.

2) Feature-Level Cooperation: Feature-level cooperation involves sharing pre-processed feature data, reducing the transmission load while retaining more information than object-level cooperation [17], [6], [5]. The performance of feature-level cooperation depends heavily on the feature fusion strategies employed, balancing between bandwidth savings and the amount of retained information.

3) Object-Level Cooperation: In object-level cooperation, only the final detection results are shared, minimizing the data transmission burden but potentially leading to information loss [24]. This method relies heavily on the accuracy and robustness of the individual perception models used on each vehicle and infrastructure component.

The emergence of large-scale cooperative perception datasets, both in simulation and real-world environments, has greatly advanced research in this domain. Notable examples include *V2X-Sim* [10], *DAIR-V2X-C* [24], and *V2X-Seq* [25], with *V2X-Seq* standing out as the first real-world dataset for vehicle-infrastructure sequential perception. These datasets provide comprehensive data for 3D object detection and tracking tasks, offering a foundation for further research in cooperative perception.

B. Multi-Object Tracking

In cooperative perception, *multi-object tracking (MOT)* plays a critical role in maintaining object identities over time by associating detected objects across frames, forming a temporal understanding of their movement. MOT is crucial for downstream tasks such as trajectory prediction and collision avoidance in autonomous driving. The process typically involves three main stages: object detection and feature representation, identity association, and trajectory management.

1) Object Detection and Feature Representation: This stage handles the initial detection and feature extraction from sensory data, which is critical for tracking objects across time. Traditionally, these tasks have been performed separately, with detection followed by feature extraction. One prominent example of this approach is *DeepSort* [19], which uses a Kalman filter for motion modeling and a deep learning-based appearance descriptor for object re-identification, offering a robust solution for tracking objects across frames while maintaining their identities. Recent methods, such as *OmniTracker* [16], focus on exploring the interplay between detection and tracking and transferring this relationship to various tracking tasks. This separation allows for independent optimization of detection and tracking, but it may lead to redundant computations.

Joint detection and feature representation, on the other hand, combines detection and tracking into a single network, improving the efficiency of multi-object tracking. Popular methods like *FairMOT* [27] and *TransTrack* [15] employ joint architectures where object detection and feature extraction are performed simultaneously, reducing the computational overhead while improving feature consistency. *TransMOT* [4] introduces a spatio-temporal graph transformer for encoding short trajectories and matching them across frames using a spatial graph decoder. Additionally, *UTM* [23] simultaneously enhances object detection and feature representation using identity-sensitive knowledge. Despite these advancements, challenges remain, especially in 3D multi-object tracking, where integrating motion cues and spatial information from point cloud data requires sophisticated handling of dynamic environments.

2) Identity Association: Identity association is critical for tracking objects across multiple frames and ensuring that each object maintains a consistent identity. Various metrics have been proposed to perform this association. Early works such as *AB3DMOT* [18] rely on simple intersection-over-union (IoU) metrics to associate objects between frames. However, this method struggles with occlusions and complex 3D environments. Recent researches such as Chiu et al. [3] and *CenterPoint* [22] improve identity association by replacing traditional IoU with more sophisticated metrics like Mahalanobis and L2 distance. These metrics help capture both spatial overlap and the underlying distribution of data points, enhancing tracking accuracy, especially in complex 3D environments where simple IoU metrics may struggle.

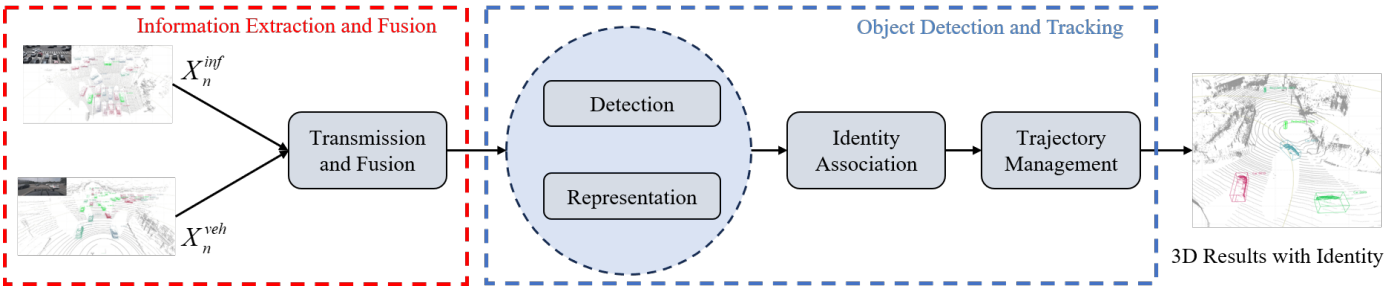

Fig. 1. The temporal perception process in vehicle-infrastructure cooperative sensing: The process comprises two main parts: data fusion and object detection-tracking, producing a 3D perception output with identity information.

Advanced association methods have been developed to handle more complex scenarios. *C-BIoU* [21] improves traditional IoU-based matching by introducing a buffer region around the bounding boxes, enhancing matching accuracy under occlusion. *EagerMOT* [7] incorporates velocity and directionality into the association process, leveraging motion cues to improve association reliability. Another advanced metric, *DIoU* [28], introduces distance and aspect ratio terms into the IoU calculation, improving matching accuracy in 2D scenarios. These identity association techniques aim to reduce false positives and increase the robustness of tracking in dynamic environments where objects may be occluded or exhibit erratic movement.

3) Trajectory Management: The final stage of MOT involves managing the trajectories of objects across frames. The goal is to maintain accurate object tracks while adding, updating, or removing object trajectories as necessary. Early methods such as *SORT* [2] use a simple bipartite graph matching approach to link detected objects between consecutive frames. This method provides a fast and efficient solution but is limited in handling occlusions and long-term tracking.

More advanced techniques, such as *DeepSort* [19], use a cascade matching strategy where recently updated trajectories are prioritized during the matching process, allowing for better handling of short-term occlusions. *ByteTrack* [26] takes this approach further by splitting detections into high and low-confidence categories, first matching high-confidence detections with tracked objects, and then using low-confidence detections to match occluded objects. This technique improves the continuity of object trajectories during occlusions. Recent methods like *SimpleTrack* [12] adapt ByteTrack's methodology to 3D tracking tasks, introducing confidence-based updates to better handle occlusion in real-world environments. Another recent method, *SimTrack* [11], criticizes the heuristic matching approach of earlier methods, arguing that it relies too heavily on manual tuning. Instead, *SimTrack* proposes using a confidence-weighted matching strategy between the predicted object locations and the detected objects, simplifying the tracking process while improving robustness. These trajectory management strategies aim to maintain object continuity across time, especially when objects undergo occlusions or disappear from the field of view temporarily.

Despite the progress made, several challenges remain in multi-object tracking, especially in 3D environments where spatial and temporal dynamics are more complex. One major limitation is the failure to fully integrate the strengths of both object detection and tracking, which are often treated as independent tasks. Furthermore, current 3D identity association metrics do not make full use of the precise pose information available in point clouds, hindering matching accuracy. Moreover, cross-frame matching strategies fail to leverage both detection and prediction dimensions to fully explore and maintain object identity, particularly in complex and dynamic environments. These gaps highlight areas for further research and improvement.

III. Methodology

In the case where the vehicle and infrastructure sensors collect data at consistent frequencies, the input for the vehicle-infrastructure cooperative 3D multi-object tracking task can be described in two parts:

1. The infrastructure-side sensor data sequence $\mathbf{S}^{\text{inf}}$, as shown in Eq. (1), where $\mathbf{X}_n^{\text{inf}}$ represents the infrastructure sensor data from the n-th pair of vehicle-infrastructure matched data. The corresponding timestamp sequence for the infrastructure data is denoted as $\mathbf{T}^{\text{inf}}$, as shown in Eq. (2).

$$\mathbf{S}^{\text{inf}} = \{\mathbf{X}_n^{\text{inf}}\}_{n=1}^{N} \tag{1}$$

$$\mathbf{T}^{\text{inf}} = \{t_n^{\text{inf}}\}_{n=1}^{N} \tag{2}$$

2. The vehicle-side sensor data sequence $\mathbf{S}^{\text{veh}}$, as shown in Eq. (3), where $\mathbf{X}_n^{\text{veh}}$ represents the vehicle sensor data from the n-th pair of vehicle-infrastructure matched data. The corresponding timestamp sequence for the vehicle data is denoted as $\mathbf{T}^{\text{veh}}$, as shown in Eq. (4).

$$\mathbf{S}^{\text{veh}} = \{\mathbf{X}_n^{\text{veh}}\}_{n=1}^{N} \tag{3}$$

$$\mathbf{T}^{\text{veh}} = \{t_n^{\text{veh}}\}_{n=1}^{N} \tag{4}$$

As shown in Fig. 1, the vehicle-infrastructure cooperative 3D multi-object tracking process is composed of two main parts: information extraction and fusion, and object detection and tracking. These are further divided into four stages:

- Data Transmission and Fusion: This stage involves pre-processing vehicle and infrastructure point cloud data. The data is transmitted over limited bandwidth

to the vehicle side, where it is fused with the vehicle's sensor data.

- Object Detection and Feature Representation: This stage detects 3D objects from the fused sensor data and represents the temporal variations in the object features.
- Identity Association: The system builds a unified identity association benchmark for objects across frames based on their features. A cost matrix is generated to match objects across frames.
- Trajectory Management: Based on the cost matrix from identity association, the system performs matching between detected objects and existing tracks. It also handles adding, deleting, and updating tracks, ultimately producing a unified 3D spatiotemporal perception result with identity information.

A. Data Transmission and Fusion Using the FF-Tracking Framework

This section leverages the FF-Tracking [25] framework as outlined in *V2X-Seq*. The **Data Transmission and Fusion** process, as illustrated in Fig. 2, is designed to facilitate sensor data exchange between the vehicle and infrastructure under limited bandwidth conditions.

1) Infrastructure Side: On the infrastructure side, sensor data at timestamp t_n^{inf} is processed through the Pillar Feature Network (PFNet) [9] to extract BEV (Bird's Eye View) features. The extracted BEV feature map is represented as:

$$\mathbf{F}_n^{\text{inf}} = \text{PFNet}(\mathbf{X}_n^{\text{inf}}), \tag{5}$$

where $\mathbf{X}_n^{\text{inf}}$ is the raw point cloud data from the infrastructure sensors, and $\mathbf{F}_n^{\text{inf}}$ is the BEV feature extracted at time t_n^{inf}.

Next, a Feature Flow Generator is employed to capture the temporal dynamics between consecutive infrastructure frames. Given the BEV feature map from the current timestamp t_n^{inf} and the previous timestamp t_{n-1}^{inf}, the feature flow $\mathbf{F}_n^{\text{flow}}$ is computed as:

$$\mathbf{F}_n^{\text{flow}} = \text{FlowGenerator}(\mathbf{F}_n^{\text{inf}}, \mathbf{F}_{n-1}^{\text{inf}}), \tag{6}$$

where the *FlowGenerator* module is built with two Backbone-FPN structures, one branch generates static features $\mathbf{F}^{\text{static}}$, and the other branch generates the feature derivative $\mathbf{F}^{\text{deriv}}$. This module computes the temporal differences between the two frames, capturing how the scene evolves over time. The feature flow is compressed using a convolutional layer to reduce bandwidth requirements:

$$\mathbf{F}_n^{\text{comp}} = \text{Conv}(\mathbf{F}_n^{\text{flow}}), \tag{7}$$

where $\mathbf{F}_n^{\text{comp}}$ represents the compressed feature flow ready for transmission.

2) Vehicle Side: Once the infrastructure-side compressed feature flow $\mathbf{F}_n^{\text{comp}}$ is transmitted to the vehicle side, it undergoes Deconvolution to reconstruct the infrastructure features at the vehicle's side:

$$\mathbf{F}_n^{\text{rec}} = \text{Deconv}(\mathbf{F}_n^{\text{comp}}), \tag{8}$$

where $\mathbf{F}_n^{\text{rec}}$ denotes the reconstructed feature map, which can be divided into two parts: one part represents the static feature $\mathbf{F}_n^{\text{static}}$, and the other part represents the feature derivative $\mathbf{F}_n^{\text{deriv}}$.

Next, the vehicle aligns the reconstructed infrastructure features to its own current timestamp t_n^{veh} using the Prediction and Affine Transform module. This module compensates for temporal misalignment between the infrastructure and vehicle data:

$$\mathbf{F}_n^{\text{align}} = \text{AffineTransform}(\mathbf{F}_n^{\text{static}} + (t_n^{\text{veh}} - t_n^{\text{inf}}) \cdot \mathbf{F}_n^{\text{deriv}}), \tag{9}$$

where $\mathbf{F}_n^{\text{align}}$ represents the aligned infrastructure features in the vehicle's frame of reference.

Simultaneously, the vehicle extracts its own features $\mathbf{F}_n^{\text{veh}}$ from its sensor data $\mathbf{X}_n^{\text{veh}}$:

$$\mathbf{F}_n^{\text{veh}} = \text{FeatureExtractor}(\mathbf{X}_n^{\text{veh}}). \tag{10}$$

Finally, the vehicle-side features and the aligned infrastructure features are concatenated and convolved to form the fused feature map $\mathbf{F}_n^{\text{fused}}$:

$$\mathbf{F}_n^{\text{fused}} = \text{Conv}(\text{Concat}(\mathbf{F}_n^{\text{veh}}, \mathbf{F}_n^{\text{align}})). \tag{11}$$

This fused feature map, containing both vehicle-side and infrastructure-side data, is passed to the detection and tracking modules for further processing.

B. Integrated Object Detection and Tracking

In this section, we describe the proposed integrated object detection and tracking framework, utilizing a Deformable DETR-based architecture for both encoding and decoding stages, as shown in Fig. 3. The architecture is composed of the following core components: **Temporal Prior Enhancement**, **Encoder**, and two parallel branches for **Object Detection** and **Object Prediction**.

1) Temporal Prior Enhancement: The input to the Temporal Prior Enhancement module consists of the fused feature maps from two consecutive frames, $\mathbf{F}_n^{\text{fused}}$ and $\mathbf{F}_{n-1}^{\text{fused}}$. To save computational resources, the fused feature map from the previous frame is stored and used in the next frame for prior enhancement. The previous frame's fused feature map, $\mathbf{F}_{n-1}^{\text{fused}}$, is downsampled and serves as the keys ($\mathbf{K}$) and values ($\mathbf{V}$) for the cross-attention mechanism. The current frame's fused feature map, $\mathbf{F}_n^{\text{fused}}$, is used as the queries ($\mathbf{Q}$).

The cross-attention mechanism computes the weighted combination of the features as follows:

$$\text{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d_k}}\right)\mathbf{V}, \tag{12}$$

where $\mathbf{Q} = \mathbf{F}_n^{\text{fused}}$ is the query from the current frame, and $\mathbf{K}$, $\mathbf{V} = \mathbf{F}_{n-1}^{\text{fused}}$ are the key and value from the previous frame. The term d_k represents the dimension of the key, which is used to scale the dot product between the query and key to prevent the result from becoming too large. After cross-attention, the resulting feature map undergoes further refinement through a Multilayer Perceptron (MLP) for introducing non-linearity and enhancing the model's capacity to capture complex relationships in the data:

$$\mathbf{F}_n^{\text{enhanced}} = \text{MLP}(\text{Attention}(\mathbf{F}_n^{\text{fused}}, \mathbf{F}_{n-1}^{\text{fused}})). \tag{13}$$

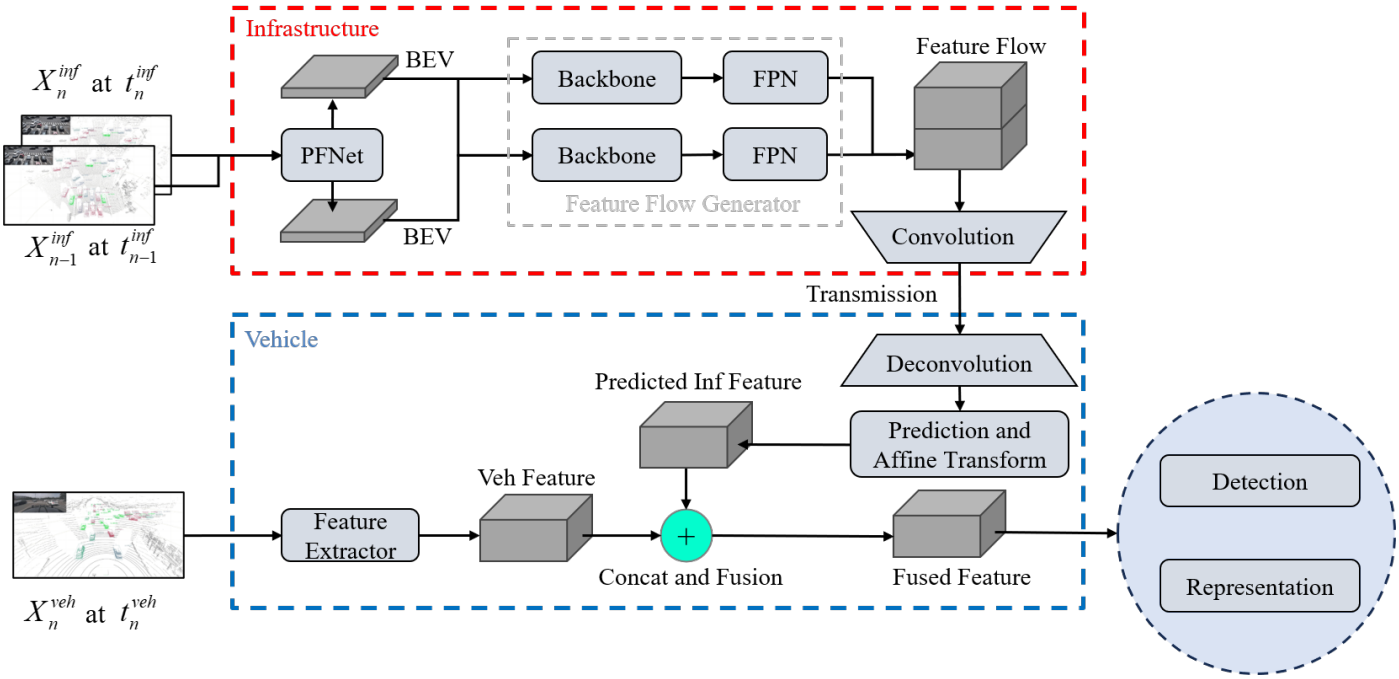

Fig. 2. The architecture of the data transmission and fusion module using the FF-Tracking framework. Infrastructure-side features are compressed and transmitted to the vehicle, where they are deconvolved, temporally and spatially aligned, and fused with vehicle-side data.

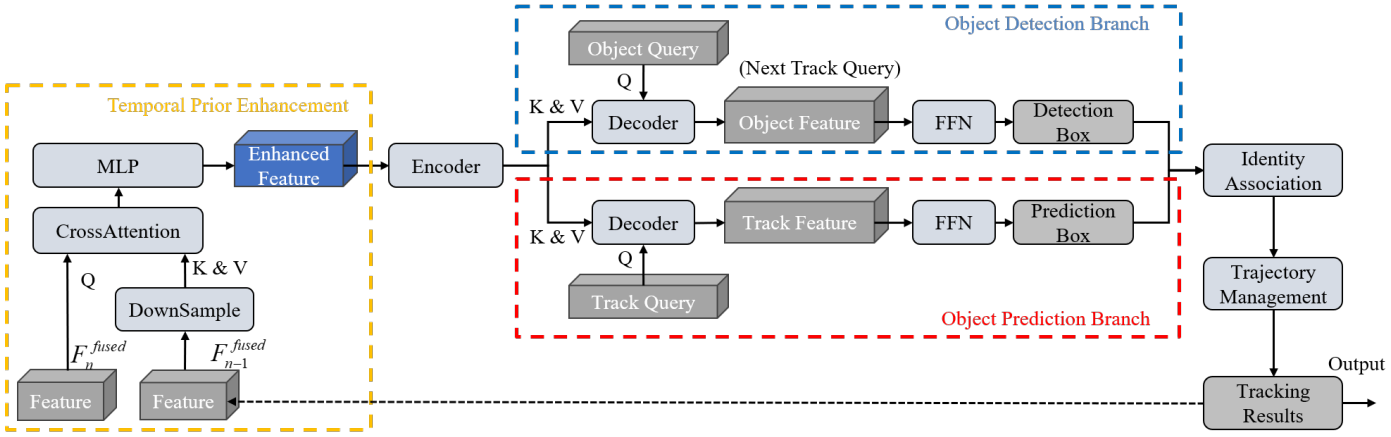

Fig. 3. The overall architecture for integrated object detection and tracking using deformable DETR. The process includes temporal prior enhancement, encoding, and two parallel branches for object detection and prediction, which jointly contribute to tracking results.

The output is the enhanced feature map $\mathbf{F}_n^{\text{enhanced}}$, which integrates temporal information from the previous frame to improve the accuracy and stability of detection and tracking in the current frame.

2) Encoder: The enhanced feature $\mathbf{F}_n^{\text{enhanced}}$ is passed to a Deformable DETR Encoder [30], which applies deformable attention over a set of reference points $\mathbf{P}$ distributed across the feature map. These reference points guide the attention mechanism, allowing it to focus on relevant regions, which is particularly suitable for sparse input data like point clouds:

$$\mathbf{Z}_n^{\text{enc}} = \text{DeformableAttention}(\mathbf{F}_n^{\text{enhanced}}, \mathbf{K}, \mathbf{V}, \mathbf{P}), \quad (14)$$

where $\mathbf{Z}_n^{\text{enc}}$ is the output of the encoder, and $\mathbf{K}$, $\mathbf{V}$ are derived from the same enhanced feature map $\mathbf{F}_n^{\text{enhanced}}$.

3) Object Detection Branch: The Object Detection Branch uses the encoded feature $\mathbf{Z}_n^{\text{enc}}$ and processes it through a Deformable DETR decoder to detect objects within the current frame. The decoder utilizes a set of learnable object queries $\mathbf{Q}^{\text{obj}}$ to retrieve object features:

$$\mathbf{F}_n^{\text{obj}} = \text{Decoder}(\mathbf{Q}^{\text{obj}}, \mathbf{Z}_n^{\text{enc}}), \quad (15)$$

where $\mathbf{F}_n^{\text{obj}}$ represents the detected object features. These features are passed to a feedforward neural network (FFN) to generate detection bounding boxes $\mathbf{B}_n^{\text{det}}$:

$$\mathbf{B}_n^{\text{det}} = \text{FFN}(\mathbf{F}_n^{\text{obj}}). \quad (16)$$

4) Object Prediction Branch: Simultaneously, the Object Prediction Branch tracks objects by predicting their locations

in the next frame based on past tracking information. The encoded feature $\mathbf{Z}_n^{\text{enc}}$ is processed with the track query $\mathbf{Q}^{\text{track}}$ which is the object features $\mathbf{F}_{n-1}^{\text{obj}}$ in the previous frame, retrieving the tracking features $\mathbf{F}_n^{\text{track}}$:

$$\mathbf{F}_n^{\text{track}} = \text{Decoder}(\mathbf{Q}^{\text{track}}, \mathbf{Z}_n^{\text{enc}}). \tag{17}$$

These features are processed through another FFN to predict the locations of objects from the previous frame in the current frame, providing the propagated positions for the tracked objects $\mathbf{B}_n^{\text{pred}}$:

$$\mathbf{B}_n^{\text{pred}} = \text{FFN}(\mathbf{F}_n^{\text{track}}). \tag{18}$$

5) Identity Association and Trajectory Management: The detected objects and predicted objects are matched using an identity association module, which computes a cost matrix between detection and prediction bounding boxes. This cost matrix is used to associate objects between frames. The tracking results are then managed in the trajectory management module, which updates existing trajectories, adds new trajectories, and deletes lost trajectories.

6) Advantages of Integrated Design: The integrated design of this architecture, based on Deformable DETR, allows simultaneous object detection and tracking within the same pipeline. By sharing the same enhanced features and attention mechanisms between detection and prediction branches, the architecture efficiently combines object detection and tracking tasks. This integration fully leverages the complementary relationship between detection and tracking, as the temporal prior information enhances the consistency of the features, allowing detection and tracking to mutually benefit from each other's information.

C. Identity Association Using XIOU Metric

In our approach, we propose a novel **XIOU** metric for identity association between detected and predicted objects. This metric incorporates three key factors: Intersection over Union (IOU), center point distance, and yaw angle difference between the two bounding boxes (as shown in Fig. 4).

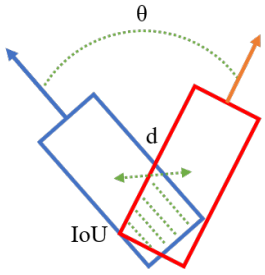

Fig. 4. Visualization of XIOU elements: IOU, center point distance, and yaw angle difference.

First, the basic IOU is calculated between the detection bounding box $\mathbf{B}^{\text{det}}$ and the prediction bounding box $\mathbf{B}^{\text{pred}}$, which measures the overlap between the two boxes:

$$\text{IOU}(\mathbf{B}^{\text{det}}, \mathbf{B}^{\text{pred}}) = \frac{V_{\text{I}}}{V_{\text{U}}}, \tag{19}$$

where V_{I} is the intersection volume of the two 3D bounding boxes, and V_{U} is their union volume.

To improve upon the limitations of IOU in cases where there is no overlap between the two boxes, the Generalized IOU (GIOU) [13] metric is used:

$$\text{GIOU}(\mathbf{B}^{\text{det}}, \mathbf{B}^{\text{pred}}) = \frac{V_{\text{I}}}{V_{\text{U}}} - \frac{V_{\text{C}} - V_{\text{U}}}{V_{\text{C}}}, \tag{20}$$

where V_{C} is the volume of the smallest convex shape enclosing both $\mathbf{B}^{\text{det}}$ and $\mathbf{B}^{\text{pred}}$. GIOU extends IOU by adding a distance-based penalty term, addressing cases where IOU is zero due to non-overlapping boxes.

Building on GIOU, our **XIOU** metric further integrates orientation and distance between the center points of the two boxes:

$$G_{\cos}(\theta_{\mathbf{B}^{\text{det}}}, \theta_{\mathbf{B}^{\text{pred}}}) = \cos(\theta_{\mathbf{B}^{\text{det}}} - \theta_{\mathbf{B}^{\text{pred}}}) + 1, \tag{21}$$

$$G_{\text{iou}}(\mathbf{B}^{\text{det}}, \mathbf{B}^{\text{pred}}) = \text{GIOU}(\mathbf{B}^{\text{det}}, \mathbf{B}^{\text{pred}}) + 1, \tag{22}$$

$$\text{XIOU}(\mathbf{B}^{\text{det}}, \mathbf{B}^{\text{pred}}) = \frac{G_{\text{iou}} \times G_{\cos}}{4}, \tag{23}$$

where $\theta_{\mathbf{B}^{\text{det}}}$ and $\theta_{\mathbf{B}^{\text{pred}}}$ represent the yaw angles (orientations) of the detection and prediction bounding boxes, respectively. This term captures the orientation difference between the boxes, while the GIOU term handles their spatial relationship.

Our **XIOU** metric offers improved performance in 3D environments by taking into account the spatial overlap, orientation similarity, and distance between objects. This makes it particularly well-suited for 3D object tracking tasks, where precise alignment of object positions and orientations is crucial for maintaining consistent identities across frames.

D. Trajectory Management with Cascade Matching

Trajectory management plays a critical role in maintaining accurate object tracking over time. This step involves matching detected objects with existing trajectories, updating object tracks, and handling the creation or deletion of tracks as necessary. Traditional approaches, such as ByteTrack, employ a two-stage matching process, starting with high-confidence detections followed by low-confidence ones. However, this approach struggles in occlusion scenarios. As shown in Fig. 5(a), object detection confidence gradually decreases during occlusion, while Fig. 5(b) shows how confidence slowly recovers when occlusion fades.

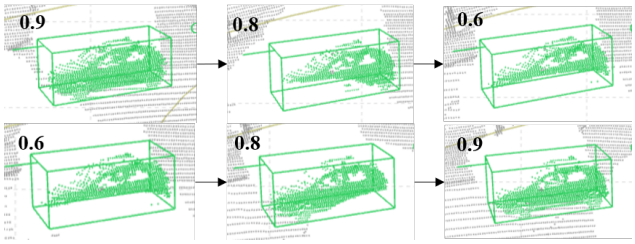

Fig. 5. (a) Confidence decline during occlusion. (b) Confidence recovery after occlusion. The figures illustrate how detection confidence changes when objects are occluded and when occlusion diminishes.

To address the challenges of tracking objects under occlusion, we introduce a **Four-Stage Cascade Matching (FSCM)** algorithm. This method improves upon previous approaches by dividing detection results and track predictions

into high-confidence and low-confidence categories, handling them across four stages. Each stage applies the Hungarian algorithm to perform the matching based on the XIOU similarity metric, which considers object overlap, orientation, and spatial alignment.

1) Stage 1: High-confidence Detection and High-confidence Track Matching: In the first stage, detections and tracks are divided into high-confidence and low-confidence sets based on predefined thresholds. A detection is considered high-confidence if its detection score s_d exceeds the detection threshold m_d, and similarly, a track is considered high-confidence if its tracking confidence score s_t exceeds the tracking threshold m_t. High-confidence detections $\mathbf{D}^{\text{high}}$ are matched with high-confidence tracks $\mathbf{T}^{\text{high}}$. The cost matrix is computed as:

$$\mathbf{C}^{\text{high}} = \mathbf{1} - \text{XIOU}(\mathbf{D}^{\text{high}}, \mathbf{T}^{\text{high}}). \tag{24}$$

The Hungarian [8] algorithm is applied to minimize the cost matrix $\mathbf{C}^{\text{high}}$. The matched detections and tracks are processed, while the unmatched ones are passed to the next stage.

2) Stage 2: Low-confidence Detection and High-confidence Track Matching: Low-confidence detections ($s_d < m_d$), denoted as $\mathbf{D}^{\text{low}}$, are matched with the high-confidence tracks ($s_t \geq m_t$) that remained unmatched from the previous stage. This is useful for handling partially occluded objects whose detection scores have decreased, while their tracking predictions remain reliable. The cost matrix is calculated as:

$$\mathbf{C}^{\text{low-high}} = \mathbf{1} - \text{XIOU}(\mathbf{D}^{\text{low}}, \mathbf{T}^{\text{high}}), \tag{25}$$

and the Hungarian algorithm matches the remaining high-confidence tracks with low-confidence detections.

3) Stage 3: High-confidence Detection and Low-confidence Track Matching: In this stage, high-confidence detections $\mathbf{D}^{\text{high}}$ are matched with low-confidence tracks $\mathbf{T}^{\text{low}}$, which were not matched in the previous stages. This helps recover objects that were previously occluded but are now detected with high confidence:

$$\mathbf{C}^{\text{high-low}} = \mathbf{1} - \text{XIOU}(\mathbf{D}^{\text{high}}, \mathbf{T}^{\text{low}}). \tag{26}$$

Again, the Hungarian algorithm is used to assign detections to tracks, updating the trajectories for reappearing objects.

4) Stage 4: Low-confidence Detection and Low-confidence Track Matching: Finally, low-confidence detections $\mathbf{D}^{\text{low}}$ are matched with low-confidence tracks $\mathbf{T}^{\text{low}}$. This step manages prolonged occlusion or potential false detections:

$$\mathbf{C}^{\text{low}} = \mathbf{1} - \text{XIOU}(\mathbf{D}^{\text{low}}, \mathbf{T}^{\text{low}}). \tag{27}$$

The Hungarian algorithm minimizes the cost matrix, and matched detections and tracks are processed.

5) Unmatched Detections and Tracks: If any high-confidence detections remain unmatched, they are used to initialize new tracks. Unmatched low-confidence detections are discarded as background noise. Tracks that remain unmatched are retained for N frames. If tracks remain unmatched beyond the threshold, they are deleted from the system.

6) Summary of FSCM Algorithm: This four-stage cascade matching algorithm divides both detections and predictions into high-confidence and low-confidence categories, leveraging the XIOU metric at each stage. The Hungarian algorithm is employed in all stages to ensure optimal matching. By progressively refining the matching process, this method ensures robust tracking, even under occlusion, and takes full advantage of detection and prediction confidences.

IV. Experiments

The primary goal of our experiments is to validate the effectiveness and robustness of the proposed method for vehicle-infrastructure cooperative 3D multi-object tracking. We conduct a series of experiments on the V2X-Seq dataset to evaluate the tracking performance under varying latency conditions and compare it with several state-of-the-art methods. Additionally, we perform an ablation study to investigate the contributions of the key modules in our architecture, including the integrated detection-tracking framework, XIOU identity association, and four-stage cascade matching (FSCM).

A. Dataset and Evaluation Metrics

Our experiments are conducted on the V2X-Seq dataset, the first large-scale real-world dataset specifically designed for vehicle-infrastructure cooperative 3D multi-object tracking. It contains over 15,000 pairs of synchronized vehicle-side and infrastructure-side frames, with each pair including 3D LiDAR point clouds and annotations with tracking IDs. All vehicle and infrastructure data in V2X-Seq are time-synchronized and spatially aligned, making it ideal for evaluating cooperative tracking performance in real-world scenarios. With over 150,000 frames across more than 200 sequences, V2X-Seq provides diverse traffic environments and object dynamics, offering a comprehensive benchmark for assessing tracking accuracy and robustness, especially under challenging conditions such as occlusions and communication delays. We use the following evaluation metrics to compare the performance of different methods:

- MOTA (Multi-Object Tracking Accuracy): MOTA reflects the overall tracking accuracy by considering three factors: false positives, missed targets, and identity switches. Higher values indicate better performance.

$$\text{MOTA} = 1 - \frac{\sum_t (\text{FN}_t + \text{FP}_t + \text{IDSW}_t)}{\sum_t \text{GT}_t}, \tag{28}$$

  where FN_t, FP_t, and IDSW_t are the false negatives, false positives, and identity switches at time t, and GT_t is the number of ground truth objects.

- MOTP (Multi-Object Tracking Precision): MOTP evaluates the localization precision of the tracked objects by computing the average distance between the predicted and ground-truth object locations.

$$\text{MOTP} = \frac{\sum_t \sum_i d_t^i}{\sum_t c_t}, \tag{29}$$

  where d_t^i is the distance between the predicted and ground-truth locations for object i at time t, and c_t is the number of matched object pairs at time t.

TABLE I. COMPARISON OF TRACKING PERFORMANCE ON THE V2X-SEQ DATASET UNDER DIFFERENT LATENCY CONDITIONS

Latency(ms)	Fusion Type	Method	MOTA↑	MOTP↑	IDS↓	BPS↓(Byte/s)
0	Object-Level	Hungarian [8]	53.18	72.35	273	3.3×10^3
	Data-Level	Concat [25]	56.03	70.17	296	1.3×10^7
	Feature-Level	FF-Tracking [25]	54.75	69.76	222	6.2×10^5
	Feature-Level	**Ours**	**57.23**	**74.64**	**206**	6.2×10^5
200	Object-Level	Hungarian [8]	50.32	71.58	260	3.3×10^3
	Data-Level	Concat [25]	51.27	69.67	234	1.3×10^7
	Feature-Level	FF-Tracking [25]	52.26	69.64	225	1.2×10^6
	Feature-Level	**Ours**	**55.76**	**74.15**	**219**	1.2×10^6

- IDS (Identity Switches): This metric tracks identity changes during tracking, with lower values indicating better performance.
- BPS (Bytes Per Second): This metric measures the bandwidth for vehicle-infrastructure communication, defined as the data exchanged per second in bytes.

B. Implementation Details

Our model employs a backbone network and FPN structure similar to that used in SECOND [20], optimized using the AdamW optimizer. The initial learning rate is set to 1×10^{-4}, and we use 500 queries during training. Both decoders are trained with identical loss functions, incorporating a classification loss and XIOU loss as the final objective. For the tracking process, we set the detection score threshold m_d to 0.5, tracking score threshold m_t to 0.4, and retain unmatched trajectories for $N = 20$ frames. The network is implemented in Pytorch and trained on an NVIDIA GeForce RTX 3090 GPU.

In the inference phase, the first frame's feature map serves as the prior frame's feature for downsampling and temporal prior enhancement. Simultaneously, the track query initializes with the object features from the first frame. From the second frame onward, the point cloud feature sequence is processed following the methodology described, outputting tracking results across all frames.

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

We compare the performance of our method against several approaches in the V2X-Seq dataset under two latency conditions: 0 ms and 200 ms. The methods include data-level, object-level, and feature-level fusion techniques, specifically:

- Concat (Data-Level) [25]: In this method, the infrastructure point cloud is transformed to the vehicle's coordinate system, where pseudo-images from both vehicle and infrastructure are concatenated. This approach uses the PointPillars detector and follows AB3DMOT's TBD (tracking-by-detection) paradigm for multi-object tracking.
- Hungarian (Object-Level) [8]: In this approach, vehicle and infrastructure detections are performed independently. Detected object sets from both vehicle and infrastructure are transmitted and matched using the Hungarian algorithm to fuse results.
- FF-Tracking (Feature-Level) [25]: This method transmits the feature flow between consecutive frames from the infrastructure to the vehicle, reducing data transmission while maintaining accuracy.

As shown in Table I, our method consistently outperforms others across key metrics, particularly in **MOTA**, achieving 57.23 at 0 ms latency—**1.2 points** higher than Concat (56.03) and **4.05 points** higher than Hungarian (53.18). This improvement reflects the combined contributions of our integrated detection-tracking framework, XIOU identity association, and four-stage cascade matching (FSCM). In terms of **MOTP**, our method achieves 74.64, surpassing Concat's 70.17 and Hungarian's 72.35, highlighting its effective use of 3D positional and orientational information. Additionally, with the lowest **IDS** score of 206, it demonstrates robust identity association and trajectory management.

Under the 200 ms delay condition, our method maintains high tracking accuracy with a **MOTA** of 55.76, outperforming Concat (51.27) and Hungarian (50.32). The **MOTP** remains at 74.15, the highest among all methods, while the **IDS** count remains low at 219. These results demonstrate the resilience of our approach to communication delays, inheriting the robustness of the FF-Tracking framework. By preserving temporal consistency in feature flow and leveraging efficient identity association, the proposed framework effectively mitigates the negative impact of delayed data transmission.

In terms of data transmission efficiency, our method achieves a transmission rate of 6.2×10^5 Byte/s, which is over **20 times** lower than Concat (1.3×10^7 Byte/s). This substantial reduction is achieved through feature-level fusion, which transmits compressed feature flows. Despite this lower bandwidth usage, our method provides a **4%** higher MOTA, highlighting its ability to optimize communication resources while improving tracking accuracy.

D. Ablation Study

We conduct an ablation study to quantify the contributions of our three key modules: the integrated detection-tracking framework, XIOU identity association, and the four-stage cascade matching (FSCM) algorithm. The baseline method, AB3DMOT, is used for comparison by systematically replacing each module in our framework with the corresponding component from AB3DMOT. The results are presented in Table II.

1) Impact of the Integrated Detection-Tracking Framework: Replacing our integrated detection-tracking framework with AB3DMOT's tracking-by-detection (TBD) paradigm results in a **2.1-point decrease in MOTA** (from 57.23 to 55.13). This indicates the unified framework's critical role in bridging detection and tracking, enabling the detection branch to benefit from tracking priors while allowing the tracking branch to leverage enhanced detection results. The observed drop in

TABLE II. ABLATION STUDY ON THE V2X-SEQ DATASET

Method	Tracking Framework	Identity Association	Trajectory Management	MOTA↑	MOTP↑	IDS↓
AB3DMOT	TBD	IOU	Single Matching	54.75	69.76	222
Ours	**Integrated**	**XIOU**	**FSCM**	**57.23**	**74.64**	**206**
1	TBD	XIOU	FSCM	55.13	72.43	208
2	Integrated	IOU	FSCM	56.72	73.58	214
3	Integrated	XIOU	Single Matching	55.38	74.16	215

MOTA demonstrates that separating detection and tracking increases errors, for example, in scenarios involving fast-moving or partially occluded objects. By integrating detection and tracking within the same pipeline, our framework effectively reduces identity switches and improves object recall, ensuring robust performance in dynamic traffic environments.

2) Impact of XIOU Identity Association: When XIOU is replaced with AB3DMOT's IOU, **MOTA drops by 0.51 points** (from 57.23 to 56.72), and **IDS increases by 3.9%** (from 206 to 214). This demonstrates XIOU's ability to capture spatial and orientation consistency, which is particularly beneficial in occlusion-heavy environments. XIOU effectively resolves ambiguous matches between detection and prediction bounding boxes by incorporating yaw angle and center-point distance, leading to improved identity consistency and reduced errors during complex interactions between vehicles. In contrast, IOU struggles to maintain identity consistency when objects overlap or move in close proximity, leading to more identity switches and reduced tracking accuracy.

3) Impact of the Four-Stage Cascade Matching (FSCM): Replacing our four-stage cascade matching (FSCM) algorithm with AB3DMOT's single matching strategy increases **IDS by 4.3%** (from 206 to 215) and reduces **MOTA by 1.85 points** (from 57.23 to 55.38). These results highlight FSCM's ability to manage trajectory updates effectively, particularly in handling occlusions and reappearing objects. FSCM dynamically adapts to the confidence levels of both detections and tracks, ensuring robust identity associations across frames. Single matching, on the other hand, lacks this flexibility, resulting in higher identity switches and degraded tracking performance, particularly in challenging scenarios with frequent occlusions or sudden object reappearances. By incorporating FSCM, our method achieves lower IDS and higher MOTA, demonstrating its importance for maintaining accurate and consistent trajectories under complex real-world conditions.

4) Module Contribution Analysis: Among the three modules, the integrated detection-tracking framework contributes the largest MOTA gain (**2.1 points**), highlighting its significant impact on overall tracking accuracy. FSCM provides the second-highest gain (**1.85 points in MOTA**), underscoring its importance in trajectory management under challenging conditions. XIOU, while contributing a relatively smaller MOTA improvement (**0.51 points**), plays a crucial role in reducing IDS, demonstrating its effectiveness in identity association tasks. Collectively, these modules form a robust system that achieves superior performance compared to traditional methods.

E. Challenges and Future Directions

1) Bandwidth Efficiency: Although our method significantly reduces data transmission compared to data-level fusion methods, the bandwidth requirement (6.2×10^5 Byte/s) remains relatively high compared to object-level methods like Hungarian. This poses challenges for large-scale deployment in real-world bandwidth-constrained environments. Future work should focus on optimizing feature extraction and compression strategies to further reduce transmission overhead while maintaining tracking accuracy.

2) Handling Long Occlusions and Disappearances: While the proposed framework effectively addresses moderate occlusions and identity switches, it struggles in scenarios involving long-term occlusions or complete object disappearances. For instance, re-associating objects after prolonged absence remains challenging. Future efforts could focus on incorporating adaptive temporal modeling techniques and improved motion prediction strategies to enhance the system's robustness in such complex and dynamic environments.

V. CONCLUSION

In this work, we proposed an innovative framework for vehicle-infrastructure cooperative 3D multi-object tracking, emphasizing three key contributions: an integrated detection-tracking framework, the XIOU identity association metric, and a four-stage cascade matching (FSCM) strategy. The integrated framework enhances both detection accuracy and tracking consistency by jointly leveraging detection and tracking information. The XIOU metric improves identity association by effectively incorporating 3D spatial information, while FSCM provides robust tracking continuity in occlusion scenarios. Experimental results on the V2X-Seq dataset validate the effectiveness of these innovations, with our method demonstrating superior tracking accuracy, reduced identity switches, and low bandwidth usage even under delayed communication conditions. These results underscore the potential of feature-level fusion and temporal prior enhancement in real-world V2X applications. Future work will focus on optimizing bandwidth efficiency through improved feature extraction and compression, and enhancing robustness in handling long-term occlusions and dynamic scenarios with adaptive temporal modeling and motion prediction, paving the way for more reliable and efficient V2X applications.

ACKNOWLEDGMENTS

This work was supported in part by the International Strategic Innovative Project of the National Key R&D Program of China (2023YFE0112500) and the Fundamental Research Funds for the Central Universities (22120230311).

REFERENCES

[1] Eduardo Arnold, Mehrdad Dianati, Robert de Temple, and Saber Fallah. Cooperative perception for 3d object detection in driving scenarios using infrastructure sensors. *IEEE Transactions on Intelligent Transportation Systems*, 23(3):1852–1864, 2020.

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

[3] Hsu-kuang Chiu, Jie Li, Rareş Ambruş, and Jeannette Bohg. Probabilistic 3d multi-modal, multi-object tracking for autonomous driving. In *2021 IEEE international conference on robotics and automation (ICRA)*, pages 14227–14233. IEEE, 2021.

[4] Peng Chu, Jiang Wang, Quanzeng You, Haibin Ling, and Zicheng Liu. Transmot: Spatial-temporal graph transformer for multiple object tracking. In *Proceedings of the IEEE/CVF Winter Conference on applications of computer vision*, pages 4870–4880, 2023.

[5] Siqi Fan, Haibao Yu, Wenxian Yang, Jirui Yuan, and Zaiqing Nie. Quest: Query stream for practical cooperative perception. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 18436–18442. IEEE, 2024.

[6] Yue Hu, Shaoheng Fang, Zixing Lei, Yiqi Zhong, and Siheng Chen. Where2comm: Communication-efficient collaborative perception via spatial confidence maps. *Advances in neural information processing systems*, 35:4874–4886, 2022.

[7] Aleksandr Kim, Aljoša Ošep, and Laura Leal-Taixé. Eagermot: 3d multi-object tracking via sensor fusion. In *2021 IEEE International conference on Robotics and Automation (ICRA)*, pages 11315–11321. IEEE, 2021.

[8] Harold W Kuhn. The hungarian method for the assignment problem. *Naval Research Logistics (NRL)*, 52(1):7–21, 2004.

[9] Alex H Lang, Sourabh Vora, Holger Caesar, Lubing Zhou, Jiong Yang, and Oscar Beijbom. Pointpillars: Fast encoders for object detection from point clouds. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12697–12705, 2019.

[10] Yiming Li, Dekun Ma, Ziyan An, Zixun Wang, Yiqi Zhong, Siheng Chen, and Chen Feng. V2x-sim: Multi-agent collaborative perception dataset and benchmark for autonomous driving. *IEEE Robotics and Automation Letters*, 7(4):10914–10921, 2022.

[11] Chenxu Luo, Xiaodong Yang, and Alan Yuille. Exploring simple 3d multi-object tracking for autonomous driving. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10488–10497, 2021.

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

[13] Hamid Rezatofighi, Nathan Tsoi, JunYoung Gwak, Amir Sadeghian, Ian Reid, and Silvio Savarese. Generalized intersection over union: A metric and a loss for bounding box regression. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 658–666, 2019.

[14] Sanbao Su, Yiming Li, Sihong He, Songyang Han, Chen Feng, Caiwen Ding, and Fei Miao. Uncertainty quantification of collaborative detection for self-driving. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pages 5588–5594. IEEE, 2023.

[15] Peize Sun, Jinkun Cao, Yi Jiang, Rufeng Zhang, Enze Xie, Zehuan Yuan, Changhu Wang, and Ping Luo. Transtrack: Multiple object tracking with transformer. *arXiv preprint arXiv:2012.15460*, 2020.

[16] Junke Wang, Dongdong Chen, Zuxuan Wu, Chong Luo, Xiyang Dai, Lu Yuan, and Yu-Gang Jiang. Omnitracker: Unifying object tracking by tracking-with-detection. *arXiv preprint arXiv:2303.12079*, 2023.

[17] Tsun-Hsuan Wang, Sivabalan Manivasagam, Ming Liang, Bin Yang, Wenyuan Zeng, and Raquel Urtasun. V2vnet: Vehicle-to-vehicle communication for joint perception and prediction. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16*, pages 605–621. Springer, 2020.

[18] Xinshuo Weng, Jianren Wang, David Held, and Kris Kitani. 3d multi-object tracking: A baseline and new evaluation metrics. In *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 10359–10366. IEEE, 2020.

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

[20] Yan Yan, Yuxing Mao, and Bo Li. Second: Sparsely embedded convolutional detection. *Sensors*, 18(10):3337, 2018.

[21] Fan Yang, Shigeyuki Odashima, Shoichi Masui, and Shan Jiang. Hard to track objects with irregular motions and similar appearances? make it easier by buffering the matching space. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 4799–4808, 2023.

[22] Tianwei Yin, Xingyi Zhou, and Philipp Krahenbuhl. Center-based 3d object detection and tracking. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11784–11793, 2021.

[23] Sisi You, Hantao Yao, Bing-Kun Bao, and Changsheng Xu. Utm: A unified multiple object tracking model with identity-aware feature enhancement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21876–21886, 2023.

[24] Haibao Yu, Yizhen Luo, Mao Shu, Yiyi Huo, Zebang Yang, Yifeng Shi, Zhenglong Guo, Hanyu Li, Xing Hu, Jirui Yuan, et al. Dair-v2x: A large-scale dataset for vehicle-infrastructure cooperative 3d object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21361–21370, 2022.

[25] Haibao Yu, Wenxian Yang, Hongzhi Ruan, Zhenwei Yang, Yingjuan Tang, Xu Gao, Xin Hao, Yifeng Shi, Yifeng Pan, Ning Sun, et al. V2x-seq: A large-scale sequential dataset for vehicle-infrastructure cooperative perception and forecasting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5486–5495, 2023.

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

[27] Yifu Zhang, Chunyu Wang, Xinggang Wang, Wenjun Zeng, and Wenyu Liu. Fairmot: On the fairness of detection and re-identification in multiple object tracking. *International journal of computer vision*, 129:3069–3087, 2021.

[28] Zhaohui Zheng, Ping Wang, Wei Liu, Jinze Li, Rongguang Ye, and Dongwei Ren. Distance-iou loss: Faster and better learning for bounding box regression. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 12993–13000, 2020.

[29] Yang Zhou, Cai Yang, Ping Wang, Chao Wang, Xinhong Wang, and Nguyen Ngoc Van. Vit-fusenet: Multimodal fusion of vision transformer for vehicle-infrastructure cooperative perception. *IEEE Access*, 2024.

[30] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable detr: Deformable transformers for end-to-end object detection. *arXiv preprint arXiv:2010.04159*, 2020.