IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 9 August 2024, accepted 22 August 2024, date of publication 27 August 2024, date of current version 5 September 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3450535

RESEARCH ARTICLE

Temporal-Based Multi-Sensor Fusion for 3D Perception in Automated Driving System

LING HUANG[1,2], YIXUAN ZENG[1], SHUO WANG[3], RUNMIN WEN[1], AND XINGYU HUANG[1]

[1]School of Civil Engineering and Transportation, South China University of Technology, Guangzhou 510641, China
[2]Jiangsu Province Collaborative Innovation Center of Modern Urban Traffic Technologies, Southeast University, Nanjing 211189, China
[3]School of Architecture, South China University of Technology, Guangzhou 510641, China

Corresponding author: Shuo Wang (shwang@scut.edu.cn)

This work was supported in part by Guangdong Basic and Applied Basic Research Foundation under Grant 2023A1515010742; and in part by the Open Project of Key Laboratory of Interactive Media Design and Equipment Service Innovation, Ministry of Culture and Tourism under Grant 202008.

ABSTRACT The 3D object detection task is a crucial subtask of the environment perception module in Automated Driving System (ADS). The accuracy of object detection directly impacts the effectiveness of downstream autonomous driving tasks such as tracking, prediction, and planning. Existing 3D object detection networks in ADS that rely on multi-sensor fusion lack the utilization of temporal information and fail to fully consider the dynamic nature of the surrounding traffic environment. Therefore, we proposed BEVTemporal for ADS, which fuses LIDAR point clouds and surrounding multi-channel images. Its unique temporal module establishes the correlation between historical data and current data, effectively leveraging the temporal information of surrounding objects. We train and validate BEVTemporal on the nuScenes datasets. After incorporating temporal module, the Average Precision (AP) metrics of the network improved by 0.3%∼1.7%, and mean Average precision (mAP) achieves 0.87% higher, nuScenes Detection Score (NDS) increased by 0.46%. The validation results on subsets of occluded objects show that the model effectively alleviates the problem of missed detection caused by sample occlusion, with significant improvements observed for heavily occluded samples. In different scenarios (sunny, rainy, daytime, night), mAP improvement ranges from 0.75% to 1.18%. Notably, in challenging scenarios such as rainy and night, AP can be improved by up to 3.6%. The experimental results show that BEVTemporal not only improves the accuracy of the 3D object detection network, but also significantly enhances the robustness of model in various scenarios and recall of objects under low visibility conditions.

INDEX TERMS Temporal information, 3D object detection, multi-sensor fusion, automated driving system.

I. INTRODUCTION

In the context of rapid development in technologies such as big data, the Internet of Things (IoT), and cloud computing, the intelligent vehicle industry, as a crucial pillar of the new round of technological revolution, contributes to the advancement of artificial intelligence, and related fields [1], [2]. The task of 3D traffic object detection serves as a key technology for environment perception in automated driving systems (ADS), providing reliable environmental information for intelligent vehicles under complex traffic scenarios.

The essence of the 3D object detection task is to classify and localize objects in a given scene by sensors data. It provides data support for perception module of autonomous driving systems, as well as subsequent tasks such as tracking and prediction. Therefore, as a critical task within the perception module, improving the accuracy of detection network enables intelligent vehicles to make accurate judgments and select rational strategies.

Due to the complexity of the driving environment, many researchers have primarily employed deep learning methods to achieve accurate detection and recognition required for ADS perception tasks [1]. Recently, people tend to designed multi-sensor fusion networks to fully leverage the advantages

The associate editor coordinating the review of this manuscript and approving it for publication was Miaohui Wang.

of different sensors. Compared to traditional machine learning methods, deep learning networks do not require manual feature engineering and can directly represent and capture complex patterns in the data. In deep learning tasks, temporal information is widely utilized to capture the important context of data in terms of its sequential order and evolution. This application of temporal information is prevalent in various tasks such as speech recognition, weather forecasting, and financial market prediction. Similarly, in the field of ADS, temporal information is often employed for tasks like trajectory prediction and intent prediction. However, there has been relatively limited research on incorporating temporal information in the context of 3D object detection tasks, mainly focusing on single sensor data types such as image-based or LiDAR-based approaches.

In the field of ADS, 3D object detection algorithms typically employ various sensor-based methods, including camera-based, LiDAR-based, and multi-sensor fusion networks. Compared to single modality netwoks, multi-sensor fusion networks leverage the complementary information from different sensors to provide more comprehensive information for intelligent vehicles. Currently, in order to fully exploit the advantages of different sensors, ADS primarily use multi-sensor fusion to achieve 3D object detection tasks, which has shown promising detection performance.

However, most of these networks treat sensors data as independent entities and fail to establish temporal correlations between consecutive frames. Due to the highly dynamic and complexity of real-world traffic scenarios, the positions, velocities, accelerations, and other attributes of various traffic participants (including motor vehicles, non-motorized vehicles, and pedestrians) surrounding an autonomous vehicle change over time. Moreover, these objects frequently occlude each other, further complicating the perception process. Therefore, if the temporal information is not taken into consideration, the model is susceptible to missed detections due to occlusions. Additionally, traffic scenes often contain a large number of repetitive static objects, such as road curbs, barriers, and other structures. If model can be specifically optimized for the characteristics of road traffic scenes, it may be capable of effectively handling the complexities and dynamics inherent in real-world traffic environments.

Based on these considerations, we propose a LiDAR-camera fusion network BEVTemporal to integrate temporal information in bird's-eye view (BEV). By fusing features over a specific time window, the model can capture the dynamic information of traffic participants more effectively and understand the temporal characteristics of the traffic scene. It also leverages the complementary information between different sensors, fully exploiting the accuracy of depth information provided by point clouds and the rich semantic and textural features of RGB data.

A series of experimental results demonstrate that our BEVTemporal model, which incorporates temporal information by associating information from consecutive frames, not only improves detection accuracy but also reduces the issue of missed detections caused by occlusions. Additionally, it exhibits better adaptability to weather and lighting variations. In summary, our contributions are as follows:

- We propose a novel 3D object detection network called BEVTemporal, which combines data from surround-view cameras and LiDAR sensors. This network effectively reduces the missed detection rate, enhances the detection capability, and improves robustness against various weather and lighting conditions.
- In order to capture the dynamic and structural characteristics of traffic scenes, we introduce a temporal module into BEVTemporal. By learning features from multiple types and angles over a period of time, BEVTemporal provides more accurate BEV local perception information for intelligent vehicles.
- To validate the adaptability of BEVTemporal in dynamic traffic environments for autonomous driving, we conducted three different object detection experiments on the nuScenes dataset: 3D multi-object detection experiment, varying occlusion level detection experiment, and different weather and light conditions detection experiment.

The results of the experiments demonstrate that after considering temporal information, the model not only improves detection accuracy but also reduces missed detection caused by sample occlusion. Additionally, the model exhibits better performance in handling weather and light variations.

II. RELATED RESEARCH

A. AUTOMATED DRIVING SYSTEM

ADS utilize sensor data and a combination of software and hardware algorithms to control vehicles. They can perceive the surrounding environment by using various sensors such as LiDAR, cameras, Radar and so on. Through real-time data processing and analysis, these systems enable functions such as navigation, driving, and control of the vehicle. ADS have the potential to enhance road safety and simplify driving for both novice and experienced drivers. The development of computing power and hardware since 2010 has greatly contributed to the research of autonomous driving systems. This has led to real-world testing of autonomous driving on public roads by several companies. Currently, autonomous driving technology is gradually becoming commercialized, with companies like Tesla and Huawei introducing vehicles with varying levels of automation. Different companies often prefer different sensor configurations for building automated driving systems. For example, Tesla tends to rely on a pure vision-based approach, while Xiaopeng Motors (Xpeng) incorporates additional data from LiDAR.

Due to the large amount of sensor data that needs to be processed, research on autonomous driving systems has focused on optimizing data processing techniques for upstream tasks, such as data compression [3], [4] and upsampling [5]. Another area of focus is improving the accuracy of subtasks within autonomous driving systems by applying new tech-

nologies. For instance, Wang and Tian [6] conducted research on the hazard perception model for for ADS, while Xu and Miyahara [7] designed an object recognition system for ADS under hazy condition. Pre-CoAD [8] enables human drivers to intervene in a validated existing ADS on public roads using gaze-based input and visual output, thereby enhancing the performance of the autonomous driving system.

B. 3D OBJECT DETECTION

1) CAMERA-BASED 3D PERCEPTION

In the early stages of camera-only 3D detection models, most networks [9], [10], [11] focused on using monocular camera. However, due to the absence of scale information, monocular approaches often predict 3D bounding boxes based on 2D bounding boxes and geometric constraints, which poses challenges in accurate depth estimation. Therefore, some works [12], [13] utilize stereo cameras to calculate disparity which can be used to generate more accurate depth maps and 3D bounding boxes. To obtain a more comprehensive understanding of the environment, DETR3D [14] extends the work of the DETR [15] to 3D space by employing surround-view cameras. Recently, some research has discovered that transferring features to BEV space can provide a unified representation of the surrounding environment. Therefore, BEVDet [16] and BEVFormer [17] further extend the work of LSS [18] to BEV space, and achieve impressive results.

Image data contains rich details and textures, but it is limited by the scale uncertainty. Therefore, estimating the required depth information solely from 2D images imposes constraints on the detection accuracy.

2) LiDAR-BASED 3D PERCEPTION

To address the challenges of sparse and unordered point clouds, LiDAR methods convert the point cloud into a more structured format, such as voxels [22], [29] or pillars [28]. These structured representations provide a grid-like structure that allows for easier processing and feature extraction. However, there are also studies [19], [20], [21], that directly extract features from the raw point cloud without any preprocessing. This approach allows for a more fine-grained analysis of the point cloud, but it may also require more computational resources.

Compared to image data, LiDAR provides more accurate depth information. However, due to the sparsity of point clouds, LiDAR is more likely to lose details in 3D space. Therefore, LiDAR-based networks often face challenges in accurately detecting small-sized and distant targets.

3) MULTI-SENSOR FUSION

In order to complement different sensor and fully leverage the advantages of various information, people have dedicated their efforts to multi-sensor fusion networks. For instance, MV3D [23] projects the 3D LiDAR point clouds onto images to extract Regions of Interest (ROIs) features. However, these features always contain irrelevant background information. ContFuse [24] projects image features onto point clouds and fuses information in feature level. Nevertheless, it may result in the loss of image semantic information due to the sparsity of point clouds. SFD [25] generates dense point clouds based on 2D images to address the issue of projection, elevating the data from 2D space to 3D space for more precise fusion.

Although the research mentioned above exhibit improved detection accuracy compared to single-modal networks, monocular image data only provides front-view and lacks comprehensive environmental information. Therefore, Transfusion [26] and BEVFusion [27] utilize both LiDAR point cloud and surround-view image to provide BEV perception. It transforms the environmental information surrounding vehicle onto a 2D plane, representing real road scene information and further enhances the accuracy of detection tasks. Recently BEV perception becomes a mainstream direction for 3D object detection.

However, current research on multi-sensor fusion networks primarily focuses on instantaneous data, lacking sufficient utilization of temporal information and optimization for traffic scenarios. Therefore, we propose a model that can effectively leverage temporal information to better handle occlusion challenges and improve the robustness of multi-sensor fusion detection networks in various weather and lighting conditions.

III. METHODOLOGY

BEVTemporal employs multiple-view surround cameras and LiDAR to accomplish BEV 3D perception tasks in traffic scenarios. Understanding the dynamic characteristics of traffic objects via temporal attention mechanisms, the model can effectively adapt to the specific attributes of traffic scenes, thereby enhancing its performance.

We present the overall architecture of BEVTemporal in Fig. 1. Our framework contains four fundamental modules, which are 1) Raw data input and preprocessing, 2) Image and point cloud feature extraction, 3) Feature scale transformation and fusion, and 4) Detection head with temporal module. In the subsequent sections of the article, we will provide detailed explanations of the latter three modules.

A. IMAGE AND POINT CLOUD FEATURE EXTRACTION

First, we need to extract feature from image and LiDAR point cloud. For point cloud feature extraction, we use VoxelNet [29] as backbone. VoxelNet is currently the most mainstream backbone in voxelized LiDAR methods, and therefore, we have chosen it as the backbone for extracting features from LiDAR point clouds. This backbone voxelizes the unordered LiDAR point cloud, and extracts features within each voxel through operations such as grouping and random sampling. Therefore, the point clouds can be converted into an ordered representation. Meanwhile, it also incorporates feature from different scales, effectively preserving abundant information.

For image feature extraction, we use Swin-Transformer [30] as backbone. It processes six-channel input image data

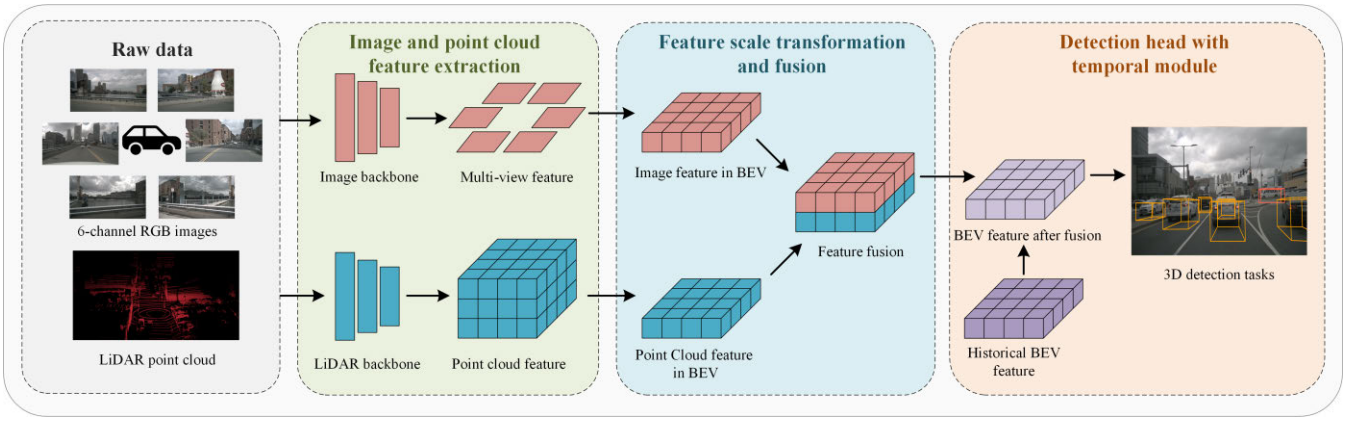

FIGURE 1. Overall architecture of BEVTemporal.

separately. Swin-Transformer achieves this by dividing each channel of the image into overlapping image windows of different sizes. This approach allows it to capture the relationships between different image patches more effectively and expand the model's receptive field, while keeping the computational cost relatively low.

B. FEATURE SCALE TRANSFORMATION AND FUSION

Due to the inconsistent feature dimensions between RGB image and LiDAR point cloud, overcoming the challenge of fusing information from different dimensions is indeed necessary. In terms of dimension selection, RGB images provide a front-view perspective and lack depth information, which causes object deformations. On the other hand, BEV representation preserves the scale of objects and avoids object distortion issues. Meanwhile, it better represents road scene information and facilitates more effective learning of 3D features by convolutional networks [23]. Therefore, we tend to transform 2D image features into BEV space to merge the information from different sensors.

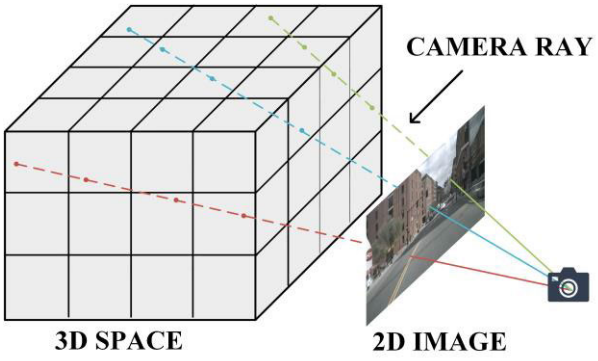

FIGURE 2. Feature projection illustration. (N, H, W, D).

The relationship between image features and depth information is not simply linear. Meanwhile, factors such as road conditions and camera vibration during data collection can cause significant errors by merely using camera intrinsic and extrinsic parameters to estimate depth. Taking inspiration from the BEVFusion model [24], in this paper, we discretize the features along the camera rays into D points and perform scaling and quantization operations. We predict the probabilities of image features appearing at different depths in 3D space. As shown in Figure 2, we obtain features with dimensions (N, H, W, D), where N represents the number of camera channels (i.e., the number of cameras in the nuScenes dataset is 6, so N is set as 6 in our experiments), $H \times W$ represents the size of the feature map, D represents the number of scales. Furthermore, we apply BEV pooling to process features within each grid in the spatial domain, mapping them to the BEV representation. We leverage precomputation and downsampling operations to accelerate the BEV pooling process [32]. After unifying the representation scales, features from different sensors are fused to obtain BEV features. These BEV features are then connected to correspo nding task heads, enabling the completion of various perception tasks.

C. TEMPORAL INFORMATION MODULE DESIGN

1) OVERALL MODULE STRUCTURE

Current multi-sensor fusion perception networks are limited to detect 3D object mostly based on instantaneous features. However, it is evident from human perception that drivers often make judgments by comprehensively considering both the historical and current road environments [33]. By referencing historical data, it is possible to mitigate false positives or false negatives caused by occlusions between objects, facilitating the model's understanding of dynamic characteristics of traffic scene. Moreover, aggregating temporal information into the model typically improves the robustness of the 3D object detection network across different scenarios, enabling autonomous vehicles to better respond to various environmental challenges.

Therefore, we propose a detection head with a temporal self-attention layer which comprehensively consider both historical BEV features and current BEV features, as shown in Fig. 3. By referencing historical features, we aim to improve

the model's ability to learn the dynamic characteristics of each objects in the traffic scene, thereby enhancing the overall performance of the model.

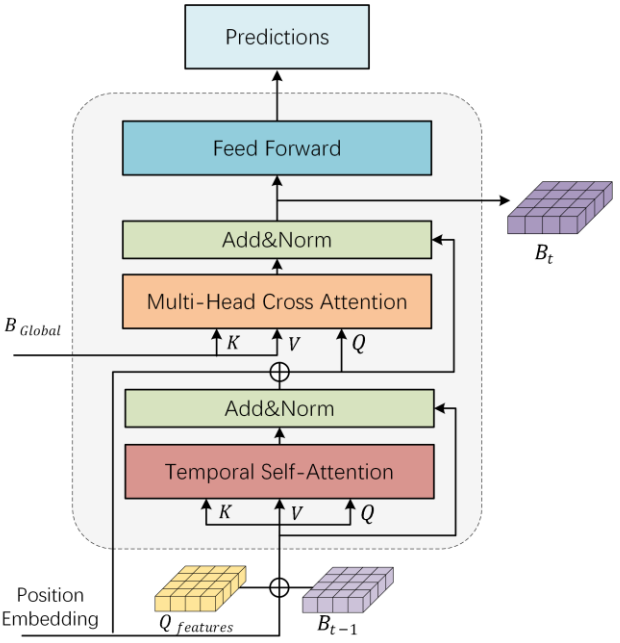

FIGURE 3. Temporal module structure.

2) CALCULATE TEMPORAL SELF-ATTENTION

To compute the temporal self-attention, object queries need to be initialized. Firstly, we concatenate the point cloud features and image features under the BEV representation to obtain the BEV feature map B_{Global}. Based on this feature map, we predict the probabilities of the center points for various object classes. We then generate a heat map, $\widehat{H} \in \mathbb{R}^{X\times Y\times K}$, where $X \times Y$ represents the feature dimensions, and K represents the number of object categories. By using the heat map to represent the center points and bounding boxes of objects, we can effectively address the issue of object overlap. Additionally, without the need for predefined anchor boxes, we can reduce the computational overhead of the network [34].

We select the top N candidates from each of the K categories in the heatmap as the initial object queries, denoted as Q_{object}, under the BEV space. We embed the category into Q_{object} to obtain the feature queries $Q_{features}$. Unlike traditional convolutional layers that possess local correlation and translation invariance, attention mechanisms only focus on the value of different elements in the sequence, without considering the relative positional information between elements. Therefore, we embed position into the sequence and use positional information of the candidate boxes to initialize the position queries $Q_{position}$ to capture the positional relationships within the sequence.

For the second frame and afterward (t $\neq$ 0), the temporal module needs to utilize both the $Q_{features}$ at time t, as well as the historical BEV features B_{t-1} at time $t-1$. By concatenating $Q_{features}$ and B_{t-1}. The attention calculation formula is as follows:

$$Attention\,(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{1}$$

where d_k represents the vector length, and Q, K, V are equal in the self-attention operation, denoted as

$$Q = Q_{features} \oplus B_{t-1} \tag{2}$$

The output of self-attention is combined by addition with the concatenated result of $Q_{features}$ and B_{t-1}. It then goes through a Layer Normalization (LN) layer to obtain Q_{self}. The LN layer normalizes each feature dimension across samples, improving the model's generalization ability and reducing internal covariate shift. By using Q_{self} as the query and feature values in B_{Global} as keys and values, the Multi-Head Cross Attention (MHCA) operation is performed. This operation is utilized to softly associate the LiDAR and image features with query, enabling BEVTemporal to adaptively determine what to choose and where to choose in B_{Global}. The MHCA calculation between Q_{self} and B_{Global} is denoted as (3).

$$\text{MHCA}\left(Q_{self}, K, V\right) = \text{Concat}\left(head_1, \ldots, head_n\right) W^O \tag{3}$$

where W^O are learnable parameters satisfying $W^O \in \mathbb{R}^{d_{model}\times nd_v}$, $head_i$ refers to the attention calculation result of each head.

$$head_i = \text{Attention}\left(Q_{self} W_i^{Q_{self}}, KW_i^K, VW_i^V\right),\ i\epsilon\,(1,n) \tag{4}$$

In (4), K, V refer to B_{Global}. $W_i^{Q_{self}} \in \mathbb{R}^{d_{model}\times d_k}$, $W_i^K \in \mathbb{R}^{d_{model}\times d_k}$, $W_i^V \in \mathbb{R}^{d_{model}\times d_v}$ are the different parameter matrices corresponding to Q_{self}, K, V, respectively. d_k, d_v, and d_{model} satisfy the following relationship:

$$d_k = d_v = \frac{d_{model}}{n} \tag{5}$$

After normalization, the result of multi-head cross attention needs to be retained as B_t for temporal attention calculation with the corresponding $Q_{features}$ at time $t+1$. At the same time, B_t is input to a Feed Forward Network (FFN) layer. The FFN layer comprehensively integrates LiDAR point cloud and image information to generate the final prediction bounding boxes, which are further used for subsequent loss calculation and parameter updating.

For data at $t = 0$, since there are no historical BEV features, the temporal self-attention operation degenerates to self-attention of $Q_{features}$. The remaining parts of the temporal module remains. Q in (1) needs to be modified to

$$Q = Q_{features} \oplus Q_{features} \tag{6}$$

This module can help BEVTemporal complete perception tasks based on current and historical data. By extracting temporal information from previous features, it comprehensively considers the influence of historical information on the current environment and the dynamic characteristics of traffic object over time. This enables BEVTemporal to learn and analyze feature information over a period of time, incorporating temporal context for more comprehensive perception. Cv x

IV. EXPERIMENTAL RESULTS ANALYSIS

A. EXPERIMENTAL DATASET AND EVALUATION METRICS

The experimental data come from the nuScenes dataset, which contains a large number of outdoor road scenes to support detection, tracking, planning and other tasks for autonomous driving systems. The data collection vehicle was equipped with a 32-line LiDAR on the roof and five radars and six monocular cameras around the vehicle. The annotated categories include cars, trucks, motorcycles, barriers etc. The amount of annotated data is approximately 1.4 million [35], which is more than seven times the size of the KITTI [36] dataset. For the 3D object detection task provided by this dataset, the accuracy metrics include AP, mAP and NDS. The mAP calculation formula is

$$\text{mAP} = \frac{1}{|\mathbb{C}|\,|\mathbb{D}|} \sum_{c\epsilon\mathbb{C}} \sum_{d\epsilon\mathbb{D}} \text{AP}_{c,d} \tag{7}$$

AP is Average Precision, $\mathbb{D} = \{0.5, 1, 2, 4\}$ meter, $\mathbb{C}$ represents categories. This metric uses the center point distance between the detected bounding box and the ground truth in the BEV plane instead of the traditional 3D Intersection of Union (IoU). It decouples the detection metrics from the objects' size and orientation, avoiding the issue where IoU scores are 0 for smaller objects [35]. The NDS is calculated as

$$\text{NDS} = \frac{1}{10}\left[5\text{mAP} + \sum_{\text{mTP}\epsilon\mathbb{TP}} (1 - \min(1, \text{mTP}))\right] \tag{8}$$

where mTP is mean True Positive, which is obtained by comprehensively calculating metrics including ATE (Average Translation Error), ASE (Average Scale Error), AOE (Average Orientation Error), AVE (Average Velocity Error), AEE(Average Attribute Error) [35], $\mathbb{TP}$ is the set of mTP. The NDS calculation is based on the results of quantifying the accuracy of position, orientation, speed and other attributes of the detected bounding boxes, in addition to the mAP.

In the following sections, we will design experiments and analyze the results based on the evaluation metrics mentioned above. Specifically, we will focus on two aspects: (1) validation of temporal module effectiveness; (2) advantages of multi-sensor fusion. Through these experiments, we aim to demonstrate the impact and benefits of the temporal module in improving perception tasks, as well as the advantages of integrating information from multiple sensor.

B. VALIDATION OF MODULE EFFECTIVENESS

To validate the effectiveness of the temporal module in BEVTemporal, we will conduct experimental analyses from three perspectives:

- 3D multi-object detection results
- Missed-detection rates on samples with different occlusion levels
- Detection accuracy under various weather and lighting conditions

1) 3D MULTI-OBJECT DETECTION EXPERIMENTS

In this section, we perform 3D object detection tasks using the nuScenes dataset, which includes LiDAR point cloud and RGB data from a set of six monocular cameras providing a panoramic view. The detection objects in road scenes are divided into two major categories: dynamic objects and static objects. The dynamic objects include four classes: cars, trucks, buses, and motorcycles, while the static objects consist of two classes: traffic cone and barriers. We train and validate BEVTemporal using samples from each class in the dataset. The experimental results are visualized in Fig. 4.

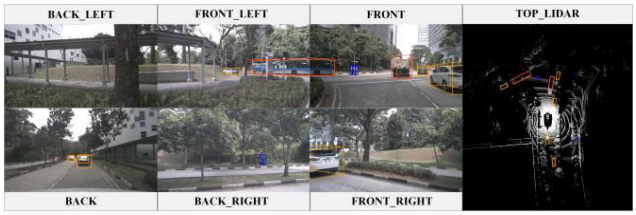

FIGURE 4. BEVTemporal 3D object detection visualization results.

According to (6), the temporal module can be degraded to a self-attention module without considering temporal information. Table 1 shows the comparison results between the model without temporal information and BEVTemporal.

As shown in Table 1, the AP for all object categories has improved to the model without temporal information. Notably, the average precision of the two static object classes is improved more significantly. This is possibly because static objects tend to remain relatively stationary in the driving environment, except for the relative displacement caused by the ego vehicle's motion. Therefore, these objects do not introduce additional deviations. And static objects like traffic cone and barriers often appear in multiple consecutive frames with similar appearances in road scenes. As a result, the network can better learn the features by leveraging the temporal information from their regular movements. On the other hand, although the motion of dynamic objects is less predictable, the model can still benefit from learning the temporal context to better understand their dynamic nature in traffic scenes. As conclusion, after adding the temporal module, the AP of each category in the BEVTemporal model has increased by approximately 0.3% to 1.7%, and the mAP has increased by 0.87%.

TABLE 1. Multi-object detection results comparison of different networks.

Method	Category (AP/%)						mAP/%	NDS/%
	Car	Truck	Bus	Motor.	Traffic Cone	Barrier		
W. T. I.	86.9	57.4	71.8	67.2	63.4	67.2	68.98	74.84
Ours	**87.2**(+0.3)	**57.7**(+0.3)	**72.2**(+0.4)	**68.9**(+1.7)	**64.8**(+1.4)	**68.3**(+1.1)	**69.85**(+0.87)	**75.29**(+0.45)

Notion of class: without temporal information (W. T. I.).

To validate the experimental analysis above and further understand how the temporal module in BEVTemporal helps the network learn image features, we visualize the heatmaps of the feature extraction network outputs for both the model without temporal information and the proposed model. We mainly focus our analysis on motorcycle, transport facilities and barrier which have shown better performance. The visualization results are shown in Fig. 5.

Fig. 5 shows that the model without temporal information tends to focus more on scattered and redundant information, such as road surface details. On the other hand, the attention of the BEVTemporal model is more directed towards dynamic objects, especially those that appear repeatedly in the scene. Moreover, BEVTemporal even pays attention to traffic signs that were not part of the training data, as shown in the box in Fig. 6. In addition, Fig. 6 also indicates that BEVTemporal also pays high attention to the traffic cone sample with significant accuracy improvement among the static samples.

By comparing the motorcycle samples in Fig. 5(a) and Fig. 5(b), we can observe that the proposed model is able to focus on the details of the motorcycle, such as the wheels and footboard. The overall heat of the motorcycle is higher than the surrounding environment. In contrast, the model without temporal information mainly focuses on the motorcycle rider, neglecting some of the motorcycle's body details. For the transport facilities in Fig. 5(c) and Fig. 5(d), the attention of the BEVTemporal model follows the shape distribution of the railing, allowing it to better focus on the contour information and regular lines in the background. This suggests that the temporal module enables the network to learn more effectively from samples with similar shapes that appear repeatedly in the scene. Fig. 5(e) and Fig. 5(f) show that without temporal information, the model tends to focus more on the near-ground region and is easy to ignore distant samples, while BEVTemporal is able to pay attention to both nearby barriers and distant dynamic objects like vehicles.

Therefore, it can be concluded from the visualization results of the feature extraction network that after adding temporal information, BEVTemporal can effectively filter out redundant information and allows the model to focus more on the global and contour features of dynamic objects. The memory capabilities of the model also enable it focus more on samples that appear repeatedly with regular shapes in images, preventing the neglect of distant samples. The temporal module helps the network learn the dynamic properties of traffic objects that appear in a regular pattern. As a result, BEVTemporal exhibits improved detection accuracy for various dynamic objects, with more significant improvements observed for samples with regular patterns.

2) EXPERIMENTS WITH DIFFERENT OCCLUSION LEVELS

Observing the visualization results on the test dataset, the model can detect some heavily occluded objects after adding temporal information, thus improving the problem of missed detections, as shown in Fig. 7.

It is evident from Fig. 7 that within the blue box in the original image (Fig. 7(a)), there are two cars driving side by side. Due to the similarity in color and significant occlusion, it is even difficult for the human eye to accurately determine the number of cars within the blue box based solely on the current frame. In the absence of temporal information, relying solely on the instantaneous data, the model's detection result is shown in Fig. 7(b), where a missed detection occurs in the red dashed box. However, when the temporal information module is incorporated, as shown in Fig. 7(c), the model can accurately detect all the samples within the red dashed box. The inclusion of temporal information helps the model better understand the dynamic characteristics of objects in the traffic scene, which assists the model in detecting objects based on historical data and reduces the probability of missed detections.

To further validate the effectiveness of the temporal module in detecting objects with different occlusion levels, the validation set are divided into three subsets based on the visibility range: 0%-40%, 40%-80%, and 80%-100% according to nuScenes dataset labels. The recall rates of each category in the subsets were calculated, and the experimental results are shown in Fig. 8. Fig. 8 illustrates that with occlusion among samples increases, detection tasks becomes more difficult, leading to a decrease in the recall rates of both models. Simultaneously, the gap between the recall of the two models widens, indicating a more significant reduction in missed detections when the temporal module is incorporated. Therefore, for samples with severe occlusion, incorporating temporal information in the model enables better feature learning, effectively reduce the probability of missed detections.

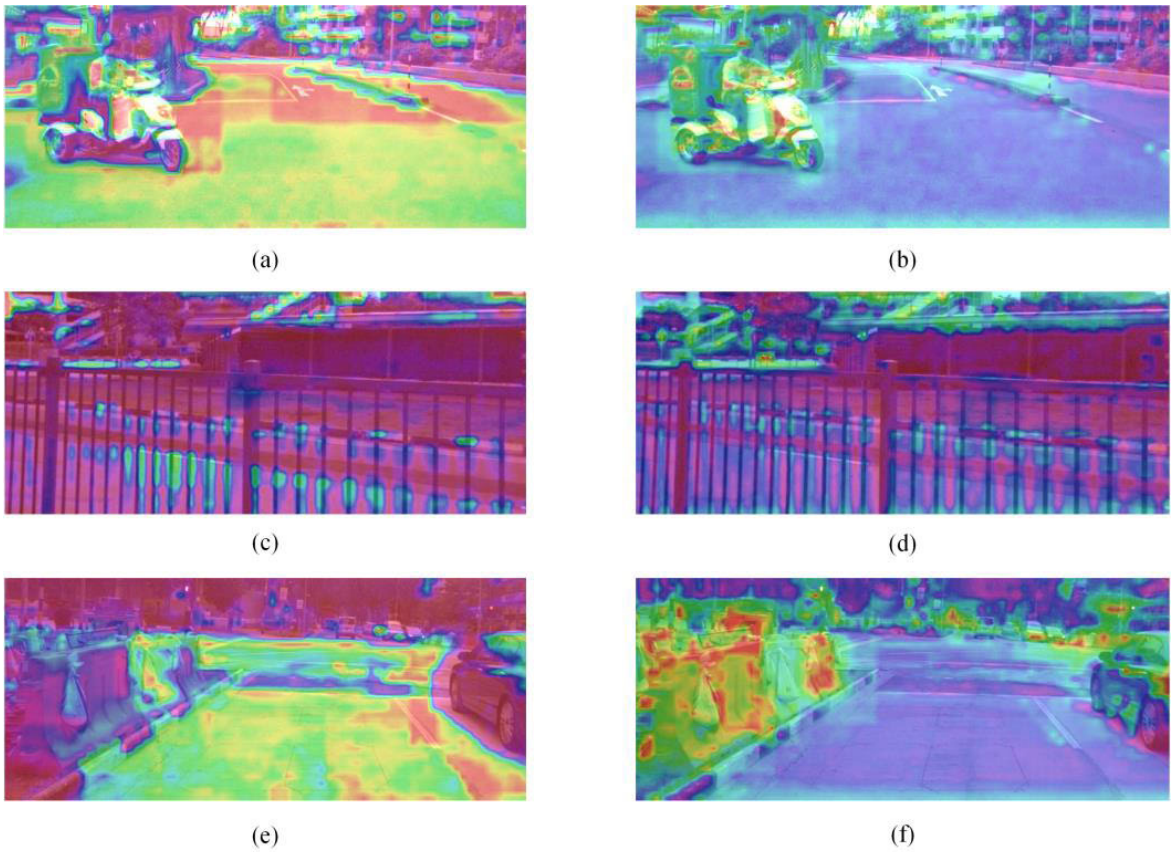

FIGURE 5. Heatmap of feature extraction network: (a) model without temporal information(motorcycle), (b) BEVTemporal(motorcycle), (c) model without temporal information (transport facilities), (d) BEVTemporal(transport facilities), (e) model without temporal information(barrier) and (f)) BEVTemporal(barrier).

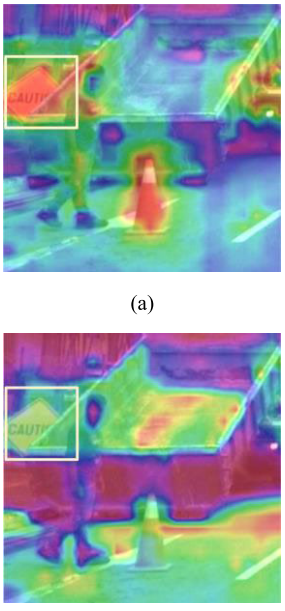

FIGURE 6. Heatmap of traffic sign: (a) model without temporal information and (b) BEVTemporal.

3) EXPERIMENTS UNDER DIFFERENT WATHER AND LIGHTING CONDITIONS

To validate the effectiveness of the temporal module under different weather and lighting conditions, we divided the validation set into two subsets based on light: day and night, as well as two subsets based on weather conditions: rainy and sunny. A pairwise comparison was conducted for each class of target categories within these scenes to evaluate the accuracy improvement of the enhanced model. The experimental results are shown in Table 2.

Table 2 demonstrates that the model incorporating temporal information outperforms the model without temporal information in most target categories across different scenes. The improvement in AP can reach a maximum of 3.6%. Furthermore, the mAP shows an overall increase in all scenes. Notably, in challenging scenarios such as rainy and night scenes, multiple target categories exhibit an improvement of over 2% in AP. Even in the presence of rain, which typically poses significant challenges to detection accuracy, the mAP can be enhanced by approximately 1.2%.

The results demonstrate that the robustness of the BEVTemporal model, enabling its adaptability to adverse

TABLE 2. Detection results comparison of different networks in different scenarios.

Category (AP/%)	Scenario							
	Day		Night		Sunny		Rainy	
	Without T.M.	Ours	Without T.M.	Ours	Without T.M.	Ours	Without T.M.	Ours
Car	86.9	**87.2**(+0.3)	88.0	**88.4**(+0.4)	86.5	**86.8**(+0.3)	88.3	**88.6**(+0.3)
Truck	57.1	**57.4**(+0.3)	**80.5**	80.3(-0.2)	**56.3**	56.0(-0.3)	59.5	**61.7**(+2.2)
Bus	71.9	**72.3**(+0.4)	0	0	70.8	**71.1**(+0.3)	80.1	**80.6**(+0.5)
Motorcycle	65.8	**67.4**(+1.6)	77.8	**79.8**(+2.0)	66.3	**67.6**(+1.3)	66.4	**70.0**(+3.6)
T.C.	63.9	**65.2**(+1.3)	0	0	63.4	**64.7**(+1.3)	66.2	**68.2**(+2)
Barrier	67.5	**68.6**(+1.1)	35.9	**38.2**(+2.3)	63.7	**65.4**(+1.7)	**81.4**	80.2(-1.2)
mAP/%	68.85	**69.68**(+0.83)	47.03	**47.78**(+0.75)	67.83	**68.60**(+0.73)	73.65	**74.83**(+1.18)

Notion of class: temporal module (T.M.), motorcycle (Motor.), traffic cone (T.C.).

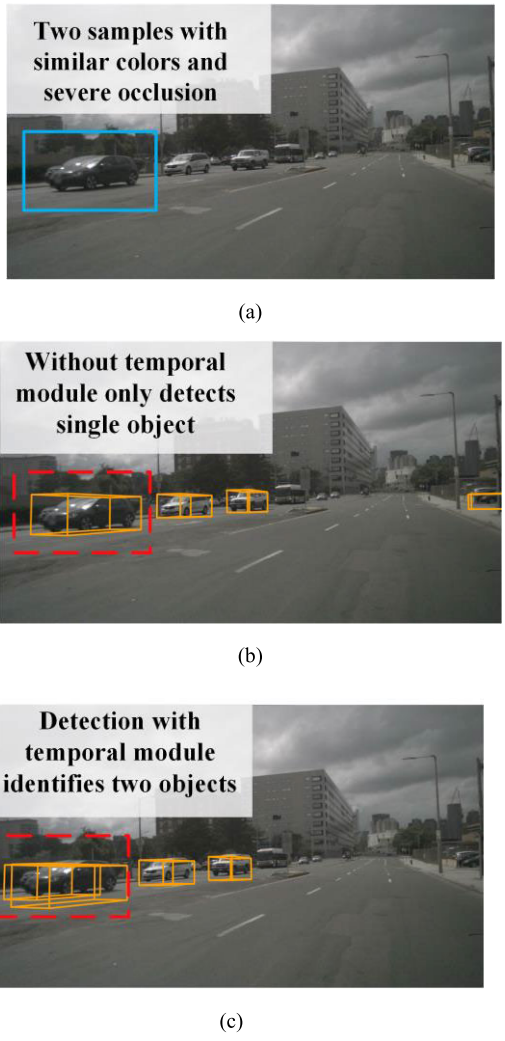

FIGURE 7. Detection of severely occluded samples: (a)original image, (b)without temporal module, and (c) with temporal module.

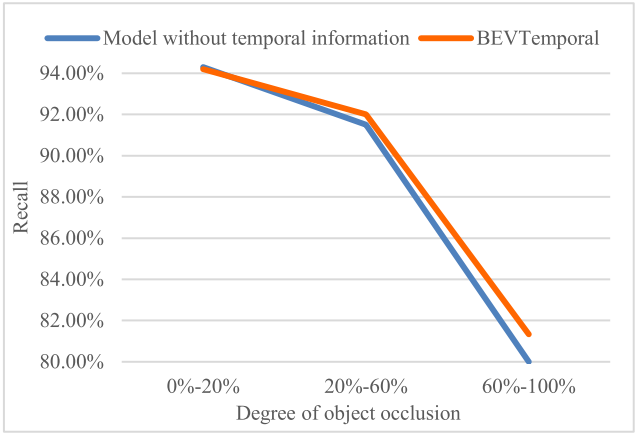

FIGURE 8. Recall of samples with different occlusion degree. Recall refers to the proportion of true positive samples among all positive samples. The higher the recall, the lower the missed detection rate.

lighting conditions and challenging autonomous driving scenarios.

C. COMPARISON OF DIFFERENT MODALITY MODELS

In addition to adding temporal information, BEVTemporal considers using multi-sensor data to accomplish the 3D object detection task. By leveraging the complementary information from different sensor types and perspectives, the model achieves accuracy improvement.

To validate the effectiveness of multi-sensor fusion detection, we select several 3D object detection networks based on different sensors and evaluate their performance on the nuScenes dataset. Then we conduct a comparative analysis with the model proposed in this paper. The results are shown in Table 3.

The experimental results in Table 3 indicate that the 3D object detection schemes based solely on laser point clouds, as well as the schemes that incorporate multi-sensor fusion, exhibit significant advantages over camera-only approaches in terms of mAP for various object categories. These schemes enable more accurate localization and classification tasks by leveraging the inherent 3D spatial information provided by point clouds. Moreover, the multi-sensor fusion networks, built upon precise spatial information from point clouds, incorporate rich semantic information from images. As a result, they demonstrate improvements in mAP metrics and AP metrics for different object categories compared to the LiDAR-only frameworks.

The series of experimental results mentioned above indicate that the model incorporating temporal information can

TABLE 3. Detections results comparison of different networks based on different sensors.

Method	Modality	Category (AP/%)						mAP/%
		Car	Truck	Bus	Motor.	T. C.	Barrier	
BEVDet[17]	Camera	64.3	35.0	35.8	44.8	60.1	61.4	50.23
BEVFormer[18]	Camera	67.7	39.2	35.7	47.9	70.3	62.5	53.83
PointPillar[28]	LiDAR	76.0	31.0	32.1	34.2	62.4	56.4	48.68
Centerpoint[38]	LiDAR	85.2	53.5	63.6	59.5	**78.4**	60.2	66.73
PointPainting[39]	Camera + LiDAR	82.6	43.6	41.6	55.9	70.8	60.2	59.12
BEVFusion[32]	Camera + LiDAR	86.9	57.4	71.8	67.2	63.4	67.2	68.98
BEVTemporal	Camera + LiDAR	**87.2**	**57.7**	**72.2**	**68.9**	64.8	**68.3**	**69.85**

Notion of class: motorcycle (Motor.), traffic cone (T.C.).

pay more attention to the global and contour features of the targets. Its memory capability allows it to focus more on samples that appear repeatedly and follow certain shape patterns, making it less likely to overlook distant objects. Therefore, compared to models without temporal information, BEVTemporal demonstrates higher detection accuracy. It also exhibits better adaptability to weather and lighting variations while reducing the occurrence of missed detections caused by sample occlusions.

V. CONCLUSION

we perform 3D object detection tasks in road traffic scenes based on LiDAR point clouds and RGB images captured by a surround-view camera system. In order to enhance the model's understanding of the temporal dynamics in traffic scenes and improve its applicability for autonomous driving perception of the surrounding environment, we propose incorporating temporal information into the 3D object detection network. By leveraging both historical features and current moment features, the model is able to comprehensively perform the detection task. The conclusions are:

- ➢ After incorporating temporal information into BEVTemporal, the detection accuracy of each traffic target has been improved. The AP metrics for different categories have increased by a range of 0.3% to 1.7%, with an overall mAP improvement of nearly 1%.
- ➢ By associating features from different time steps, the BEVTemporal model with temporal information can alleviate the issue of missed detections caused by occlusion. Moreover, the more severe the sample occlusion, the more significant the advantages of this model become.
- ➢ The temporal information module contributes to improving the robustness of the model. After adding temporal information, the model exhibits mAP improvement ranging from 0.75% to 1.18% across different weather and lighting conditions. In challenging scenarios such as rainy and night scenes, AP for various object categories can be improved by up to 3.6%, enhancing the network's ability to handle different scene variations.
- ➢ By employing a fusion approach that combines LiDAR data with image data, the 3D object detection task benefits from accurate spatial information in point cloud and rich textural details in image data. Compared to detection schemes based on a single sensor, the fusion of multi-sensor information enables a more comprehensive and overall higher detection accuracy.

Based on thses experimental results, we believe that BEVTemporal effectively captures the temporal correlation between features at different time steps through the temporal information module. This enables the model to process time-series data more efficiently and better understand the dynamic characteristics of traffic scenes, leading to improved performance. Furthermore, based on the visualization results, we infer that BEVTemporal utilizes temporal information to filter out transient or false-positive detections, thereby enhancing the detection accuracy of the 3D object detection network.. The temporal module enhances the model's stability in different scenarios and improves its effectiveness in detecting occluded samples by incorporating temporal memory. Besides, BEVTemporal achieves complementary multi-sensor fusion to provide more comprehensive and accurate information, reducing misjudgments and improving accuracy.

LIMITATION

BEVTemporal requires spatial relevance among samples. However, nuScenes dataset's keyframe are collected at 2Hz, leading to a time difference of 0.5 seconds between frames. For those target categories, which may have significant speed differences compared to the ego vehicle, it may lead to spatial errors and reducing the detection accuracy and training stability of the model.

The key to addressing this issue is to increase the data collection frequency and reduce the frame interval. Therefore, in further research, it would be beneficial to validate the approach using a dataset with a higher sampling frequency or design a spatial correlation module within the model to align the spatial information of various objects in the scene and mitigate the impact of spatial errors.

REFERENCES

[1] K. Li, J. Dai, S. Li, and M. Bian, "State-of-the-art and technical trends of intelligent and connected vehicles," *Automot. Saf. Energy*, vol. 8, no. 1, pp. 1–14, 2017.

[2] D. Feng, C. Haase-Schütz, L. Rosenbaum, H. Hertlein, C. Gläser, F. Timm, W. Wiesbeck, and K. Dietmayer, "Deep multi-modal object detection and semantic segmentation for autonomous driving: Datasets, methods, and challenges," *IEEE Trans. Intell. Transp. Syst.*, vol. 22, no. 3, pp. 1341–1360, Mar. 2021.

[3] X. Sun, S. Wang, M. Wang, Z. Wang, and M. Liu, "A novel coding architecture for LiDAR point cloud sequence," *IEEE Robot. Autom. Lett.*, vol. 5, no. 4, pp. 5637–5644, Oct. 2020.

[4] X. Sun, M. Wang, J. Du, Y. Sun, S. S. Cheng, and W. Xie, "A task-driven scene-aware LiDAR point cloud coding framework for autonomous vehicles," *IEEE Trans. Ind. Informat.*, vol. 19, no. 8, pp. 8731–8742, Jun. 2022.

[5] T.-Y. Chen, C.-C. Hsiao, and C.-C. Huang, "Density-imbalance-eased LiDAR point cloud upsampling via feature consistency learning," *IEEE Trans. Intell. Vehicles*, vol. 8, no. 4, pp. 2875–2887, Jun. 2022.

[6] Y. Wang and Y. Tian, "Study of the hazard perception model for automated driving systems," in *Proc. Int. Conf. Hum.-Comput. Interact.* Springer, 2022, pp. 435–447.

[7] Y. Xu and K. Miyahara, "Object recognition system under hazy condition for automated driving systems," in *Proc. IEEE Int. Conf. Cybern. Intell. Syst. (CIS) IEEE Conf. Robot., Autom. Mechatronics (RAM)*, Jun. 2023, pp. 25–29.

[8] C. Wang, D. Chu, M. Martens, M. KrüGer, and T. H. Weisswange, "Hybrid eyes: Design and evaluation of the prediction-level cooperative driving with a real-world automated driving system," in *Proc. 14th Int. Conf. Automot. User Interfaces Interact. Veh. Appl.*, Sep. 2022, pp. 274–284.

[9] C. Yan and E. Salman, "Mono3D: Open source cell library for monolithic 3-D integrated circuits," *IEEE Trans. Circuits Syst. I, Reg. Papers*, vol. 65, no. 3, pp. 1075–1085, Mar. 2018.

[10] T. Wang, X. Zhu, J. Pang, and D. Lin, "FCOS3D: Fully convolutional one-stage monocular 3D object detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 913–922.

[11] B. Xu and Z. Chen, "Multi-level fusion based 3D object detection from monocular images," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 2345–2353.

[12] X. Chen, K. Kundu, Y. Zhu, H. Ma, S. Fidler, and R. Urtasun, "3D object proposals using stereo imagery for accurate object class detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 5, pp. 1259–1272, May 2018.

[13] P. Li, X. Chen, and S. Shen, "Stereo R-CNN based 3D object detection for autonomous driving," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 7636–7644.

[14] Y. Wang, V. C. Guizilini, T. Zhang, Y. Wang, H. Zhao, and J. Solomon, "DETR3D: 3D object detection from multi-view images via 3D-to-2D queries," in *Proc. Conf. Robot Learn.*, 2022, pp. 180–191.

[15] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2020, pp. 213–229.

[16] J. Huang, G. Huang, Z. Zhu, Y. Ye, and D. Du, "BEVDet: High-performance multi-camera 3D object detection in bird-eye-view," 2021, *arXiv:2112.11790*.

[17] Z. Li, "Bevformer: Learning bird's-eye-view representation from multi-camera images via spatiotemporal transformers," in *Proc. 17th Eur. Conf. Comput. Vis. (ECCV)*, 2022, pp. 1–18.

[18] J. Philion and S. Fidler, "Lift, splat, shoot: Encoding images from arbitrary camera rigs by implicitly unprojecting to 3D," in *Proc. 16th Eur. Conf. Comput. Vis. (ECCV)*, Glasgow, U.K. Springer, 2020, pp. 194–210.

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

[20] C. R. Qi, H. Su, K. Mo, and L. J. Guibas, "PointNet: Deep learning on point sets for 3D classification and segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 652–660.

[21] C. R. Qi, L. Yi, H. Su, and L. J. Guibas, "PointNet++: Deep hierarchical feature learning on point sets in a metric space," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 1–14.

[22] Y. Yan, Y. Mao, and B. Li, "SECOND: Sparsely embedded convolutional detection," *Sensors*, vol. 18, no. 10, p. 3337, Oct. 2018.

[23] X. Chen, H. Ma, J. Wan, B. Li, and T. Xia, "Multi-view 3D object detection network for autonomous driving," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 1907–1915.

[24] M. Liang, B. Yang, S. Wang, and R. Urtasun, "Deep continuous fusion for multi-sensor 3D object detection," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 641–656.

[25] X. Wu, L. Peng, H. Yang, L. Xie, C. Huang, C. Deng, H. Liu, and D. Cai, "Sparse fuse dense: Towards high quality 3D detection with depth completion," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 5408–5417.

[26] X. Bai, Z. Hu, X. Zhu, Q. Huang, Y. Chen, H. Fu, and C.-L. Tai, "TransFusion: Robust LiDAR-camera fusion for 3D object detection with transformers," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 1080–1089.

[27] T. Liang, "BEVFusion: A simple and robust LiDAR-camera fusion framework," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 35, 2022, pp. 10421–10434.

[28] 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.*, 2019, pp. 12697–12705.

[29] Y. Zhou and O. Tuzel, "VoxelNet: End-to-end learning for point cloud based 3D object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 4490–4499.

[30] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 9992–10002.

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

[32] Z. Liu, H. Tang, A. Amini, X. Yang, H. Mao, D. L. Rus, and S. Han, "BEVFusion: Multi-task multi-sensor fusion with unified bird's-eye view representation," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, May 2023, pp. 2774–2781.

[33] L. Huang, Z. Cui, F. You, P. Hong, H. Zhong, and Y. Zeng, "Vehicle trajectory prediction model for multi-vehicle interaction scenario," *J. Jilin Univ. Eng. Technol. Ed.*, pp. 1–10.

[34] H. Law and J. Deng, "CornerNet: Detecting objects as paired keypoints," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 734–750.

[35] H. Caesar, V. Bankiti, A. H. Lang, S. Vora, V. E. Liong, Q. Xu, A. Krishnan, Y. Pan, G. Baldan, and O. Beijbom, "NuScenes: A multimodal dataset for autonomous driving," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 11618–11628.

[36] A. Geiger, P. Lenz, and R. Urtasun, "Are we ready for autonomous driving? The KITTI vision benchmark suite," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2012, pp. 3354–3361.

[37] L. Huang, J. Wu, R. Zhang, D. Zhao, and Y. Wang, "Comparative analysis & modelling for Riders' conflict avoidance behavior of E-bikes and bicycles at un-signalized intersections," *IEEE Intell. Transp. Syst. Mag.*, vol. 13, no. 4, pp. 131–145, May 2021.

[38] 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. 11779–11788.

[39] 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. 4603–4611.

LING HUANG received the Ph.D. degree in systems engineering from Beijing Jiaotong University, in 2007. From 2017 to 2018, she conducted her visiting scholar program with the University of Washington, USA. She is currently an Associate Professor with the School of Civil and Transportation Engineering, South China University of Technology. She has led or participated in more than ten national and provincial-level projects, and more than 20 government and industry applied projects. She has published more than 50 high-quality academic papers in transportation-related journals and conferences. Her main research interests include intelligent transportation systems, traffic video analysis, driver behavior analysis and modeling, and microscopic traffic simulation. She has received three provincial-level awards, including the First Prize for Natural Science from China Simulation Society. She has authored one Chinese monograph, which was selected for the 2018 Major Technological Innovations in Transportation, and one English monograph. She has also served as a Committee Member of the 2nd Traffic Modeling and Simulation Technology Committee of China Simulation Society.

YIXUAN ZENG received the bachelor's degree in intelligent science and technology from the South China University of Technology, Guangdong, China, in 2021, where she is currently pursuing the master's degree in intelligent transportation systems. Supervised by L. Huang, she focuses on deep learning and computer vision.

SHUO WANG received the B.Sc. (Eng.) degree from the Department of Civil Engineering, China University of Mining and Technology, Xuzhou, China, in 1998, and the M.Arch. and Ph.D. (Eng.) degrees in architecture from the South China University of Technology, Guangzhou, China, in 2003 and 2014, respectively. He is currently a Lecturer with the School of Architecture, South China University of Technology. His current research interests include computer aided architecture design, and application of virtual reality.

RUNMIN WEN received the B.S. degree in information engineering from the School of Electronic and Information Engineering, South China University of Technology, in 2022. He is currently pursuing the master's degree in intelligent transportation system with the South China University of Technology. Supervised by L. Huang, he focuses on deep learning and computer vision.

XINGYU HUANG received the degree from the South China University of Technology, in 2023, where he is currently pursuing the master's degree in intelligent transportation system. Supervised by L. Huang, he focuses on deep learning and computer vision in autonomous vehicle.