MixedFusion: An Efficient Multimodal Data Fusion Framework for 3-D Object Detection and Tracking

Cheng Zhang, Hai Wang, *Senior Member, IEEE*, Long Chen, Yicheng Li, and Yingfeng Cai, *Senior Member, IEEE*

***Abstract*— The performance of environmental perception is critical for the safe driving of intelligent connected vehicles (ICVs). Currently, the most prevalent technical solutions are based on multimodal data fusion to achieve a comprehensive perception of the surrounding environment. However, existing fusion perception methods suffer from issues such as low sensor data utilization and unreasonable fusion strategies, which severely limit their performance in adverse weather conditions. To address these issues, this article proposes a novel multimodal data fusion framework called MixedFusion. In this framework, we introduce two innovative fusion strategies for the data characteristics of each sensor: high-level semantic guidance (HLSG) and multipriority matching (MPM). It not only realizes the efficient utilization of the multimodal data but also further realizes the complementary fusion between the multimodal data. We perform extensive experiments on the nuScenes and K-radar datasets. The experimental results demonstrate that the fusion framework proposed in this article significantly improves the performance of 3-D object detection and tracking in severe weather conditions.**

***Index Terms*— Environmental perception, intelligent connected vehicle (ICV), multimodal data fusion.**

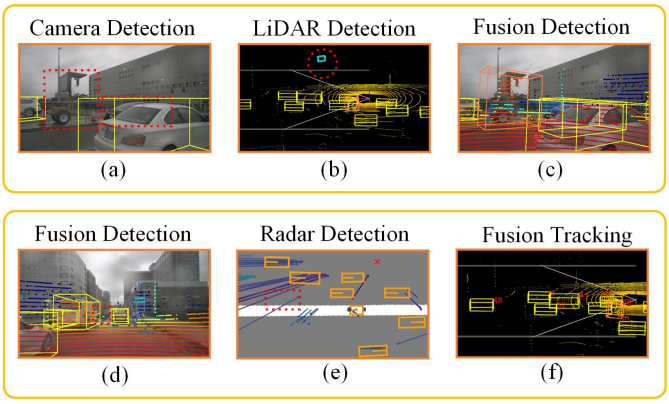

Fig. 1. Comparison of single-modal and multimodal methods. The dotted boxes and circles indicate the missed and false detections, respectively. The multimodal fusion-based methods can effectively overcome the shortcomings of single-modal methods to achieve accurate and robust detection and tracking. (a) Camera detection. (b) LiDAR detection. (c) and (d) Fusion detection. (e) Radar detection. (f) Fusion tracking.

I. Introduction

With the development of intelligent driving, the safety of intelligent connected vehicles (ICVs) has received increasing attention [1]. As the foundation of intelligent driving technology, environmental perception tasks play a crucial role in ensuring the driving safety of ICVs [2]. Currently, ICVs are usually equipped with a variety of sensors, including cameras, light detection and ranging (LiDAR), and radars. The perceptual system based on multimodal data fusion can effectively use complementary information from different types of sensors to achieve robust perception in adverse weather [3].

There are obvious advantages and disadvantages associated with each sensor type. Camera images contain rich colors, texture information, and dense semantic information, but the lack of depth information is caused by the principle of perspective imaging is inevitable [4]; the LiDAR point clouds can precisely capture the 3-D structural information of the environment, but their sparseness and disorder can lead to a lack of semantics [5]. The radar point clouds are hardly affected in foggy weather due to their strong penetration ability, but their resolution is low, and the false alarm is serious [6], [7]. In some working conditions, relying solely on single-modal information to infer will inevitably result in inaccurate perceptions. Therefore, a multimodal method that can complement camera images, LiDAR point clouds, and radar point clouds has a significant advantage [8], [9], [10].

The current fusion perception tasks are primarily divided into fusion detection and fusion tracking, which we refer to as multiobject detection (MOD) and multiobject tracking (MOT). In tracking by detection (TBD) architecture, tracking takes the results of detection as input. Therefore, MOT is usually a subsequent task of MOD. As shown in Fig. 1(a)–(c), the 3-D MOD methods based on camera images often resulted in missed detection due to the lack of depth information and the issue of obstruction. The 3-D MOD methods based on LiDAR point clouds will also result in significant false detection due to a lack of semantics; in contrast, the fusion detection method can use the complementary information of images and point clouds to achieve precise perception. As shown in Fig. 1(d)–(f), the fusion detection method failed to detect the distant object due to the distortion of images

Manuscript received 17 October 2022; revised 18 June 2023; accepted 15 October 2023. This work was supported in part by the National Natural Science Foundation of China under Grant 52225212, Grant U20A20333, and Grant 52072160; and in part by the Key Research and Development Program of Jiangsu Province under Grant BE2020083-3. *(Corresponding author: Hai Wang.)*

Cheng Zhang and Hai Wang are with the School of Automotive and Traffic Engineering, Jiangsu University, Zhenjiang 212013, China (e-mail: 2112004009@stmail.ujs.edu.cn; wanghai1019@163.com).

Long Chen, Yicheng Li, and Yingfeng Cai are with the Automotive Engineering Research Institute, Jiangsu University, Zhenjiang 212013, China (e-mail: chenlong@ujs.edu.cn; liyucheng070@163.com; caicaixiao0304@126.com).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TNNLS.2023.3325527.

Digital Object Identifier 10.1109/TNNLS.2023.3325527

and LiDAR point clouds caused by the rain. However, the missed object is detected by the detection method based on radar point clouds. Although its low resolution has led to a certain amount of missed detection, the fusion tracking method, which combines the two detection results, can achieve complementary advantages and accurate tracking.

A variety of fusion perception methods have been proposed by researchers to realize robust perception in adverse weather conditions. However, existing methods do not thoroughly explore the differences and complementarities of multimodal data. In articles such as FUTR3D [8] and RVF-Net [11], the three-modal fusion method of the camera image, LiDAR point clouds, and radar point clouds was proposed. Although these methods realize data fusion formally, the differences in heterogeneous data are not considered. LiDAR point clouds and camera images are typically more accurate, while radar suffers from significant clutter interference, resulting in low confidence for certain point clouds. Simply including radar point clouds in data fusion not only fails to improve the robustness of the algorithm but actually reduces the confidence of the fused data. For this reason, some scholars believe that radar itself comes with uncertainty that should be eliminated from multimodal data fusion as soon as possible. In many publications, including TransFusion [12], Fast-CLOCs [13], BEVFusion [14], FusionPainting [15], DeepFusion [16], EagerMOT [17], and mmMOT [18], radar was abandoned and only camera images and LiDAR point clouds were used to detect and track objects. This clearly falls into another extreme. The advantages of radar in rainy and foggy weather are incomparable to those of other sensors. Discarding radar is not conducive to taking advantage of the complementary fusion of multimodal data.

In addition, the current perception methods have the problems of inefficient utilization of multimodal data and unreasonable fusion strategies to some extent. In articles such as PointPainting [19], BiProDet [20], and VPFNet [21], fusion detection methods based on proposal level and point level have been proposed. These methods achieve data alignment of multimodal features through the joint calibration matrix of multiple sensors. Subsequently, data fusion is accomplished using projection-based techniques, essentially employing hard association fusion. However, hard association fusion methods are often limited by the sparsity of point clouds, leading to inefficient utilization of dense semantic information from images. In fusion tracking methods such as CBMOT [22], Be-MOT [23], and YONTD-MOT [24], fusion detection is abandoned to fulfill the requirement of observation independence. Only camera images and LiDAR point clouds are used for independent detection results. These methods require tracking the state of objects separately in the image pixel coordinate and the 3-D point cloud coordinate, failing to fully leverage the advantages of multimodal data fusion. Hence, such a simplistic fusion strategy is unreasonable.

To address the aforementioned issues, an efficient multimodal data fusion perception framework called MixedFusion is proposed in this article. Due to the coupled characteristics of the TBD architecture, our proposed framework consists of cascaded fusion detection and fusion tracking methods. Furthermore, according to the data characteristics of heterogeneous sensors, we hierarchically fuse the camera image, LiDAR point clouds, and radar point clouds in the detection and tracking algorithm. This approach not only avoids interference from low-confidence radar point clouds on camera images and LiDAR point clouds but also retains the advantages of radar in adverse weather conditions, fully exploiting the complementary fusion advantages of multimodal data.

In this framework, we also introduce two innovative fusion strategies, high-level semantic guidance (HLSG) and multipriority matching (MPM), to overcome the limitations of current fusion perception methods. By using high-level semantic data from images as a guide, we are able to achieve an adaptive soft-weighted association of camera image and LiDAR point cloud features in the shared bird's eye view (BEV) space. This not only resolves the issue of sparse point clouds causing low utilization of image semantic information, but it also makes it easier to combine dense semantics and spatial structural information in a complementary manner. In fusion tracking, through the cascaded matching of fusion detection results and radar detection results, we perform object tracking and state updating in a unified 3-D world coordinate. This approach effectively utilizes radar data, supplements the limitations of fusion detection, and enhances the robustness of the algorithm.

We conducted extensive tests on nuScenes [6] and K-radar [7] datasets. The experimental results show that our proposed method effectively improves the performance of fusion detection and tracking methods in adverse weather conditions.

The main contributions of this work are given as follows.

1) We propose an efficient multimodal data fusion perception framework, which realizes the hierarchical fusion of camera images, LiDAR point clouds, and radar point clouds and the combination optimization of detection and tracking algorithms.
2) In the framework, two novel fusion strategies, namely, HLSG and MPM, are introduced to improve the utilization of multimodal data and further realize the complementary fusion of the multimodal data.
3) Our proposed multimodal data fusion framework significantly improved the performance of the 3-D MOD and MOT methods, which effectively solved the problem of environment perception in adverse weather conditions.

II. Related Works

A. 3-D Object Detection Based on Multimodal Data Fusion

Currently, the predominant fusion detection methods can be divided into two categories: proposal level and point level. Among them, proposal-level works, such as VPFNet [21] and CenterFusion [25], are relatively nascent. The fusion strategy of these methods is to create 3-D proposals based on the LiDAR point clouds and then project the proposals onto other modal data for refinement. Essentially, this fusion strategy only processes data from single modalities. It does not effectively use multimodal complementary information. Moreover, it is difficult to take advantage of data fusion. In contrast, point-level methods, such as MVP [26], Focals-Conv [27], and AutoAlign [28], are more popular.

As shown in Fig. 2, these methods first project the point clouds into the image plane based on the joint calibration matrix of multiple sensors to obtain the corresponding pixel values. Then, the features of the point clouds and the corresponding image pixels are extracted. Finally, the obtained feature vectors are concatenated to achieve multimodal data fusion. This fusion strategy makes use of multimodal complementary information to a certain extent. Compared with the first strategy, its detection accuracy is improved. However,

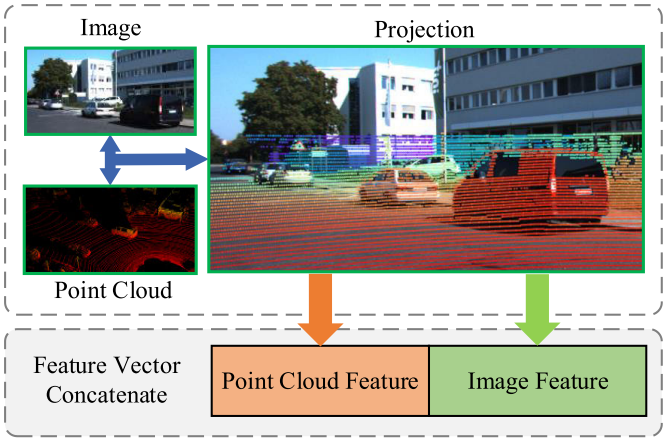

Fig. 2. Mainstream strategy of point-level fusion detection methods.

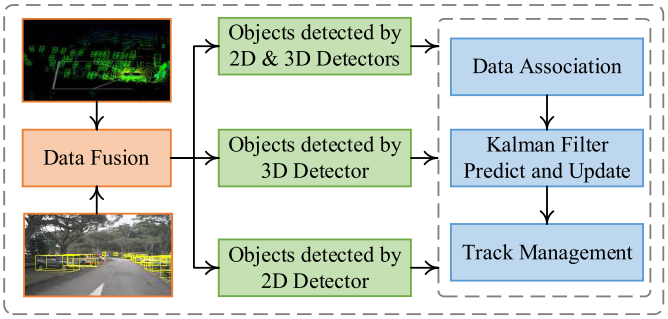

Fig. 3. Mainstream strategy of fusion tracking methods.

this method may result in low utilization of image semantic information [29], [30]. For example, the point clouds are sparse when the algorithm detects distant objects. In this case, the image features obtained based on the projection matrix are forced to become sparse. As a result, most of the image features are discarded. The image feature itself contains dense semantic information, and it is difficult to retain the semantic information after being forced to be sparse. This makes it difficult for the network to understand the information contained in the image features to some extent. Therefore, we propose a fusion strategy based on HLSG to effectively solve the problem of low utilization of image semantic information in remote detection. Moreover, the strategy can maximize the retention of image features and realize the complementary fusion of images and point clouds.

B. 3-D Object Tracking Based on Multimodal Data Fusion

With the rapid development of MOD based on deep learning, 3-D MOT methods using the TBD framework have gradually become mainstream in recent years [31], [32], [33], [34], [35]. To satisfy the assumption of observation independence, fusion tracking algorithms, such as DeepFusion [16], EagerMOT [17], and JRMOT [36], use 2-D MOD results from camera images and 3-D MOD results from LiDAR point clouds for fusion tracking.

As shown in Fig. 3, the core concept of the fusion strategy is cascaded data association. These methods first match the intersection over union (IoU) of the detection object list from different modalities. Then, they associate the data based on the priority of the detection results. After that, the Kalman filter

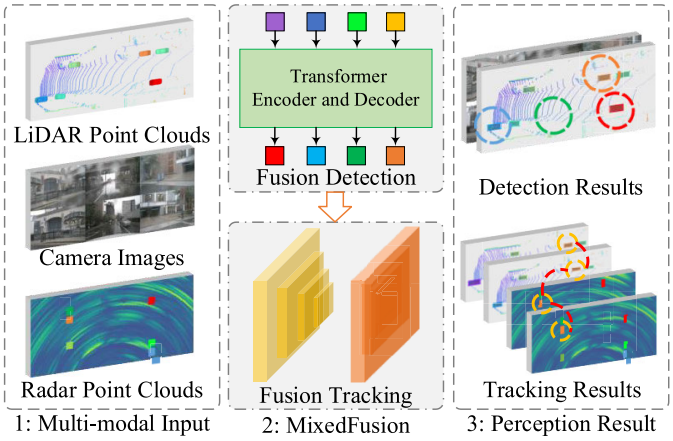

Fig. 4. Overall architecture of MixedFusion.

(KF) is used to predict and update the status, and finally, the lifecycle management of the track is performed. This method only realizes the fusion in form, and there are fatal flaws. The accuracy of 2-D MOD based on camera images is typically much higher than that of 3-D MOD based on LiDAR point clouds due to the sparsity of remote LiDAR point clouds. This results in the fact that the remote objects detected in the camera image are hardly detected by the LiDAR point clouds. In the tracking algorithm, these objects can only be predicted and updated in 2-D pixel coordinates. Since the camera images cannot provide accurate depth information for remote objects, it is difficult to obtain the corresponding 3-D world coordinates via back projection. Even if tracking is performed, this part of the tracking results has no practical significance. Until the LiDAR point clouds detect these objects, the 3-D world coordinates can be updated. The fusion tracking algorithm is almost completely limited by the detection results of the LiDAR point clouds. However, the advantages of the fusion algorithm are not reflected. Therefore, we introduce radar detection results into fusion tracking to track all detection results simultaneously in 3-D world coordinates. On this basis, the complementary fusion of multimodal data is realized through the MPM strategy, which enhances the adaptability of the algorithm in adverse weather conditions.

III. Proposed Method

The overall architecture of our proposed MixedFusion is shown in Fig. 4, which is composed of a fusion detection algorithm and a fusion tracking algorithm. The algorithm realizes the hierarchical fusion of LiDAR point clouds, multiview camera images, and radar point clouds, and outputs MOD and MOT results.

A. Fusion Detection Method Based on HLSG

1) Overall Architecture: Camera and LiDAR images and LiDAR point clouds are typically reliable because of their high resolution and accuracy. However, radar has low resolution, and the problems of clutter interference and false alarms are serious. Therefore, the confidence of radar point clouds is relatively low. To avoid the interference of radar on camera and LiDAR, we only use LiDAR point clouds and multiview camera images in the fusion detection. Referring to DETR [37], the fusion detection framework has been optimized with the bipartite matching training strategy.

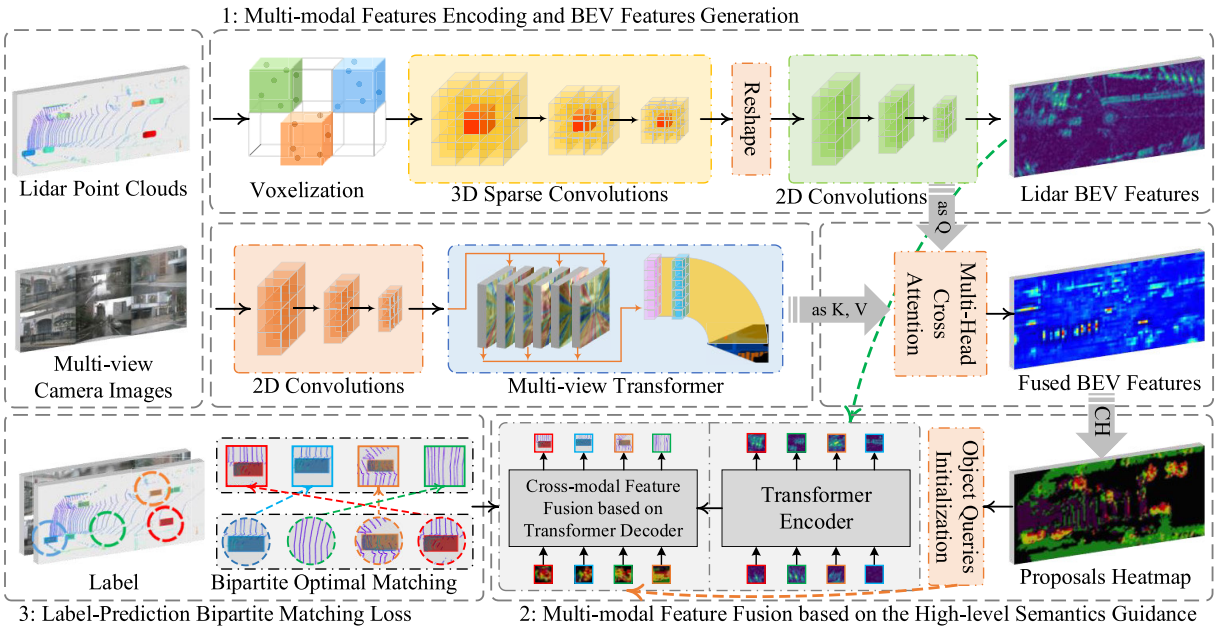

Fig. 5. Overall architecture of the fusion detection algorithm.

The overall architecture of the algorithm is shown in Fig. 5, which is mainly composed of multimodal feature fusion based on HLSG and label-prediction bipartite matching loss (LPBM-Loss).

2) Multimodal Features Encoding and BEV Features Generation: The LiDAR point clouds describe the environment from 3-D world coordinates, whereas the camera images describe the environment from the 2-D foreground view. To eliminate the difference in spatial semantic information and unify the expression of multimodal features, the point clouds and images are processed in two parallel channels. Finally, the multimodal features are fused on the shared BEV.

Compared to the point-based method, the voxel-based method can generally account for the accuracy and efficiency of algorithms at the same time [38]. In the point cloud channel, the first step is to voxelize the point clouds. Then, 3-D sparse convolutions are used for feature encoding. Next, the obtained voxel features are compressed in the height direction. Finally, 2-D convolution is used again for feature encoding to obtain the point clouds BEV features. The above process can be simply described as follows:

$$F_{\text{lid}} = \text{Voxel(PC)} \circ \text{SPConv} \circ \text{Reshape} \circ \text{Conv} \tag{1}$$

where $\circ$ denotes the cascade operation and PC denotes the input LiDAR point clouds. Voxel, SPConv, Reshape, and Conv represent voxelization, 3-D sparse convolution, compress in the height direction, and 2-D convolution, respectively.

In the image channel, 2-D convolutions are first used for feature encoding, and then, the obtained feature maps are sent to the multiview transformer (MVT) module. Finally, the high-level semantic features of the image BEV are generated by MVT. The above process can be simply described as

$$F_{\text{cam}} = \text{ResNet}\left(\text{Img}^0, \text{Img}^1, \ldots, \text{Img}^5\right) \circ \text{MVTrans} \tag{2}$$

where ResNet denotes the ResNet-50 backbone, Img^0, Img^1, and Img^5 represent multiview camera images from 0 to 5, and MVTrans denotes MVT. Next, the point clouds BEV is interacted with the image BEVs using a cross-modal attention mechanism to generate the final fused BEV features.

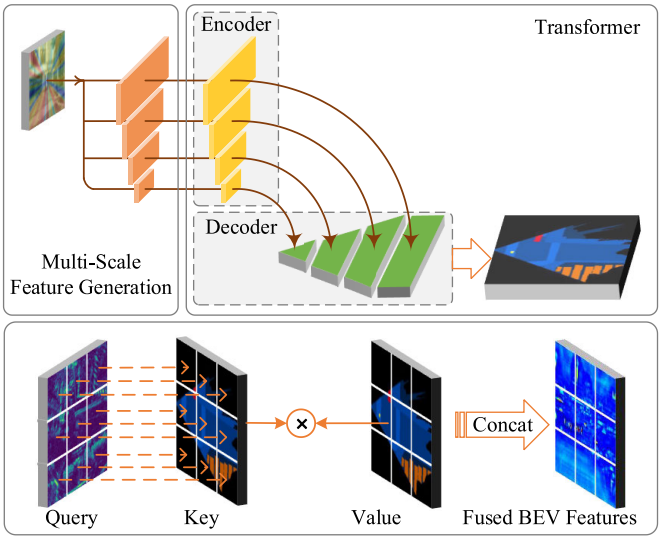

Fig. 6. Process of image BEV and fused BEV generation.

The detailed process of image and fusion BEV generation is shown in Fig. 6. According to the principle of camera perspective imaging, the image is a cone model. This causes the scenes in the camera image to exhibit near-large and far-small characteristics. The close-range objects have dense pixels and rich depth information. However, the pixels of remote objects are relatively rare, and the depth information is seriously missing. It is not easy to construct BEV from camera images, particularly for remote objects. To construct a semantic BEV with relatively accurate depth information, the multiscale image features retained low-dimensional fine information and high-dimensional semantic information are generated by ResNet50. Inspired by translating images into

maps [39], we designed the MVT structure, which can simultaneously process multiview image features. By inputting the multiscale foreground view features to the encoder, the multiview BEV features can be obtained from the decoder. After the multiscale BEV features are concatenated, the semantic BEV is available. Notably, we only present the BEV generation process from one camera's perspective in Fig. 6. In actual application, six multiscale encoders and decoders are set up in parallel in MVT, which can simultaneously output the semantic BEV of six cameras' perspectives. However, since the unknown sharing and complementarity between multiview BEVs, it is difficult to combine them into a complete BEV. Thus, we further designed the multihead cross-modal attention (MHCA) module to address the problems. Although a complete image BEV is hard to build, the point cloud BEV is naturally complete. Thus, the point cloud BEV can be seen as a BEV mask to combine the image BEVs. In particular, for each feature on point clouds BEV, the corresponding image BEV feature is found according to the spatial position. Then, all the acquired image features are combined to obtain the complete BEV.

Following the above idea, the point clouds BEV is used as a query, while the semantic BEVs of the six perspectives are used as the key and value in MHCA. In a specific implementation, six independent MHCAs are set up to interact the point clouds BEV with the six semantic BEVs in parallel. It is worth noting that the spatial structure information of the point clouds BEV is accurate and comprehensive. During the interaction, image semantic information can be directed to precisely match the corresponding point clouds. After that, the multiview image BEV features can be combined into complete image BEV and finally concatenated with the LiDAR BEV to obtain the fused BEV.

The above process can be expressed as

$$Q_{\text{lid}} = W^q F_{\text{lid}}, \quad K^i_{\text{cam}} = W^{k(i)}_{\text{cam}} F^i_{\text{cam}}, \quad V^i_{\text{cam}} = W^{v(i)}_{\text{cam}} F^i_{\text{cam}} \tag{3}$$

$$F^i_{\text{cam}} = \text{softmax}\left(\frac{Q_{\text{lid}}\left(K^i_{\text{cam}}\right)^T}{\sqrt{d}} + B\right) V^i_{\text{cam}} \tag{4}$$

$$F^{\text{BEV}}_{\text{cam}} = \text{Combine}\left(F^0_{\text{cam}}, F^1_{\text{cam}}, \ldots, F^5_{\text{cam}}\right) \tag{5}$$

$$F^{\text{BEV}}_{\text{fused}} = \text{Concat}\left(F^{\text{BEV}}_{\text{cam}}, F^{\text{BEV}}_{\text{lid}}\right) \tag{6}$$

where Q_{lid}, K^i_{cam}, and V^i_{cam} represent the LiDAR object query, camera BEV key, and camera BEV value, respectively. W^q, $W^{k(i)}_{\text{cam}}$, and $W^{v(i)}_{\text{cam}}$ represent the weight matrix of Q_{lid}, K^i_{cam}, and V^i_{cam}, respectively. F_{lid} and F^i_{cam} represent the LiDAR BEV features and the ith camera image features, respectively. $F^{\text{BEV}}_{\text{cam}}$ and $F^{\text{BEV}}_{\text{lid}}$ represent the image BEV and LiDAR BEV, respectively. B is the bias, and $F^{\text{BEV}}_{\text{fused}}$ denotes the fused BEV features.

3) Multimodal Feature Fusion Based on HLSG: In DETR [37], the initialization of the object query is completely random. In subsequent research, it was discovered that better initialization of object queries effectively reduces the difficulty of optimization and increases learning efficiency. Referring to CenterPoint [31], the CenterHead module is utilized to convert the fused BEV features into the heatmap of proposals. Then, the high response zones in the heatmap are selected as the proposals. The proposals contain abundant prior information, such as the spatial position and the high-level semantics of objects. We believe that using proposals as the initialization of object query may be a good choice because accurate spatial information and high-level semantics are very important for 3-D MOD.

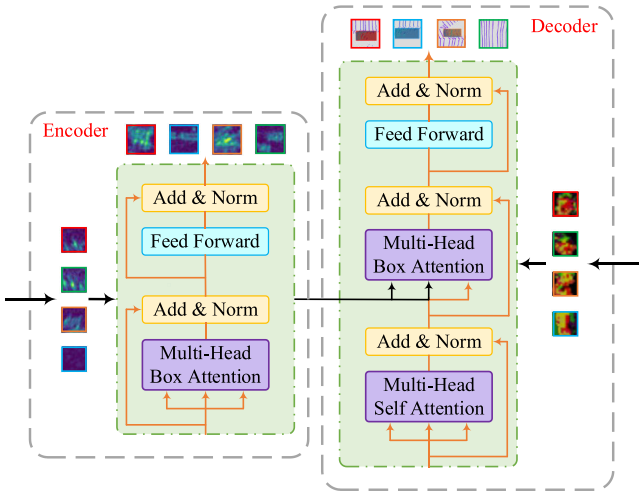

Fig. 7. Transformer encoder and decoder based on box-attention.

It is worth noting that the fused BEV features are only applied to initialize the object queries and are not further used as input to the Transformer encoder. Instead, the LiDAR point clouds BEV features are used as the encoder input. Although encoding the fused BEV features can provide more comprehensive information for subsequent decoder optimization, it introduces tremendous redundancy and computation. The point cloud BEV features, which have rich spatial structure information, are adequate for the needs of 3-D MOD. More than that, it is sufficient that the object queries contain high-level semantics that are only required to indicate the position of the potential objects.

In addition, the Transformer can introduce square-level computational complexity. However, both real-time and accuracy are essential to 3-D MOD. As shown in Fig. 7, the Transformer encoder and decoder based on box-attention [40] are designed to simultaneously resolve these two important problems. The main idea of box-attention is focusing on the boxes by predicting their transformation from reference windows. Compared with the original attention mechanism, box-attention merely concerns on the keys around the queries instead of all the keys. This approach is similar to convolutional neural networks (CNNs), which can decrease the computational complexity from square level to linear. In the encoder, the multihead self-box-attention (MHSBA) is designed to encode the point clouds BEV features in the form of reference boxes. The specific process is given as follows:

$$\begin{aligned} &\text{MHSBA}(Q_{\text{lid}}, K_{\text{lid}}, V_{\text{lid}}) \\ &\quad = \text{Concat}(h_1, \ldots, h_l) W^O \end{aligned} \tag{7}$$

$$\begin{aligned} h_i &= \text{BoxAttention}\left(Q_{\text{lid}}, K^i_{\text{lid}}, V^i_{\text{lid}}\right) \\ &= \sum_{m \times m} \text{softmax}\left(Q_{\text{lid}}\left(K^i_{\text{lid}}\right)^T\right) * V^i_{\text{lid}} \end{aligned} \tag{8}$$

where Q_{lid}, K_{lid}, and V_{lid} represent the shared object query, key, and value from LiDAR point clouds, respectively. h_i denotes the ith head of MHSBA. W^O is a learnable projection matrix. K^i_{lid} and V^i_{lid} represent ith head's component of the key and value, and m denotes the number of grids. The encoded

point clouds BEV features are then used as keys and values in subsequent decoders.

The multimodal feature fusion based on HLSG is realized in the decoding process. As stated above, the object queries contain rich high-level semantics and spatial prior information, thereby eliminating the requirement to set multiple decoders for cascaded optimization. In this article, a single-layer decoder is employed to predict the final 3-D bounding box (Bbox) parameters and category confidence. The process of the interaction between the object queries follows the conventional Transformer. However, the multihead cross-modal box-attention (MHCMBA) is designed to interact the object queries with the keys and values in the core decoding process. The specific process is given as follows:

$$\begin{aligned}&\mathrm{MHCMBA}(Q_{\mathrm{fused}}, K_{\mathrm{lid}}, V_{\mathrm{lid}})\\&\quad=\mathrm{Concat}(h_1,\ldots,h_l)W^O \qquad (9)\\&h_i=\mathrm{BoxAttention}\left(Q_{\mathrm{fused}}, K^i_{\mathrm{lid}}, V^i_{\mathrm{lid}}\right)\\&\quad=\sum_{m\times m}\mathrm{softmax}\left(Q_{\mathrm{fused}}\left(K^i_{\mathrm{lid}}\right)^T\right)*V^i_{\mathrm{lid}} \qquad (10)\end{aligned}$$

where Q_{fused} represents the shared object query from the fused BEV features. The subsequent process of enhancing features is the same as the encoder.

In conclusion, the multimodal feature fusion based on MHCMBA is essentially a global matching and adaptive filtering of the fused BEV features containing high-level semantics to the point cloud BEV features. Whether the point clouds are sparse or not, the image features can be fully utilized, thereby fundamentally avoiding the limitations of point-based fusion methods.

B. Fusion Tracking Method Based on MPM

1) Overall Architecture: Given the uncertainty of the motion model and the observation model, fusion tracking typically employs KF for optimal state estimation to predict and update tracking trajectories. However, the premise of applying KF is the assumption of observational independence. This means that the data used for the fusion must be independent of each other. Therefore, almost all fusion tracking methods combine the 2-D MOD results based on images and the 3-D MOD results based on LiDAR point clouds, giving up the high-accuracy results of fusion detection. The original intention of introducing radar into fusion tracking is to take advantage of its benefits in adverse weather conditions. However, if the fusion detection results are to be utilized, introducing radar is also necessary to satisfy the assumption of observation independence. Since the focus of this article is on the fusion framework, the existing RADDet [41] is chosen as the radar MOD algorithm.

The overall architecture of the algorithm is shown in Fig. 8, which is mainly composed of the appearance features and Bbox features generation module and the MPM module.

2) Appearance Features and Bbox Features Generation: Different from the traditional MOT in the mathematics field, the MOT based on TBD architecture typically simplifies typically simplify data association to the problem of optimal bipartite matching between the detected objects and the existing tracks. Generally, the Hungarian and greedy matching algorithms are employed to associate the Bbox parameters of the detected objects with the existing tracks. Whether the association is successful is judged by a constant threshold value. However, this method oversimplifies the data association problem and is far from the matching process of the human brain. Obviously, human beings can make comprehensive judgments by considering the objects' appearance, features, and spatial positions. Therefore, we simply extended FANTrack [42] and applied it to the fusion tracking.

In fusion detection method, the Transformer encoder and decoder are used to fuse the features of LiDAR point clouds and camera images. Here, the fused BEV features $F^{\mathrm{BEV}}_{\mathrm{fused}}$ generated by Transformer are used as the appearance features of the detected objects. In the radar pipeline, the appearance features of objects are generated by concatenating the multiscale features of radar point clouds. Since radar has low resolution in the height direction, the prediction of height information is not accurate. Thus, the height is ignored and x, y, z, l, w, and θ are only used as the Bbox parameters to generate the spatial features. The above process can be expressed as

$$F_{\mathrm{App}}=\mathrm{Concat}\left(F^{\mathrm{BEV}}_{\mathrm{fused}}, F^{\mathrm{BEV}}_{\mathrm{rad}}\right) \qquad (11)$$

$$F_{\mathrm{Bbox}}=\mathrm{FC}\circ\mathrm{Conv}(\mathrm{Bbox}_0,\mathrm{Bbox}_1,\ldots,\mathrm{Bbox}_i) \qquad (12)$$

where F_{App} and F_{Bbox} denote the appearance features and 3-D Bbox features of the detected objects, respectively, $F^{\mathrm{BEV}}_{\mathrm{rad}}$ denotes the multiscale radar BEV features, and Bbox_i represents the parameters of the ith Bbox.

Next, the Siamese network (SN) consisting of similarity estimate network (SEN) and data association network (DAN) is set up to associate the trackers and the detected objects. However, the appearance features and spatial features are not always equally important for data association. Sometimes people can track object only according to the color of objects or spatial features are enough if the running speed of the objects is slow. The mainstream methods can hardly describe this situation except using the constant threshold to weigh the importance of the appearance features and spatial features. Here, two weight coefficients w_{App} and w_{Bbox} are introduced to adaptively adjust the importance of these features. The above process can be expressed as

$$\mathrm{SimMap}=\mathrm{SEN}\left(w_{\mathrm{App}}F_{\mathrm{App}}+w_{\mathrm{Bbox}}F_{\mathrm{Bbox}}\right) \qquad (13)$$

$$\mathrm{AssoMap}=\mathrm{DAN}(\mathrm{SimMap}) \qquad (14)$$

where SimMap and AssoMap represent the similarity map and association map, respectively. It is worth noting that there is no clear threshold in our associate method. w_{App} and w_{Bbox} are generated by a multilayer perception (MLP) network. The inputs of the MLP are F_{App} and F_{Bbox}. The importance weights w_{App} and w_{Bbox} are normalized to sum up to unity. Thereby, the network can make a comprehensive judgment by adaptive learning, which has stronger generalization capability.

3) Data Association Based on MPM: The above content is the detailed process of the certain level of the proposed data association. However, the core of fusion tracking is to prioritize detection results from different sources based on confidence and then associate the data according to the priority. Different from the mainstream methods, the detection results introduced in fusion tracking are all from the 3-D coordinates of the fusion detection and radar detection algorithms, which can be directly matched in the world coordinate system. In particular, 3-D generalized IoU (3-DGIoU) is applied to evaluate the coincidence degree of the fusion detection results

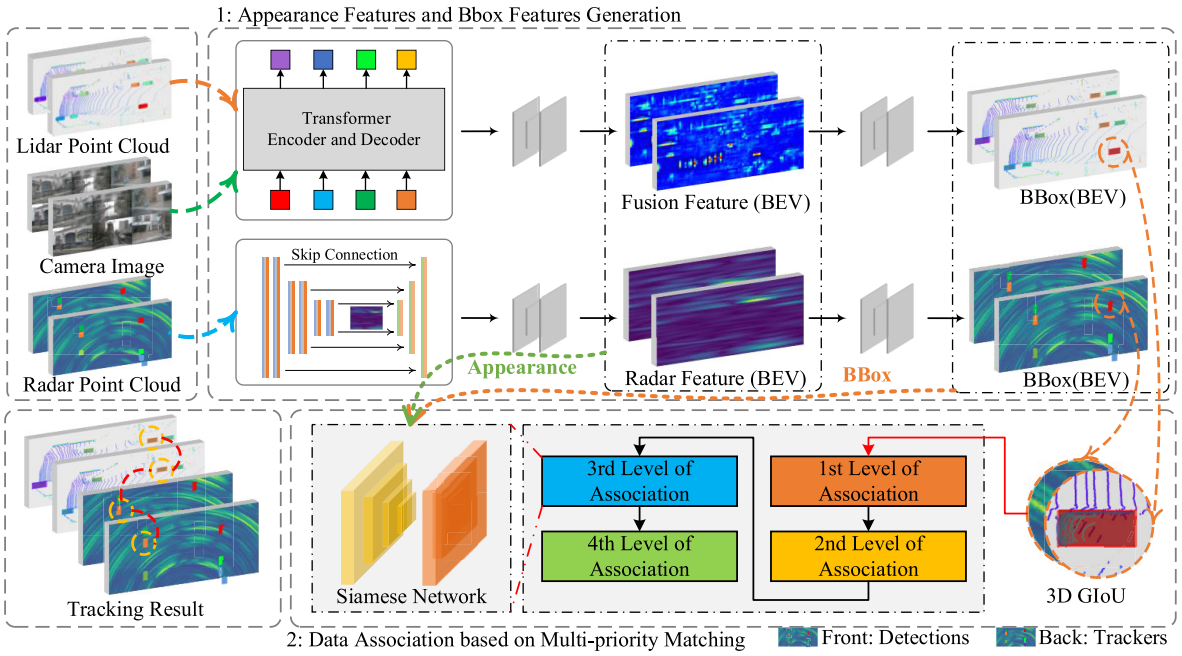

Fig. 8. Overall architecture of the fusion tracking algorithm.

and the radar detection results, which is expressed as

$$S_U = S_{B_1} + S_{B_2} - S_I \tag{15}$$

$$3\text{DGIoU}(B_1, B_2) = S_I/S_U - (S_C - S_U)/S_C \tag{16}$$

where S_U and S_I represent the area of union and intersection, respectively, B_1 and B_2 represent the detection result list of fusion detection and radar, respectively, and S_C is the area of the enclosing convex hull of U.

Several studies have demonstrated that high-accuracy matching can be achieved using only 3DIoU. However, the experiment revealed that the detection results from the radar point clouds contain an excessive amount of noise, resulting in a greater number of false matching results. Therefore, 3DGIoU is chosen as the cost function for the greedy algorithm, which is relatively complex but more robust.

For the current mainstream fusion tracking methods, detection results are typically divided into two categories: those detected by LiDAR point clouds (with 3-D Bbox parameters) and those detected by the camera image alone (only with 2-D Bbox parameters). Consequently, the tracking algorithm contains two trackers to maintain the respective tracks of 3-D and 2-D objects. In this article, the matching results are divided more precisely into three categories: 1) D_{Fusion}—the objects detected simultaneously by fusion detection and the radar algorithm; 2) D_{LC}—the objects detected only by the fusion detection algorithm; and 3) D_{Ra}—the objects detected only by radar. As shown in Table I, the four-level data association between the detection results and the existing track T is performed based on priority.

In the first-level association, only the association between D_{Fusion} and T is considered. We believe that objects detected by both fusion detection and radar algorithms have the highest confidence. Associating D_{Fusion} with T first can reduce the impact of low-quality objects on the subsequent data association. Next, the track $T^{\text{1st}}_{\text{Matched}}$ that has been successfully associated is reserved, and the track $T^{\text{1st}}_{\text{Unmatched}}$ that has failed to be associated will be associated at the next level. Since the fusion detection results are more accurate than the radar detection results, the Bbox parameters of the fusion detection results are preferred to update the state of $T^{\text{1st}}_{\text{Matched}}$ in the KF algorithm.

In the second-level association, only the association between D_{LC} and $T^{\text{1st}}_{\text{Unmatched}}$ is considered. The fusion detection results have higher confidence than the radar detection results. Thus, it is set as the second priority. The subsequent state update process is the same as the first level and is not described again.

In the third-level association, only the association between D_{Ra} and $T^{\text{2nd}}_{\text{Unmatched}}$ is considered. The confidence of D_{Ra} is relatively low. However, it can usually detect remote objects that cannot be detected by the fusion detection algorithm, which is conducive to the initialization of the track at a very far distance. On rainy and foggy days, it can also supplement the objects that the fusion detection algorithm cannot detect. At this time, only the Bbox parameters of D_{Ra} are known, and their accuracy is relatively low. However, the objects are relatively far in most cases, which typically has a minor impact on intelligent driving. Once the distance is close, the fusion detection algorithm will detect it. At this time, it will skip to the first- and second-level associations and update the states utilizing the high-accuracy Bbox parameters of the fusion detection.

In the fourth level of association, the final association between the detection objects of all the remaining uncompleted associations $D^{\text{Left(1st,2nd,3rd)}}_{\text{Unmatched}}$ and $T^{\text{3rd}}_{\text{Unmatched}}$ is tried to perform. Although most of the high-quality detection results have been matched in the previous three-level association, we believe that the remaining $D^{\text{Left(1st,2nd,3rd)}}_{\text{Unmatched}}$ still has effective objects of the incomplete association. Therefore, the association objects with relatively high similarity but has failed to be associated in DN are kept and output the final $D^{\text{4th}}_{\text{Unmatched}}$ and $T^{\text{4th}}_{\text{Unmatched}}$.

IV. Experiments

We performed extensive experiments on nuScenes and K-radar datasets. The detailed ablation experiments were

TABLE I
DETAILED PROCESS OF MPM

Priority	Detections	Trackers	Matched (Updated)	Unmatched
1st Association	D_{Fusion}	T_{Inital}	$T_{Matched}^{1\text{st}} \leftarrow D_{Matched}^{1\text{st}}$	$T_{Unmatched}^{1\text{st}}, D_{Unmatched}^{1\text{st}}$
2nd Association	D_{LC}	$T_{Unmatched}^{1\text{st}}$	$T_{Matched}^{2\text{nd}} \leftarrow D_{Matched}^{2\text{nd}}$	$T_{Unmatched}^{2\text{nd}}, D_{Unmatched}^{2\text{nd}}$
3rd Association	D_{Ra}	$T_{Unmatched}^{2\text{nd}}$	$T_{Matched}^{3\text{rd}} \leftarrow D_{Matched}^{3\text{rd}}$	$T_{Unmatched}^{3\text{rd}}, D_{Unmatched}^{3\text{rd}}$
4th Association	$D_{Unmatched}^{Left(1\text{st},2\text{nd},3\text{rd})}$	$T_{Unmatched}^{3\text{rd}}$	$T_{Matched}^{4\text{th}} \leftarrow D_{Matched}^{4\text{th}}$	$T_{Unmatched}^{4th}, D_{Unmatched}^{4th}$

further performed on each architecture component to verify the fusion framework's effectiveness.

A. Dataset

Currently, most of the datasets for intelligent driving only contain camera images and LiDAR point cloud data. For this reason, nuScenes and K-radar datasets containing camera images, LiDAR point clouds, and radar point clouds are chosen for experimental verification.

NuScenes is a large-scale autonomous driving dataset collected in Boston and Singapore with complex traffic environment. The complete dataset contains 1000 driving scenes in different locations, weather conditions, vehicle types, vegetation, and road markings. Therefore, it has become one of the most recognized datasets in the field [6].

In addition, K-radar [7] is a novel large-scale dataset and benchmark that contains 35k frames of 4-D radar, camera, and LiDAR data. More importantly, it includes challenging driving conditions such as adverse weather (fog, rain, and snow) on various road structures (urban and suburban roads, alleyways, and highways).

B. Implementation Details

1) Fusion Detection:

a) Data augmenting: For the nuScenes dataset, the point clouds are sparse because 32-line LiDAR is used to collect data. After space-to-time conversion, the keyframe point clouds were superimposed with the nonkeyframe point clouds to generate the dense LiDAR point clouds. Following conventions [32], we set the X and Y detection ranges to $[-54, 54]$ m and the Z detection range to $[-5, 3]$ m. The voxel size of point clouds is set to [0.075, 0.075, 0.2] m. In the pretraining process of the point clouds algorithm, we used random flipping along the X- and Y-axes, global scaling with a scaling coefficient of [0.95, 1.05], random global rotation with a range of $[-\pi/8, \pi/8]$, and ground truth sample (GT-Sample).

For the K-radar dataset, 64-line LiDAR is provided, and the LiDAR point clouds are dense enough. Following the parameter settings of the baseline provided by the official platform, we set the X detection range to [0, 72] m, the Y detection range $[-6.4, 6.4]$ m, and the Z detection range to $[-2, 6]$ m. In addition, the other parameters of data augmenting were set the same as the nuScenes dataset.

b) Training process: For easy comparison, the parameters of the training process for the two datasets are set to the same in most cases. The pretraining model of the image pipeline is derived from the ResNet50 pretraining weight. Also, the Adam optimizer is used to train the network. The maximum learning rate is 0.001, the weight decay is 0.01, and the momentum is 0.85.

We used 8× Tesla V100 GPUs for training, and the batch size is set to 16. First, we conduct 40 epochs of pretraining for the point cloud pipeline. Then, we freeze the pretraining weights of the image pipeline and conduct ten epochs of fine-tuning training. GT-Sample is not used in the last ten training epochs since the network's generalization ability may be influenced. Since the maximum number of objects in the two datasets' scenes is 142, the number of object queries is set to 180 to ensure the balance of positive and negative samples. For box-attention, the parameter settings follow the BoxeR-3D [40]. Three reference windows of 16 features on BEV with three angles $\{-2\pi/3, 0, 2\pi/3\}$ are set for one query. Each attention head is assigned a reference window of one angle. For LPBM-Loss, the optimal matching between the label and the prediction can be achieved when α, β, and γ are 0.2, 0.4, and 0.4, respectively.

2) Fusion Tracking:

a) Training process: Similar to fusion detection, the parameters for the two datasets are set to the same. The training batch size of the SEN is set to 64. The input data consists of all single-frame object detection list parameters, including object number, category, Bbox parameters, appearance features, and label information. Furthermore, the Adam optimizer and exponential decay learning rate are utilized. The maximum learning rate was set to $1e^{-5}$ and then decreased every 200 epochs, with a base of 0.95. Similarly, the DAN also uses the Adam optimizer and exponential-decay learning rate. The maximum learning rate was set to $1e^{-6}$ and then decreased by 0.95 per 20 epochs.

b) Threshold setting: Since the detection results of the radar contain noise, the confidence threshold is set to 0.2. Conversely, the confidence of the fusion detection results is very high, and the confidence threshold is set to 0. The 3DGIoU threshold for matching fusion and radar detection results has been set to -0.5. According to the parameter settings of AB3DMOT [32], the maximum life cycle for all categories is set at 3. In addition, the minimum number of matches is determined based on the data association priority. The minimum data hits in the first-, second-, third-, and fourth-level associations are set to 1, 2, 4, and 6, respectively.

C. Dataset Experiments

We conducted a large number of comparative experiments with other state-of-the-art methods on nuScenes and K-radar datasets to demonstrate the superiority of the proposed method. Compared with the radar dataset, the nuScenes dataset does not contain many kinds of severe weather scenes. For this

TABLE II
PERFORMANCE ON NUSCENES VALIDATION SET BENCHMARK (ONLY RAIN AND NIGHT SCENES)

Methods	Modality	mAP↑	NDS↑	Latency(ms)
BEVFormer [43]	C	36.6	42.0	-
PointPillars [44]	L	48.1	57.3	34.4
SECOND [45]	L	49.3	59.9	69.8
CenterPoint [31]	L	56.7	62.6	80.7
FUTR3D [8]	C+L	61.5	65.3	321.4
PointPainting [19]	C+L	62.8	67.9	185.8
MVP [26]	C+L	63.2	68.4	187.1
FusionPainting [15]	C+L	63.6	68.7	-
AutoAlign [28]	C+L	64.7	68.9	-
TransFusion [12]	C+L	65.4	69.3	312.5
BEVFusion [14]	C+L	65.7	69.8	238.1
CMT [46]	C+L	65.9	70.1	166.7
Ours	**C+L**	**66.3**	**70.5**	**143.4**

TABLE III
PERFORMANCE ON K-RADAR TEST SET BENCHMARK

Methods	Modality	Normal	Overcast	Fog	Rain	Sleet	Light Snow	Heavy Snow	Total
RTNH [7]	R	44.40	55.60	46.90	34.00	20.20	41.60	40.20	41.1
Part-A2 [47]	L	57.82	41.87	44.50	49.37	39.26	49.48	36.21	42.94
PV-RCNN [48]	L	78.19	66.74	44.54	44.32	45.18	47.34	37.10	44.57
Voxel R-CNN [38]	L	81.82	69.59	48.81	47.08	46.88	54.84	37.17	46.37
CasA [49]	L	82.23	65.58	44.39	53.68	44.90	62.72	36.90	50.87
TED-S [50]	L	74.29	68.79	45.66	53.60	44.81	63.37	36.70	51.01
Fast-CLOCs [13]	C+L	52.32	46.07	44.43	47.07	44.61	46.92	34.78	42.97
3D Dual-Fusion [51]	C+L	74.77	72.33	43.91	52.50	44.06	58.59	36.30	47.78
BiProDet [20]	C+L	78.60	68.80	44.71	53.77	44.99	62.60	37.01	51.25
Focals Conv [27]	C+L	80.49	73.64	46.21	53.81	45.12	62.75	37.02	52.00
VPFNet [21]	C+L	81.12	76.25	46.32	53.66	44.92	63.07	36.93	52.17
TED-M [50]	C+L	77.16	69.69	47.40	54.29	45.20	64.29	36.78	52.31
Ours	**C+L**	**84.49**	**76.59**	**53.31**	**55.32**	**49.60**	**68.65**	**44.86**	**55.08**

reason, only rain and night scenes in the nuScenes dataset are chosen for experiments.

1) Fusion Detection: For the nuScenes dataset, mean average precision (mAP) and nuScenes detection score (NDS) are used to evaluate the detection accuracy of the algorithm. For the K-radar dataset, the mAP of different weather conditions and the mAP of all weather conditions are used to evaluate the detection accuracy.

The comparative experimental results of fusion detection in nuScenes and K-radar datasets are shown in Tables II and III, respectively. On the nuScenes dataset, the proposed method only showed a marginal improvement compared to the state-of-the-art approach CMT [46]. The mAP and NDS are both improved by 0.4%, and the latency of the algorithm is improved by 23.3 ms. This can be attributed to two primary reasons. First, the adverse weather conditions in nuScenes dataset contain only rain and night scenes, and most of the scenes are structured road scenes. This results in a lower difficulty for detection. Second, the forward and backward detection distances in the nuScenes dataset are both limited to 53 m, with most objects being in the medium to close range. This limitation prevents our method from leveraging its advantages in long-range detection.

Consequently, the performance of almost all algorithms is quite similar, and our advantages are not prominent. Moreover, the proportion of distant objects is lower than that of nearby objects. Nearly, all methods perform well with nearby objects. Therefore, the weak improvement can explain the algorithm's advantages in the face of detection at long-range distances to a large extent.

On the radar dataset, the proposed method demonstrates significant improvements over state-of-the-art approach TED-M [50] across all weather conditions, particularly in extreme weathers such as overcast, fog, and heavy snow. The mAP for all scenes increased by 2.77%, effectively showcasing the robustness of our method in handling various adverse weather conditions. It is worth noting that the radar dataset has a forward detection range of 72 m, which nearly reaches the limit of LiDAR detection algorithms. At this range, the LiDAR point clouds are sparse, making it difficult for LiDAR-based algorithms to make reliable inferences. However, the image semantics utilization of the proposed method is much higher than that of the mainstream fusion algorithms, which can make an effective complement to the sparse point clouds. Therefore, our method exhibits significant advantages in detecting targets at medium-to -long-range distances.

To intuitively show the performance of the proposed method, the rain (left) and the night scene (right) are randomly selected for 3-D Bbox visualization, as shown in Fig. 9. From top to bottom, they are the multiview camera image method, the LiDAR point cloud method, and the fusion method. In the left figure, the multiview camera image detection method fails to detect distant cars and buses due to occlusion and a lack of depth information, while image distortion in rainy weather causes even nearby cars to be missed in detection. In addition, the rainy weather aggravates the sparsity of the point clouds, and the laser propagation in the rainy weather brings a certain degree of distortion. Thus, the LiDAR point clouds detection method is mistakenly the trees on both sides of the road as cars and cyclists. In contrast, our fusion detection algorithm eliminates both missed and false detections. The multiview camera image detection method in the right figure misses two riders due to low exposure. Although the LiDAR point clouds are unaffected by exposure at night, the

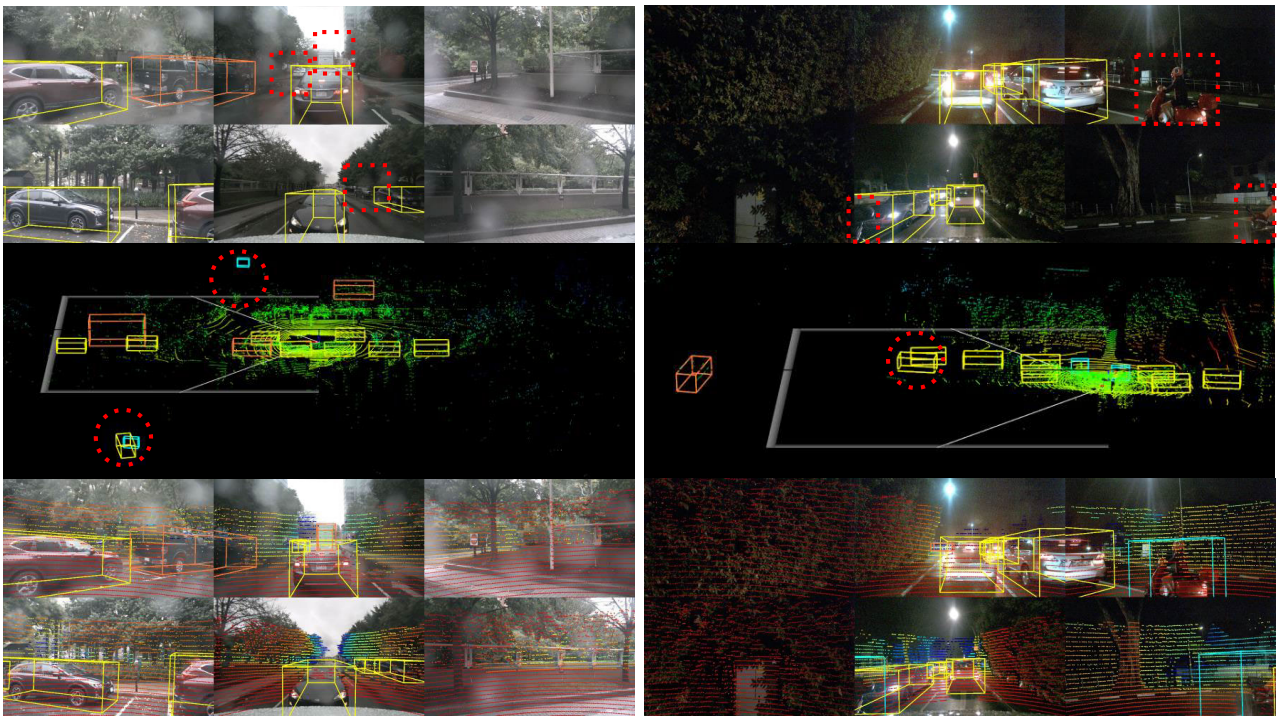

Fig. 9. 3-D Bbox visualization of the single-modal methods and the proposed multimodal methods. The first line shows the detection results generated by the multiview camera image method, the second line shows the detection results generated by the LiDAR point clouds (CenterPoint) method, and the third line shows the detection results generated by MixedFusion. The dotted boxes and circles indicate the missed and false detections, respectively. Our proposed fusion detection method eliminates both missed and false detections.

TABLE IV
PERFORMANCE ON NUSCENES VALIDATION SET BENCHMARK (ONLY RAIN AND NIGHT SCENES)

Method	Modality	AMOTA↑	AMOTP↓	MOTA↑	MOTP↓	IDS↓
CBMOT* [22]	C	15.2	169.0	13.2	72.3	256
AB3DMOT* [32]	L	54.5	83.6	48.6	39.8	178
MPN-Base. [52]	L	56.3	85.7	49.3	34.7	198
AB3DMOT [32]	L	57.4	79.1	51.2	33.6	165
CRF -MOT [53]	L	59.8	97.8	53.3	34.1	132
CenterTrack [31]	L	64.7	59.4	53.8	32.5	144
NEBP [54]	L	65.9	60.7	54.1	32.1	99
IPRL-TRI [55]	C+L	53.2	83.0	47.3	81.7	111
Prob -MOT [55]	C+L	59.7	67.5	52.1	39.6	125
CBMOT [22]	C+L	65.3	61.3	55.6	41.0	115
EagerMOT [19]	C+L	66.1	59.1	57.8	36.9	103
MLPMOT [22]	C+L	66.9	51.2	59.9	31.7	101
Ours	**C+L+R**	**67.9**	**59.2**	**62.1**	**36.8**	**97**

LiDAR point clouds method generates false alarms at remote distances due to the front occlusion of the vehicle. In contrast, our fusion detection algorithm eliminates missed and false detections.

2) Fusion Tracking: For the nuScenes dataset, averaged MOT Accuracy (AMOTA), averaged MOT precision (AMOTP), MOT accuracy (MOTA), MOT precision (MOTP), and ID Switching (IDS) are used to evaluate the performance of the algorithms. For the K-radar dataset, the AMOTA of different weather conditions and the AMOTA of all weather conditions are used to evaluate the tracking accuracy.

The comparative experimental results of fusion tracking in nuScenes and K-radar datasets are shown in Tables IV and V, respectively. On the nuScenes dataset, the proposed method exhibits improvements over the state-of-the-art method MLP-MOT [22]. The AMOTA and MOTA increased by 1% and 2.2%, respectively, and IDS decreased by 4. However, the AMOTP and MOTP decreased by 8% and 5.1%, respectively. This indicates that our proposed method, which incorporates radar data, has advantages and disadvantages. The NuScenes dataset utilizes a lower-resolution 3-D radar, which introduces noise and false alarms in the detection results.

On the one hand, the radar demonstrates significant advantages in adverse weather conditions, effectively detecting medium-to-long-range objects that a camera or LiDAR may not detect. It serves as a complement to fusion-based detection. On the other hand, the noise and false alarms introduce a considerable number of low-confidence and invalid detection results. These results may affect the data association process. Despite our efforts to mitigate the interference caused by false alarms through MPM strategies, it is inevitable that they still impact the tracking results.

On the radar dataset, the proposed method shows significant improvements over the state-of-the-art approach YONTD-MOT [24], specifically in all adverse weather conditions. The AMOTA for all scenes increased by 2.16%, with the most

TABLE V
PERFORMANCE ON K-RADAR TEST SET BENCHMARK

Methods	Modality	Normal	Overcast	Fog	Rain	Sleet	Light Snow	Heavy Snow	Total
AB3DMOT [32]	L	41.76	31.85	32.52	46.05	31.45	52.90	39.71	44.24
CenterTrack [31]	L	53.26	35.98	33.70	41.19	32.09	51.18	42.42	45.85
TrackMPNN [33]	L	62.09	42.45	37.55	38.68	34.94	49.37	44.66	46.40
PC3T [34]	L	65.00	55.64	36.46	46.96	33.95	41.32	38.53	47.03
UG3DMOT [35]	L	68.97	45.65	48.65	47.20	38.26	54.67	36.41	48.21
mmMOT [18]	C+L	59.97	42.49	36.80	37.95	37.75	51.13	46.94	48.42
JRMOT [36]	C+L	70.08	53.93	39.40	43.84	32.48	45.93	33.46	49.13
Be-Track [23]	C+L	62.75	64.94	44.70	44.66	31.98	55.14	35.44	50.25
DeepFusion [16]	C+L	62.79	70.87	40.71	48.43	42.00	53.88	44.21	50.70
EagerMOT [17]	C+L	69.99	74.91	36.14	46.40	33.49	53.46	36.53	51.61
YONTD-MOT [24]	C+L	72.12	73.24	46.87	42.87	39.67	55.53	42.42	52.63
Ours	**C+L+R**	**72.23**	**73.45**	**49.97**	**48.68**	**42.46**	**60.65**	**49.75**	**54.79**

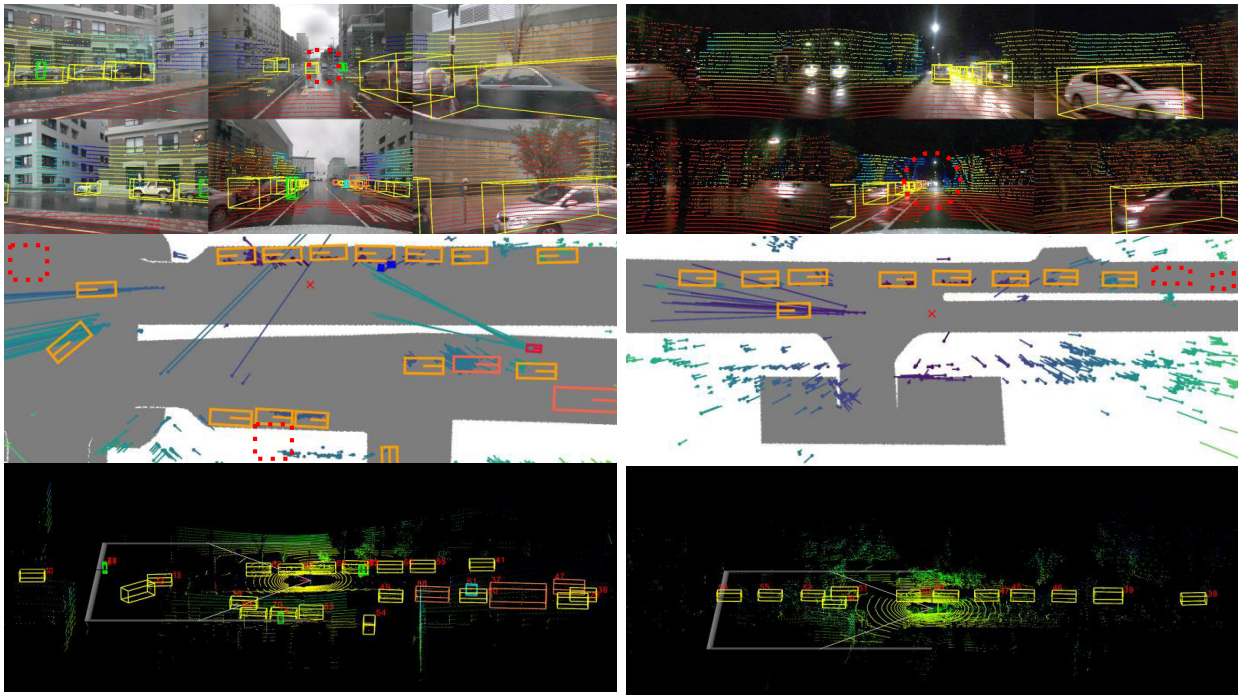

Fig. 10. 3-D Bbox visualization of fusion tracking results in rain (left) and night (right) scenes. The first line shows the 3-D Bboxes generated by the fusion detection of MixedFusion, the second line shows the 3-D Bboxes generated by the radar point clouds detection method, and the third line shows the 3-D Bboxes generated by the proposed fusion tracking of MixedFusion. Our proposed fusion tracking method eliminates both missed and false detections in severe weather conditions.

notable improvements observed in extreme weather conditions such as light snow and heavy snow, where camera and LiDAR distortions are severe. The K-radar dataset utilizes a 4-D high-resolution imaging radar, effectively reducing the probability of false alarms compared to radar. Therefore, the introduction of radar significantly improves across all weather conditions.

Similarly, to demonstrate the performance of the proposed method intuitively, the rain (left) and the night scenes (right) are randomly selected for 3-D Bbox visualization, as shown in Fig. 10. From top to bottom, they are the fusion detection results, the radar detection results, and the fusion tracking results. In the left figure, the fusion detection method can detect most objects. However, the distortion of sensor data caused by rain prevents the detection of remote vehicles. In addition, the radar detection method misses two pedestrians due to its low resolution. However, it has strong penetration ability and long propagation distance, and it successfully detected the objects that missed detection by the fusion detection method. The final fusion tracking algorithm successfully tracks all objects in the scene through the matching of the two. In the right figure, the fusion detection algorithm successfully detected all objects, but false detection occurred at a remote distance due to the influence of car lights in the night scene and the occlusion of point clouds in the front fleet. However, although radar missed the detection of the two vehicles behind due to the shielding, there was no false detection in the detection results. By matching the two, the final fusion tracking algorithm also successfully tracks all objects in the scene.

TABLE VI
FUSION DETECTION ABLATION STUDY IN NUSCENES VALIDATION SET (ONLY RAIN AND NIGHT SCENES)

HLSG	DETR (Box-Attention)	mAP↑	NDS↑
×	×	58.5	65.4
√	×	63.7(↑5.2)	68.5(↑3.1)
×	√	62.6(↑4.1)	67.2(↑1.8)
√	√	66.3(↑7.8)	70.5(↑5.1)

Baseline: CenterPoint

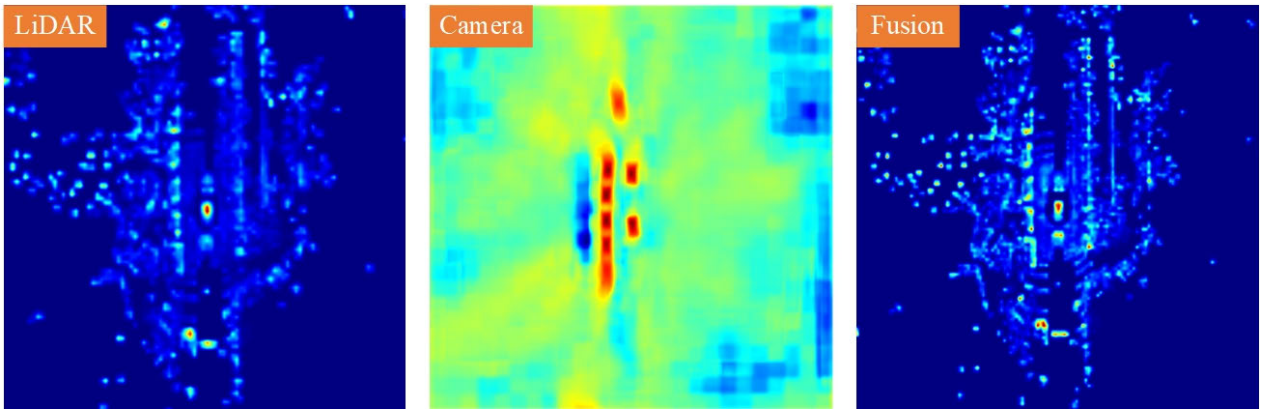

Fig. 11. Visualization of BEV features before and after fusion in randomly selected three scenes. The first figure shows the LiDAR BEV features derived from the baseline (CenterPoint), the second figure shows the image semantic BEV features derived from multiview camera images before fusion, and the third figure shows the fused BEV features derived from both multiview camera images and LiDAR point clouds after fusion. Compared with the first column, the confidence of the foreground points (objects to be detected) in the third column is higher. This means that our proposed HLSG fusion strategy can effectively fuse the LiDAR BEV with the image semantic BEV to generate the fused BEV with stronger feature representation capability.

D. Ablation Experiments

Similar to the dataset experiments, only rain and night scenes in the nuScenes validation set are selected to accurately verify the adaptability of each component in the fusion framework to severe weather.

1) Fusion Detection: We selected the powerful point clouds detection method CenterPoint as the baseline for the ablation experiment of fusion detection. To ensure fairness, the point cloud pipeline of the architecture has identical parameters to CenterPoint.

As shown in Table VI, if we only add the fusion module based on HLSG, the mAP and NDS will increase by 5.2% and 3.1%, respectively, which is a significant increase. In order to reduce the effect caused by other network architecture, the most common transformer encoder and decoder structures are used in this ablation study. As shown in Fig. 11, a scene is randomly selected for BEV feature visualization to illustrate the performance improvement of HLSG further. The color depth in the figure indicates the level of neural network attention. The first figure shows the LiDAR BEV features derived from the baseline (CenterPoint), the second figure shows the image semantic BEV features derived from the multiview camera images before fusion, and the third figure shows the fused BEV features derived from both the multiview camera images and LiDAR point clouds after fusion. Through observation, it can be found that the difference between foreground and background color in the image semantic BEV is very large, which obviously contains very rich semantic information. Compared to the LiDAR BEV, the foreground points in the fused BEV are noticeably brighter and can even outline the general shape to some extent. This means that the fused BEV has a stronger feature representation ability, proving the superior performance advantage of the proposed fusion strategy.

If only the DETR framework based on box-attention is utilized, the mAP and NDS will increase by 4.1% and 1.8%, respectively. Unlike the hard correlation between image and point clouds based on the joint calibration matrix used in the traditional method, our DETR framework is an adaptive soft-weighted association fusion. To further validate the effect of adaptive fusion based on box-attention, we also randomly selected two scenes for feature visualization, as shown in Fig. 12. The first row shows the heatmap of LiDAR point clouds features derived from the baseline (CenterPoint); the second row shows the heatmap of multiview camera image features derived from the camera pipeline of the proposed fusion detection method; the third row shows the heatmap of fused features derived from the proposed fusion detection method. Please note that during the entire training process of the network, only label information of 3-D Bbox from the dataset is used. The image pipeline has only the ResNet50 pretraining weights; we did not train them separately in the subsequent process. Although the label information of 2-D Bbox is not provided to constrain the image pipeline precisely, the network still accurately focuses on the foreground points in the multiview camera images. This means that the image pipeline is well-trained through adaptive soft-weighted association fusion based on the MHBA decoder in DETR. In addition, the foreground points in the third line are significantly clearer than those in the first line, further indicating that the image pipeline contributes to the final prediction. What is more, it is strong evidence to prove the effectiveness of the fusion strategy.

2) Fusion Tracking: The combination of CenterPoint and AB3DMOT was chosen as the baseline for subsequent ablation studies. To ensure fairness, all parameters of life cycle management in fusion tracking refer to AB3DMOT settings.

The ablation studies of fusion tracking are shown in Table VII. If only the fourth-level data association is removed, AMOTA and MOTA will decrease by 0.9% and 1.8%, respectively, AMOTP and IDS will increase by 0.1% and 4, respectively. The performance of the algorithm only decreases marginally. In fact, the majority of detected objects have been matched with the tracks at the first three levels of data association. However, the slight decrease in accuracy indicates that there are still effective objects and tracks that cannot be matched. This means that three levels of data association are insufficient for complete matching, and the fourth level of data association is required.

At this time, objects detected by radar but not by fusion detection are completely discarded, which means that the tracking algorithm almost entirely depends on the fusion detection results. The radar detection results are only used to adjust the priority of the fusion detection result and are rarely used effectively. The significant drop in accuracy can also explain two problems: first, radar detects some remote objects

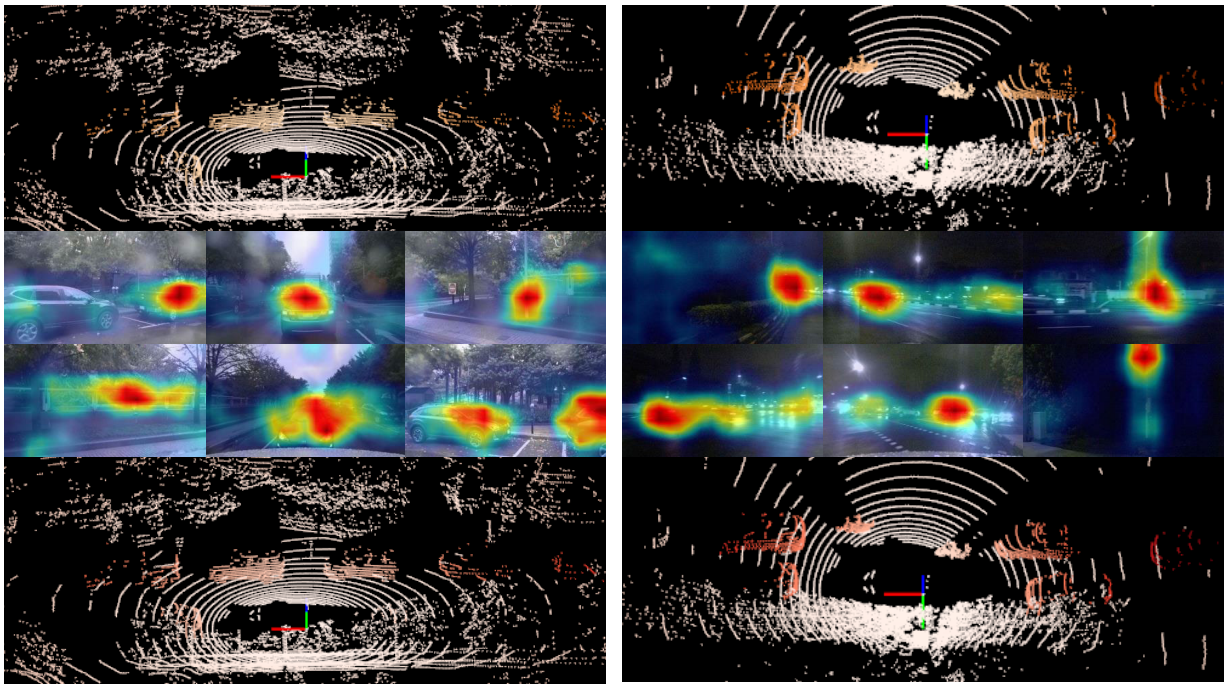

Fig. 12. Feature visualization at the level of the original input data. The first row shows the heatmap of LiDAR point clouds features derived from the baseline (CenterPoint), the second row shows the heatmap of multiview camera image features derived from the camera pipeline of the proposed fusion detection method, and the third row shows the heatmap of fused features derived from the proposed fusion detection method. The foreground points in the third line are significantly clearer than those in the first line, indicating that our proposed fusion strategy enables the network to infer more accurately whether a position is a foreground point. The second line shows that the network effectively focuses on the features of the foreground points in multiview images, demonstrating that even without providing 2-D label constraints, the image pipeline can still be well trained and contribute to the final prediction.

TABLE VII

FUSION TRACKING ABLATION STUDY IN NUSCENES VALIDATION SET (ONLY RAIN AND NIGHT SCENES)

Level	AMOTA↑	AMOTP↓	MOTA↑	IDs↓
Full	67.9	59.2	62.1	97
No 4th association	67.0(↓0.9)	59.3(↑0.1)	60.3(↓1.8)	101(↑4)
No 3rd association	63.8(↓4.1)	60.8(↑1.6)	55.2(↓6.9)	116(↑19)
No 1st and 2nd association	63.3(↓4.6)	61.0(↑1.8)	53.1(↓9)	120(↑23)
No fusion detection	60.6(↓7.3)	63.3(↑4.1)	52.3(↓9.8)	144(↑47)

No fusion detection (Baseline: CenterPoint + AB3DMOT)

that cannot be detected by the fusion detection algorithm, and second, radar has obvious advantages in adverse weather conditions. It can identify the objects missed by the fusion detection algorithm because of the data distortion and effectively supplement them in fusion tracking to achieve true complementary fusion.

Next, if the first-level and the second-level data associations are eliminated, AMOTA and MOTA will decrease by 0.5% and 2.1%, respectively, while AMOTP and IDS will increase by 0.2% and 4, respectively, which means that the algorithm's performance decreases slightly. Due to the fact that the essence of the first and second-level data association is to match fusion detection results based on their priority, it only adds a simple hierarchical association compared to AB3DMOT. Therefore, we merged them in the ablation experiment. The experimental results demonstrate that by giving a higher priority to objects jointly detected by both fusion detection and radar, it is possible to improve the accuracy of association and reduce the difficulty of subsequent data association.

Finally, if all data associations are removed, this means that we did not change the fusion tracking method and employed AB3DMOT to track only the fusion detection results. AMOTA and MOTA will decrease by 2.7% and 0.8%, respectively, while AMOTP and IDS will rise by 2.3% and 24, respectively. The performance of the algorithm will decrease significantly. The experimental results indicate that the high-precision detection results afforded by fusion detection can significantly enhance the MOT algorithm. Even if the tracking algorithm is not optimized, the accuracy of MOT will be improved as long as enough accurate detection results can be provided. It also demonstrates that the effective utilization of fusion detection results is crucial and that we must introduce the radar to meet the assumption of observation independence.

V. CONCLUSION

In this article, we propose an efficient multimodal data fusion framework called MixedFusion. Two new fusion strategies based on HLSG and MPM are proposed in the framework to optimize the fusion detection and the fusion tracking method. It achieves the efficient utilization of multimodal data and realizes the complementary fusion on this basis. A large number of experimental results indicate that the proposed fusion framework significantly improves the performance of both 3-D object detection and tracking in adverse weather conditions.

REFERENCES

[1] C. Lin, D. Tian, X. Duan, J. Zhou, D. Zhao, and D. Cao, "3D-DFM: Anchor-free multimodal 3-D object detection with dynamic fusion module for autonomous driving," *IEEE Trans. Neural Netw. Learn. Syst.*, early access, May 13, 2022, doi: 10.1109/TNNLS.2022.3171553.

[2] Y. Chen, H. Li, R. Gao, and D. Zhao, "Boost 3-D object detection via point clouds segmentation and fused 3-D GIoU-L_1 loss," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 33, no. 2, pp. 762–773, Feb. 2022, doi: 10.1109/TNNLS.2020.3028964.

[3] S.-C. Huang, Q.-V. Hoang, and T.-H. Le, "SFA-Net: A selective features absorption network for object detection in rainy weather conditions," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 34, no. 8, pp. 5122–5132, Aug. 2023, doi: 10.1109/TNNLS.2021.3125679.

[4] X. Chang, P. Ren, P. Xu, Z. Li, X. Chen, and A. Hauptmann, "A comprehensive survey of scene graphs: Generation and application," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 1, pp. 1–26, Jan. 2023, doi: 10.1109/TPAMI.2021.3137605.

[5] Y. Li et al., "Deep learning for LiDAR point clouds in autonomous driving: A review," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 32, no. 8, pp. 3412–3432, Aug. 2021, doi: 10.1109/TNNLS.2020.3015992.

[6] H. Caesar et al., "NuScenes: A multimodal dataset for autonomous driving," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Seattle, WA, USA, Jun. 2020, pp. 11618–11628, doi: 10.1109/CVPR42600.2020.01164.

[7] D.-H. Paek, S.-H. Kong, and K. Tirta Wijaya, "K-radar: 4D radar object detection for autonomous driving in various weather conditions," Jan. 2023, *arXiv:2206.08171*. Accessed: Jun. 16, 2023.

[8] X. Chen, T. Zhang, Y. Wang, Y. Wang, and H. Zhao, "FUTR3D: A unified sensor fusion framework for 3D detection," Mar. 2022, *arXiv:2203.10642*. Accessed: Mar. 25, 2022.

[9] M. Li, P.-Y. Huang, X. Chang, J. Hu, Y. Yang, and A. Hauptmann, "Video pivoting unsupervised multi-modal machine translation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 3, pp. 3918–3932, Mar. 2023, doi: 10.1109/TPAMI.2022.3181116.

[10] R. Zhou, X. Chang, L. Shi, Y.-D. Shen, Y. Yang, and F. Nie, "Person reidentification via multi-feature fusion with adaptive graph learning," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 31, no. 5, pp. 1592–1601, May 2020, doi: 10.1109/TNNLS.2019.2920905.

[11] F. Nobis, E. Shafiei, P. Karle, J. Betz, and M. Lienkamp, "Radar voxel fusion for 3D object detection," *Appl. Sci.*, vol. 11, no. 12, p. 5598, Jun. 2021, doi: 10.3390/app11125598.

[12] X. Bai et al., "TransFusion: Robust LiDAR-camera fusion for 3D object detection with transformers," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2022, pp. 1090–1099.

[13] S. Pang, D. Morris, and H. Radha, "Fast-CLOCs: Fast camera-LiDAR object candidates fusion for 3D object detection," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. (WACV)*, Waikoloa, HI, USA, Jan. 2022, pp. 3747–3756, doi: 10.1109/WACV51458.2022.00380.

[14] Z. Liu et al., "BEVFusion: Multi-task multi-sensor fusion with unified bird's-eye view representation," May 2022, *arXiv:2205.13542*. Accessed: May 28, 2022.

[15] S. Xu, D. Zhou, J. Fang, J. Yin, Z. Bin, and L. Zhang, "FusionPainting: Multimodal fusion with adaptive attention for 3D object detection," Aug. 2021, *arXiv:2106.12449*. Accessed: Sep. 25, 2021.

[16] X. Wang, C. Fu, Z. Li, Y. Lai, and J. He, "DeepFusionMOT: A 3D multi-object tracking framework based on camera-LiDAR fusion with deep association," *IEEE Robot. Autom. Lett.*, vol. 7, no. 3, pp. 8260–8267, Jul. 2022, doi: 10.1109/LRA.2022.3187264.

[17] A. Kim, A. Ošep, and L. Leal-Taixé, "EagerMOT: 3D multi-object tracking via sensor fusion," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, Jun. 2021, pp. 11315–11321, doi: 10.1109/ICRA48506.2021.9562072.

[18] W. Zhang, H. Zhou, S. Sun, Z. Wang, J. Shi, and C. Change Loy, "Robust multi-modality multi-object tracking," Sep. 2019, *arXiv:1909.03850*. Accessed: Feb. 27, 2022.

[19] S. Vora, A. H. Lang, B. Helou, and O. Beijbom, "PointPainting: Sequential fusion for 3D object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 4604–4612.

[20] Y. Zhang, Q. Zhang, J. Hou, Y. Yuan, and G. Xing, "Bidirectional propagation for cross-modal 3D object detection," May 2023, *arXiv:2301.09077*. Accessed: Jun. 16, 2023.

[21] H. Zhu et al., "VPFNet: Improving 3D object detection with virtual point based LiDAR and stereo data fusion," Dec. 2021, *arXiv:2111.14382*. Accessed: Dec. 11, 2021.

[22] N. Benbarka, J. Schröder, and A. Zell, "Score refinement for confidence-based 3D multi-object tracking," Jul. 2021, *arXiv:2107.04327*. Accessed: Feb. 27, 2022.

[23] M. Dimitrievski, P. Veelaert, and W. Philips, "Behavioral pedestrian tracking using a camera and LiDAR sensors on a moving vehicle," *Sensors*, vol. 19, no. 2, p. 391, Jan. 2019, doi: 10.3390/s19020391.

[24] X. Wang, J. He, C. Fu, T. Meng, and M. Huang, "You only need two detectors to achieve multi-modal 3D multi-object tracking," 2023, *arXiv:2304.08709*.

[25] R. Nabati and H. Qi, "CenterFusion: Center-based radar and camera fusion for 3D object detection," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2021, pp. 1526–1535, doi: 10.1109/WACV48630.2021.00157.

[26] T. Yin, X. Zhou, and P. Krähenbühl, "Multimodal virtual point 3D detection," Nov. 2021, *arXiv:2111.06881*. Accessed: Nov. 18, 2021.

[27] Y. Chen, Y. Li, X. Zhang, J. Sun, and J. Jia, "Focal sparse convolutional networks for 3D object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*. New Orleans, LA, USA: IEEE, Jun. 2022, pp. 5418–5427, doi: 10.1109/CVPR52688.2022.00535.

[28] Z. Chen et al., "AutoAlign: Pixel-instance feature aggregation for multi-modal 3D object detection," Jan. 2022, *arXiv:2201.06493*. Accessed: Feb. 23, 2022.

[29] L. Zhang et al., "TN-ZSTAD: Transferable network for zero-shot temporal activity detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 3, pp. 3848–3861, Mar. 2023, doi: 10.1109/TPAMI.2022.3183586.

[30] L. Zhang et al., "Weakly aligned feature fusion for multimodal object detection," *IEEE Trans. Neural Netw. Learn. Syst.*, early access, Aug. 26, 2021, doi: 10.1109/TNNLS.2021.3105143.

[31] T. Yin, X. Zhou, and P. Krähenbühl, "Center-based 3D object detection and tracking," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 11784–11793.

[32] X. Weng, J. Wang, D. Held, and K. Kitani, "AB3DMOT: A baseline for 3D multi-object tracking and new evaluation metrics," Aug. 2020, *arXiv:2008.08063*. Accessed: Aug. 30, 2022.

[33] A. Rangesh, P. Maheshwari, M. Gebre, S. Mhatre, V. Ramezani, and M. M. Trivedi, "TrackMPNN: A message passing graph neural architecture for multi-object tracking," May 2021, *arXiv:2101.04206*. Accessed: Jun. 16, 2023.

[34] H. Wu, W. Han, C. Wen, X. Li, and C. Wang, "3D multi-object tracking in point clouds based on prediction confidence-guided data association," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 6, pp. 5668–5677, Jun. 2022, doi: 10.1109/TITS.2021.3055616.

[35] J. He, C. Fu, and X. Wang, "3D multi-object tracking based on uncertainty-guided data association," 2023, *arXiv:2303.01786*.

[36] A. Shenoi et al., "JRMOT: A real-time 3D multi-object tracker and a new large-scale dataset," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*. Las Vegas, NV, USA: IEEE, Oct. 2020, pp. 10335–10342, doi: 10.1109/IROS45743.2020.9341635.

[37] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *Computer Vision—ECCV*, vol. 12346, A. Vedaldi, H. Bischof, T. Brox, and J.-M. Frahm, Eds. Cham, Switzerland: Springer, 2020, pp. 213–229, doi: 10.1007/978-3-030-58452-8_13.

[38] J. Deng, S. Shi, P. Li, W. Zhou, Y. Zhang, and H. Li, "Voxel R-CNN: Towards high performance voxel-based 3D object detection," in *Proc. AAAI Conf. Artif. Intell.*, 2021, vol. 35, no. 2, pp. 1201–1209.

[39] A. Saha, O. M. Maldonado, C. Russell, and R. Bowden, "Translating images into maps," Mar. 2021, *arXiv:2110.00966*. Accessed: Aug. 30, 2022.

[40] D.-K. Nguyen, J. Ju, O. Booij, M. R. Oswald, and C. G. M. Snoek, "BoxeR: Box-attention for 2D and 3D transformers," May 2022, *arXiv:2111.13087*. Accessed: May 31, 2022.

[41] A. Zhang, F. Erlik Nowruzi, and R. Laganiere, "RADDet: Range-Azimuth-Doppler based radar object detection for dynamic road users," May 2021, *arXiv:2105.00363*. Accessed: Feb. 28, 2022.

[42] E. Baser, V. Balasubramanian, P. Bhattacharyya, and K. Czarnecki, "FANTrack: 3D multi-object tracking with feature association network," in *Proc. IEEE Intell. Vehicles Symp. (IV)*, Paris, France, Jun. 2019, pp. 1426–1433, doi: 10.1109/IVS.2019.8813779.

[43] Z. Li et al., "BEVFormer: Learning bird's-eye-view representation from multi-camera images via spatiotemporal transformers," Jul. 2022, *arXiv:2203.17270*. Accessed: Aug. 31, 2022.

[44] A. H. Lang, S. Vora, H. Caesar, L. Zhou, J. Yang, and O. Beijbom, "PointPillars: Fast encoders for object detection from point clouds," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 12697–12705.

[45] Y. Yan, Y. Mao, and B. Li, "SECOND: Sparsely embedded convolutional detection," *Sensors*, vol. 18, no. 10, p. 3337, Oct. 2018, doi: 10.3390/s18103337.
[46] J. Yan et al., "Cross modal transformer: Towards fast and robust 3D object detection," Mar. 2023, *arXiv:2301.01283*. Accessed: Jun. 16, 2023.
[47] S. Shi, Z. Wang, J. Shi, X. Wang, and H. Li, "From points to parts: 3D object detection from point cloud with part-aware and part-aggregation network," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 8, pp. 2647–2664, Aug. 2021, doi: 10.1109/TPAMI.2020.2977026.
[48] S. Shi et al., "PV-RCNN: Point-voxel feature set abstraction for 3D object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*. Seattle, WA, USA: IEEE, Jun. 2020, pp. 10526–10535, doi: 10.1109/CVPR42600.2020.01054.
[49] H. Wu, J. Deng, C. Wen, X. Li, C. Wang, and J. Li, "CasA: A cascade attention network for 3-D object detection from LiDAR point clouds," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5704511, doi: 10.1109/TGRS.2022.3203163.
[50] H. Wu, C. Wen, W. Li, X. Li, R. Yang, and C. Wang, "Transformation-equivariant 3D object detection for autonomous driving," Dec. 2022, *arXiv:2211.11962*. Accessed: Jun. 16, 2023.
[51] Y. Kim, K. Park, M. Kim, D. Kum, and J. Won Choi, "3D dual-fusion: Dual-domain dual-query camera-LiDAR fusion for 3D object detection," Feb. 2023, *arXiv:2211.13529*. Accessed: Jun. 16, 2023.
[52] G. Braso and L. Leal-Taixe, "Learning a neural solver for multiple object tracking," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Seattle, WA, USA, Jun. 2020, pp. 6246–6256, doi: 10.1109/CVPR42600.2020.00628.
[53] J.-N. Zaech, D. Dai, A. Liniger, M. Danelljan, and L. Van Gool, "Learnable online graph representations for 3D multi-object tracking," Apr. 2021, *arXiv:2104.11747*. Accessed: May 8, 2022.
[54] V. G. Satorras and M. Welling, "Neural enhanced belief propagation on factor graphs," in *Proc. Int. Conf. Artif. Intell. Statist.*, 2021, pp. 685–693.
[55] H.-K. Chiu, J. Li, R. Ambrus, and J. Bohg, "Probabilistic 3D multi-modal, multi-object tracking for autonomous driving," Oct. 2021, *arXiv:2012.13755*. Accessed: Aug. 30, 2022.

Cheng Zhang received the B.S. degree from Jiangsu University, Zhenjiang, China, in 2020, where he is currently pursuing the Ph.D. degree.

His research interests include computer vision, deep learning, and intelligent vehicles.

Hai Wang (Senior Member, IEEE) received the B.S., M.S., and Ph.D. degrees from the School of Instrument Science and Engineering, Southeast University, Nanjing, China, in 2006, 2008, and 2012, respectively.

In 2012, he joined the School of Automotive and Traffic Engineering, Jiangsu University, Zhenjiang, China, where he is currently working as a Professor. He has published more than 50 articles in the field of machine vision-based environment sensing for intelligent vehicles. His research interests include computer vision, intelligent transportation systems, and intelligent vehicles.

Long Chen received the Ph.D. degree in vehicle engineering from Jiangsu University, Zhenjiang, China, in 2002.

His research interests include intelligent automobiles and vehicle control systems.

Yicheng Li received the Ph.D. degree in vehicle engineering from the Wuhan University of Technology, Wuhan, China, in 2018.

He is currently an Assistant Professor at the Automotive Engineering Research Institute, Jiangsu University, Zhenjiang, China. His research interests include intelligent vehicle localization, intelligent transportation systems, computer vision, and 3-D data processing.

Yingfeng Cai (Senior Member, IEEE) received the B.S., M.S., and Ph.D. degrees from the School of Instrument Science and Engineering, Southeast University, Nanjing, China, in 2006, 2009, and 2013, respectively.

In 2013, she joined the Automotive Engineering Research Institute, Jiangsu University, Zhenjiang, China, where she is currently working as a Professor. She has published more than 100 articles in high-level journals, including IEEE TRANSACTIONS ON INTELLIGENT TRANSPORTATION SYSTEMS, IEEE TRANSACTIONS ON INTELLIGENT VEHICLES, IEEE TRANSACTIONS ON IMAGE PROCESSING, IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS, and IEEE TRANSACTIONS ON VEHICULAR TECHNOLOGY in the field of sensing and control for intelligent vehicles. Her research interests include computer vision, intelligent transportation systems, and intelligent automobiles.

Dr. Cai got the National Fund for Distinguished Young Scholars of China.