Computers, Materials & Continua

Tech Science Press

Doi:10.32604/cmc.2025.063205

REVIEW

Research Progress on Multi-Modal Fusion Object Detection Algorithms for Autonomous Driving: A Review

Peicheng Shi[1,*], Li Yang[1], Xinlong Dong[1], Heng Qi[2] and Aixi Yang[3]

[1]School of Mechanical and Automotive Engineering, Anhui Polytechnic University, Wuhu, 241000, China
[2]State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing, Wuhan University, Wuhan, 430072, China
[3]Polytechnic Institute, Zhejiang University, Hangzhou, 310015, China

*Corresponding Author: Peicheng Shi. Email: shipeicheng@126.com

Received: 08 January 2025; Accepted: 11 March 2025; Published: 19 May 2025

ABSTRACT: As the number and complexity of sensors in autonomous vehicles continue to rise, multimodal fusion-based object detection algorithms are increasingly being used to detect 3D environmental information, significantly advancing the development of perception technology in autonomous driving. To further promote the development of fusion algorithms and improve detection performance, this paper discusses the advantages and recent advancements of multimodal fusion-based object detection algorithms. Starting from single-modal sensor detection, the paper provides a detailed overview of typical sensors used in autonomous driving and introduces object detection methods based on images and point clouds. For image-based detection methods, they are categorized into monocular detection and binocular detection based on different input types. For point cloud-based detection methods, they are classified into projection-based, voxel-based, point cluster-based, pillar-based, and graph structure-based approaches based on the technical pathways for processing point cloud features. Additionally, multimodal fusion algorithms are divided into Camera-LiDAR fusion, Camera-Radar fusion, Camera-LiDAR-Radar fusion, and other sensor fusion methods based on the types of sensors involved. Furthermore, the paper identifies five key future research directions in this field, aiming to provide insights for researchers engaged in multimodal fusion-based object detection algorithms and to encourage broader attention to the research and application of multimodal fusion-based object detection.

KEYWORDS: Multi-modal fusion; 3D object detection; deep learning; autonomous driving

1 Introduction

Environmental perception technology [1] is the core technology of autonomous driving systems, which is responsible for helping vehicles identify and locate surrounding obstacles, and is essential for applications such as autonomous vehicles, robotics, and intelligent traffic management. Autonomous driving technology has made significant progress in recent years, but deploying reliable autonomous driving systems in the real world remains challenging. Autonomous vehicles [2] must perform tasks such as perception, prediction, decision-making, planning, and execution in a complex and uncontrolled environment, and any small mistake can lead to serious consequences. Fig. 1 shows the system architecture for autonomous driving.

Challenges for autonomous driving systems include dealing with occlusion, complex backgrounds, and sensor noise. In order to overcome these challenges, researchers have begun to explore multimodal data fusion strategies [3], combining data sources from different sensors, such as lidar, camera, and radar, to

improve detection performance by mining their complementary characteristics. Lidar can provide high-precision three-dimensional spatial information, but it may face problems of sparse data and lack of semantic features in dense scenes; although camera images are rich in color and texture information, they have limitations in expressing three-dimensional geometric structures. By integrating the spatial information of point clouds and the semantic features of images, it is possible to accurately identify and locate targets in complex environments, significantly improving detection accuracy and robustness. However, as the number of sensors installed in autonomous vehicles increases, problems arise one after another, such as differences in data distribution between different modalities, and alignment errors in time and space. How to efficiently extract multimodal features, achieve accurate modal alignment, and dynamically adjust the differences in contributions of different modalities to tasks are still key issues that urgently need to be addressed.

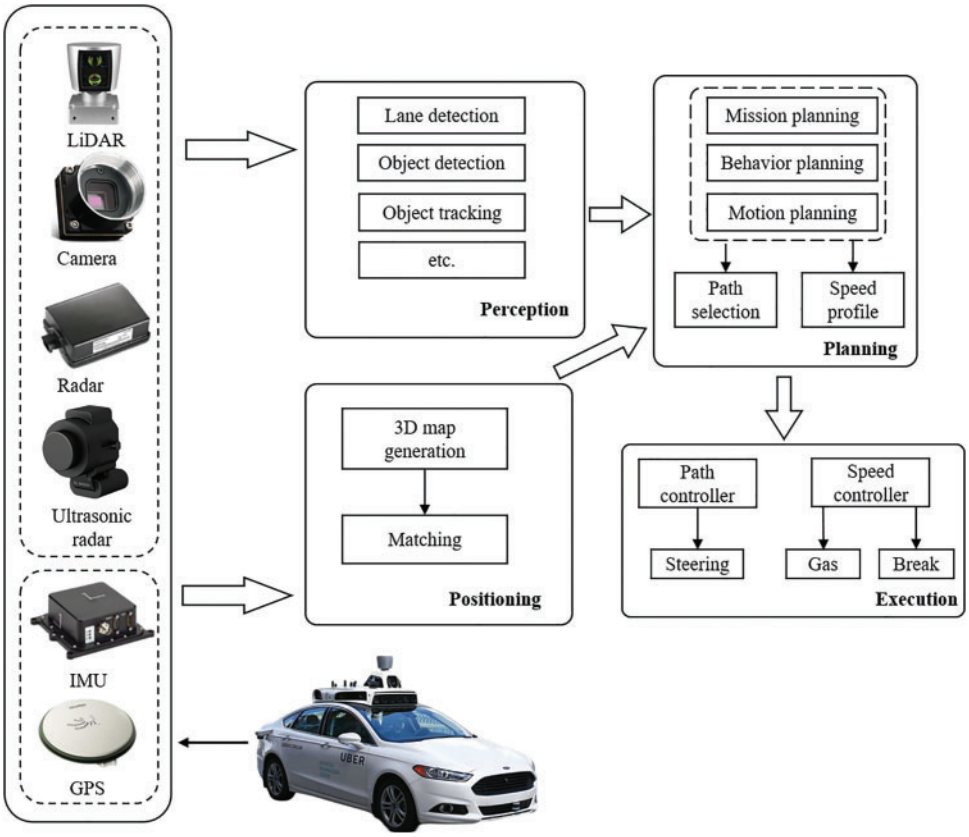

Figure 1: Auto drive system architecture, composed of four subsystems: Perception; Positioning; Planning and Execution

Although the existing reviews comprehensively summarize some key technologies in the field of autonomous driving (such as object detection, semantic segmentation, sensor fusion, etc.), there are still some obvious limitations, such as:

(1) Narrow technical coverage: Some reviews only focus on a single technology (such as convolutional neural network (CNN) or Transformer) and lack a systematic summary of multimodal fusion methods. For example, the application of CNNs in image perception is discussed in detail in Ref. [4], but does not address the processing methods of lidar or radar data. Ref. [5] focuses on the advantages of Transformer in sequence modeling, while ignoring its potential for fusion with other modalities such as images or point clouds. The

limitations of this single technology make it difficult for the existing reviews to fully reflect the research status of multimodal perception of autonomous driving.

(2) Failure to cover the latest technological advances: Although some reviews cover the basic methods of multimodal fusion, they fail to incorporate the latest technological advances in a timely manner. For example, perception algorithms based on self-supervised learning have made significant progress in recent years, which can improve model performance by using unlabeled data, but the existing reviews [6] are still mainly limited to supervised learning methods. In addition, some emerging technologies, such as multimodal fusion methods based on graph neural networks, have not been fully discussed.

(3) Lack of in-depth analysis of practical application scenarios: Existing reviews often focus on the theoretical description of algorithms, but lack of in-depth analysis of practical application scenarios. For example, the performance of multimodal perception algorithms often decreases significantly in extreme weather (e.g., rain, snow, haze) or complex traffic scenarios (e.g., urban congestion, highways), but existing reviews fail to systematically summarize these challenges and their solutions.

These shortcomings provide important research directions for this review. In this paper, we will review the latest research progress of object detection algorithms based on multimodal fusion, and conduct in-depth analysis of different solutions and algorithms. We will explore how these advanced algorithms and technologies can solve the technical challenges of multimodal data fusion and improve the detection performance of autonomous driving systems in complex environments. By summarizing and analyzing these research results, we can better understand the potential and prospect of multimodal data fusion in the field of autonomous driving, and help us build a safer and more reliable autonomous driving system. The main contributions of this paper are as follows:

(1) This paper systematically summarizes the application of multimodal fusion methods in autonomous driving, which fills the gap in existing research. Focusing on three-dimensional object detection in autonomous driving scenarios, this paper comprehensively and systematically reviews the research in the field of multimodal fusion object detection. Compared with the existing review, we not only cover the classical methods from the early days to the latest cutting-edge technologies, but also pay special attention to the multimodal fusion methods based on deep learning and Transformer architectures in recent years. This paper comprehensively discusses the latest representative work of 3D object detection based on multi-modal fusion, introduces relevant datasets and evaluation indicators, summarizes the background technology of multi-modal fusion, including the characteristics of sensors in autonomous driving and the advantages and disadvantages of single-modal sensor object detection, and deeply analyzes the technical paths of different algorithms.

(2) The potential of the latest technologies in autonomous driving is explored in depth. In this paper, the multimodal fusion object detection methods are classified and analyzed from multiple dimensions, including data types (such as images, point clouds, radars, etc.), fusion strategies (such as data-level fusion, feature-level fusion, decision-level fusion, etc.), application scenarios (such as autonomous driving, robot navigation, etc.), and model architecture (such as Transformer, CNN, etc.). The three core issues in multimodal fusion are discussed in detail: what to fuse, when to fuse, and how to fuse, and the processing methods of the original data are reasonably classified. The technical methods of multimodal fusion are comprehensively discussed, including camera-lidar fusion, camera-millimeter-wave radar fusion, camera-lidar-millimeter-wave radar fusion, and other sensor fusion.

(3) It condenses the future research direction of multimodal fusion algorithm, and provides new ideas for technological progress in the field of autonomous driving. In this paper, we pay special attention to the latest research progress in this field in recent years, including some emerging technologies that have not yet been widely used, analyze the innovation and performance improvement of these methods in detail, and put

forward the key directions for future research. Including: (a) Convolutional operators suitable for feature extraction of non-Euclidean data; (b) The exploration of semi-supervised learning vs. unsupervised learning; (c) Multimodal target detection in polar coordinates; (d) Adaptive implicit multimodal spatial fusion; (e) Efficient real-time detection.

2 Dataset and Evaluation Indicators

Datasets are typically made up of a variety of scenarios, each with a different number in a different dataset. We have summarized the multiple datasets available for autonomous driving environment perception tasks, including issuance time, region, sensor data, and the number of annotations for different perception tasks (object detection; lane line segmentation; and semantic segmentation). Among them, four large-scale autonomous driving benchmark datasets, KITTI [7], NuScenes [8], Waymo Open Dataset (WOD) [9], and ONCE [10] datasets, are introduced in detail, which have become the focus of this paper due to their importance and wide application in autonomous driving research. There are also smaller-scale datasets that focus on specific individual perception tasks, which are not covered in this article. For example, OpenLane [11] is created for a single task such as lane markings; DeepAccident [12] is a set of accidents dedicated to the safety of autonomous driving research.

2.1 Datasets

To reduce development costs and ensure the safety of experiments, researchers often use open-source in-vehicle datasets to train and validate the detection algorithms they build. High-quality datasets not only enable the training of fast and accurate detection algorithms, but also provide a fair algorithm evaluation platform and benchmark, which helps researchers make horizontal and vertical comparisons to develop better models. Table 1 summarizes the data of the current mainstream autonomous driving perception datasets.

Table 1: Autonomous driving perception dataset

Dataset	Year	Area	Scenario	Time	Point Clouds	Image	Frames	3D box
KITTI [7]	2012	EU	22	1.5	15 K	15 K	15 K	80 K
NuScenes [8]	2019	NA/AS	1000	5.5	390 K	1.4 M	40 K	1.4 M
Waymo [9]	2019	NA	1150	6.4	230 K	12 M	230 K	12 M
ONCE [10]	2021	AS	†	144	1M	7 M	15 K	417 K
DeepAccident [12]	2022	Sim	464	†	131K	786 K	131 K	1.8 M
ApolloScape [13]	2018	AS	103	2.5	29 K	144 K	144 K	70 K
Lyft L5 [14]	2019	AF	366	2.5	46 K	240 K	46 K	1.3 M
A* 3D [15]	2019	AS	†	55	39 K	39 K	39 K	230 K
H3D [16]	2019	NA	160	0.8	27 K	83 K	27 K	1.1 M
A2D2 [17]	2020	EU	†	†	12.5 K	41 K	12.5 K	43 K
Cityscapes 3D [18]	2020	–	†	2.5	–	5 K	5 K	40 K
Argoverse [19]	2019	NA	113	0.6	22 K	490 K	22 K	993 K
AIODrive [20]	2021	Sim	100	2.8	100 K	1 M	100 K	26 M
KITTI-360 [21]	2020	EU	11	†	80 K	320 K	80 K	68 K

Note. “AS” stands for Asia, “EU” stands for Europe, “NA” stands for North America, “Sim” stands for simulated data, and “3D box” represents the number of annotation instances for 3D inspection tasks. “†” means that the statistics are not available, and “–” means that the column does not exist.

KITTI: KITTI dataset [7] is a groundbreaking autonomous driving dataset proposed by Karlsruhe Institute of Technology (KIT) and Toyota Technological Institute Chicago (TTIC) in 2012, and is the first comprehensive dataset for multiple autonomous driving tasks. The collection vehicle of the dataset carries 2 grayscale cameras, 2 color cameras, 1 64-line lidar, 4 optical lenses and a global positioning system (GPS) navigation system to obtain scene data, including 389 stereo image pairs, optical flow maps, 39.2 km visual range sequences, 15,000 point cloud frames and 200,000 artificially labeled 3D target frames. Each frame can contain up to 15 vehicles and 30 pedestrians with varying degrees of occlusion and truncation, and also includes 7481 training images and 7518 test images. The KITTI dataset classifies object detection tasks into three levels based on the size, visibility, and truncation of the target: easy, medium, and difficult. There are two types of evaluation for object detection tasks: 3D Object Detection Evaluation (3D AP) and Bird's-Eye View Detection Evaluation (BEV AP). Compared with other mainstream datasets, KITTI has a small data scale, making it suitable for research scenarios with limited resources, such as algorithm prototyping, rapid verification, lightweight model training, and beginner learning and teaching. Its high-quality 2D/3D annotation data makes it ideal for basic object detection research, especially in object detection tasks with monocular cameras or lidar point clouds.

NuScenes: The NuScenes dataset [8] is a large-scale autonomous driving dataset developed by the Motional team in 2019, containing 1000 driving scenarios in two cities, Boston and Singapore, including complex scenarios such as sunny, rainy, and dark nights. Among them, 850 scenes were used for training and validation, and 150 scenes were used for testing, each with a 20-s time and 40,000 keyframes, totaling approximately 1.4 million camera images, 390,000 lidar scans, 1.3 million radar scans, and 1.4 million 3D object annotations. Compared to other datasets, all of its sensors provide a 360 degree view. The sensors used to collect data include 6 cameras, 1 lidar, and 5 radars, with a camera image resolution of 1600 × 900. At the same time, it also has corresponding HD-Map and controller area network (CAN) bus data to explore multi task perception with multiple inputs. The NuScenes dataset is currently the most widely used dataset for 3D object detection tasks. NuScenes has a medium data scale (about 40,000 frames) and provides multi-sensor data (camera, lidar, radar) and timing information, making it suitable for complex tasks such as multi-modal fusion and time-series object detection and tracking. Its diverse scenarios and rich annotations make it the first choice for medium-scale deep learning model training and multi-task learning.

Waymo: The Waymo Dataset [9] is a dataset released by Google's Waymo autonomous driving company in 2020, and the dataset tasks are divided into three categories: 2D object detection, 3D object detection, and target tracking. It collects data from multiple cities like San Francisco, Phoenix, etc. Includes different scenarios in various driving conditions, such as day, night, dawn, dusk, and rain. The dataset consists of 1150 scenarios, including 798 scenarios in the training set, 202 scenarios in the validation set, and 150 scenarios in the test set. Each scene has 20 s and a total of 230,000 frames of data, of which 12 million objects are manually labeled. The acquisition equipment in the scene is 5 high-resolution cameras and 5 high-quality lidar sensors, with an image resolution of 1920 × 1280 or 1920 × 886 pixels. The Waymo dataset has a large data scale (about 1950 scenes, 20 s per scene), providing high-resolution lidar and multi-camera data, which is suitable for high-precision 3D object detection, large-scale deep learning model training, and multi-target tracking tasks in complex scenes. Its fine annotation and large-scale data support the development and verification of high-precision algorithms.

ONCE: The ONCE dataset [10] is one of the largest and most diverse autonomous driving datasets collected by Huawei's Noah's Ark Lab for 144 h of driving in China. The dataset contains 16,000 typical scenes, including 417,000 3D frames and 769,000 2D frames, covering various labeling categories such as cars, pedestrians, buses, trucks, and cyclists. The weather conditions of the collection scenes were varied, including sunny, cloudy, and rainy days, and the time span ranged from morning to evening, covering

different time periods. The large data size of the ONCE dataset (1,000,000 scenarios) makes it suitable for data-driven deep learning methods, especially unsupervised or semi-supervised learning tasks. Although there are few annotations, its massive data provides a solid foundation for large-scale pre-training and diverse scenario research.

2.2 Evaluation Indicators

The most commonly used evaluation metrics for multimodal fusion 3D object detection tasks are the average precision (AP) for a single class, the mean average precision (mAP) across all classes, and the nuscenes detection score (NDS).

AP: Average-precision AP is a commonly used metric in object detection tasks, which is suitable for evaluating the detection performance of a model on a specific class (such as vehicles, pedestrians, and bicycles). In static or simple dynamic scenes, AP can better reflect the accuracy of the model for target localization and classification. Its limitation is that it only works with a single class and does not directly reflect the overall performance of the model in a multi-class task. In the evaluation of 3D object detection performance, the prediction frame with the intersection union ratio (IoU) of the detection frame and the truth frame is greater than a certain threshold, otherwise it is judged to be false. IoU represents the degree of overlap between the detection frame (box_1) and the truth box (box_2). The AP value can be obtained by calculating the area enclosed by the accuracy-recall curve and the coordinate axis, and the formulas for Precision, Recall, AP and IoU are as follows:

$$\text{Precision} = \frac{TP}{TP + FP} \tag{1}$$

$$\text{Recall} = \frac{TP}{TP + FN} \tag{2}$$

$$AP = \int_0^1 P(R)\, dR \tag{3}$$

$$IoU(box_1, box_2) = \frac{|box_1 \cap box_2|}{|box_1 \cup box_2|} \tag{4}$$

where TP is a true example (predicted to be positive, actual is positive), FP is a false positive example (predicted to be positive, actual is negative), FN is a false negative example (predicted to be negative, actual is positive), and R is to represent the recall rate.

LET-3D-APL: An evaluation index specially designed for 3D object detection tasks aims to solve the problem that traditional 3D detection indicators (such as 3D AP) are too sensitive to positioning errors. By introducing a tolerance mechanism for positioning error, the performance of the model in complex scenarios can be evaluated more comprehensively, especially in the detection task of vehicles, pedestrians and other objects in the field of autonomous driving. However, it depends on the setting of positioning error tolerance threshold and longitudinal affinity, and improper parameter selection may affect the fairness of the evaluation results. In addition, additional computational steps (e.g., longitudinal affinity calculations) are introduced, resulting in higher computational complexity than traditional 3D APs. The formula for calculating LET-3D-APL is as follows:

$$LET-3D-APL = \int_0^1 p_L(r)\, dr = \int_0^1 \overline{a}_l \cdot p(r)\, dr \tag{5}$$

where $p_L(r)$ is the longitudinal affinity-weighted precision value, $p(r)$ is the accuracy value at the recall r, and $\overline{a}_l$ is the average longitudinal affinity of the matching prediction that is considered to be the TP.

mAP: Mean Average Accuracy (mAP) is an extension of AP, which evaluates the overall performance of the model in multi-class object detection tasks by averaging the APs of all classes, and is a common indicator to evaluate the comprehensive performance of the model in multiple categories (such as vehicles, pedestrians, traffic signs, etc.). It is suitable for complex scenarios with a variety of targets, and can fully reflect the detection ability of the model. The limitation is that it may not be sensitive enough to datasets with unbalanced categories, and it cannot directly reflect the model's ability to predict attributes such as target direction and velocity. The matching strategy of the mAP is to replace the interaction ratio (IoU) of the 3D bounding box with the 2D center distance on the BEV plane. According to different distance thresholds, it can be divided into 0.5, 1, 2 and 4 m, and the mAP is averaged according to the AP in the above thresholds. It is calculated as follows in Eq. (6).

NDS: It is a composite scoring metric proposed by the NuScenes dataset that combines weighted scores from mAP and other tasks such as target direction, velocity, and attribute prediction. It is suitable for evaluating the comprehensive performance of the model in multiple tasks such as object detection, direction prediction, and velocity estimation. In dynamic scenarios, NDS can better reflect the model's ability to predict the target motion state (e.g., velocity, direction). Its limitations lie in its high computational complexity, the need for additional attribute prediction tasks, and its strong dependence on datasets, which is mainly applicable to specific datasets such as NuScenes. NDS is a weighted combination of mAP and other object attribute detection results, and consists of Mean Translation Error (ATE), Mean Scale Error (ASE), Mean Directional Error (AOE), Mean Velocity Error (AVE), and Mean Attribute Error (AAE). mAP is the average of the matching thresholds for the bird's-eye center distance D = $\{0.5, 1, 2, 4\}$ meters and the class C set, and the NDS indicator is calculated as follows:

$$mAP = \frac{1}{CD} \sum_{c \epsilon C} \sum_{d \epsilon D} AP_{c,d} \tag{6}$$

$$NDS = \frac{1}{10} \left[5mAP + \sum_{mTP \epsilon TP} (1 - \min(1, mTP)) \right] \tag{7}$$

$$mTP = \frac{1}{C} \sum_{c \epsilon C} TP_c \tag{8}$$

where $AP_{c,d}$ is the average precision, TP_c is the metric set of five average errors, and mTP is the average TP metric for all classes.

3 Background Technology

In this section, we provide a background overview of typical sensors used in autonomous driving and introduce camera-based object detection methods and lidar-based object detection methods.

3.1 Mainstream Sensors for Autonomous Driving

There are many different working principles and data collection methods of in-vehicle sensors, and their adaptability in different driving environments is also different. Due to the unique advantages and limitations of each mainstream sensor, a single sensor cannot fully meet the perception needs of unmanned vehicles. Therefore, multi-sensor fusion technology has emerged, which can integrate the advantages of each sensor and overcome the limitations of a single sensor, so as to provide more accurate information for smart cars

for decision-making planning and vehicle control, and enhance the safety of smart cars. Fig. 2 illustrates the mainstream sensors used in autonomous driving.

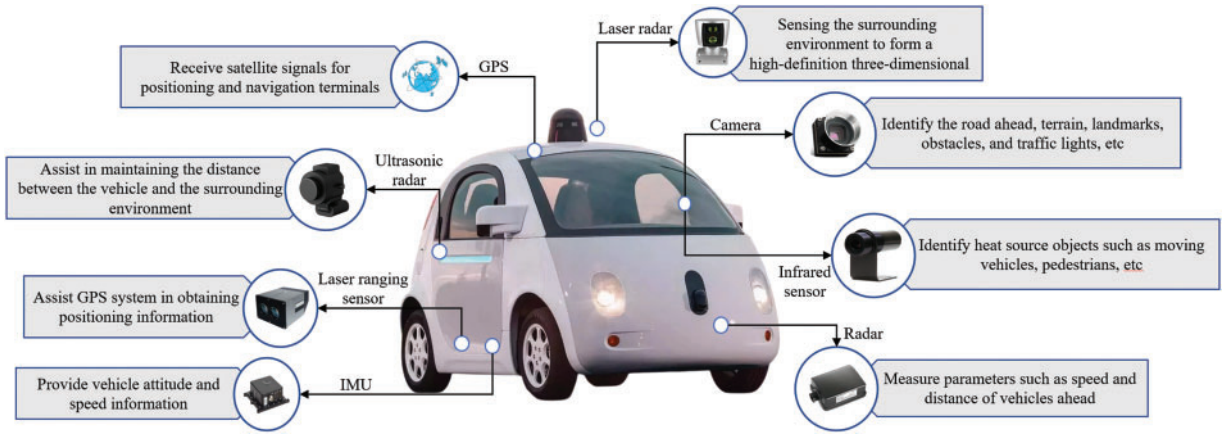

Figure 2: Sensors commonly used for autonomous driving

(1) Camera: The camera enhances the perception ability of the system by providing high-resolution image data in 3D target detection and fusing it with sensor data such as lidar. Not only does it have rich color and texture information to identify and classify targets, but it can also improve depth perception and build bird's-eye view features through stereo vision technology to achieve more accurate and robust target positioning in a variety of environmental conditions. However, the camera cannot directly obtain the 3D structural information of the scene, which limits its ability to accurately target in 3D space.

(2) LiDAR: The role of lidar in 3D object detection is mainly reflected in its ability to provide 3D structural information of the scene. By emitting a laser beam and measuring the reflection information, the lidar is able to obtain accurate depth images and point cloud data, with each point containing information such as depth, azimuth, and inclination. These data not only contain rich three-dimensional spatial information, but also compared with cameras, lidar is less susceptible to time and weather changes, and is suitable for the detection of targets in 3D space. The high accuracy and resolution of lidar make it play an important role in the perception of the vehicle environment in autonomous driving, especially in scenarios where precise target localization is required. However, lidar is unable to acquire color and texture information, which limits its application in target recognition and classification.

(3) Radar: The main advantages of radar in three-dimensional target detection are its strong penetration ability, high-precision ranging ability, and low cost. It detects targets by emitting electromagnetic waves and receiving signals reflected from the target, and it can also calculate information such as distance, velocity, and azimuth between the target and the radar. The high resolution of radar allows it to accurately detect and identify target objects, while its strong penetration allows it to work normally in adverse weather conditions such as rain, snow, and smoke. However, compared with lidar, radar has lower detection accuracy, smaller detection range, and relatively low angular resolution, which limits its recognition and positioning capabilities in target detection.

(4) Ultrasonic radar: Ultrasonic radar is mainly used in short-distance obstacle detection in three-dimensional target detection, and its functions include providing accurate short-range ranging and auxiliary parking. It has the advantages of low cost, simplicity in ranging methods, and the ability to triangulate with multiple probes. However, its detection range is limited, usually between 15 and 500 cm; Low update

frequency and propagation delay, suitable for low-speed scenarios; it cannot be accurately located, and it is greatly disturbed by the environment, such as signal transmission and echo co-channel interference.

Ultrasonic Radar: Ultrasonic radar is primarily utilized for short-distance obstacle detection in three-dimensional target detection applications, with key functions including precise short-range ranging and parking assistance. Its advantages include low cost, a straightforward ranging method, and the capability to perform triangulation using multiple probes. However, it has several limitations: its detection range is typically restricted to 15 to 500 cm; it exhibits a low update frequency and propagation delay, making it suitable only for low-speed scenarios; it lacks precise localization capabilities; and it is highly susceptible to environmental interference, such as signal transmission and echo co-channel interference.

(5) GNSS and high-definition maps: GNSS (Global Navigation Satellite System) provides precise global positioning information to assist vehicles in determining their precise location around the world. High-definition maps provide detailed road and environmental information. When integrated with GNSS data, this information enhances positioning accuracy and supports vehicle object detection and path planning. However, GNSS signals can be affected by factors such as buildings, weather, etc., resulting in reduced positioning accuracy. At the same time, high-definition maps need to be updated regularly to reflect road changes, which can be a disadvantage when map data is not updated in a timely manner.

(6) IMU and odometer: The IMU (Inertial Measurement Unit) is able to provide high-precision attitude and velocity information in a short period of time by measuring linear acceleration and angular velocity, while the odometer estimates the distance and direction traveled by the vehicle by measuring the rotation of the wheels. However, there is a cumulative error in the IMU, which can lead to large positioning deviations when used for a long time; The odometer is affected by wheel slippage and different ground conditions, resulting in measurement errors. In the target detection task, the data of IMU and odometer can help correct the change of viewing angle caused by vehicle movement, help to realize the tracking and positioning of dynamic targets, and improve the accuracy of detection.

3.2 Single-Modal Target Detection Methods

In this section, we introduce the background overview of the use of single-modal sensors for object detection, which is divided into two modes: image-based and point cloud, and introduce some classic detection methods according to different data modalities, sort out the method context, and analyze their advantages and disadvantages. Based on the detection method of image mode, we introduce the detection methods of monocular camera and binocular camera. Based on the detection method of point cloud mode, we divide it into projection method, voxel method, point cluster method, columnar method and graph structure method according to the technical path of point cloud feature processing.

3.2.1 Image-Based 3D Object Detection

With its small size and low price, car cameras have attracted wide attention from the industry. While it is capable of generating RGB images rich in texture and color information, it also lacks depth information, which makes it critical to recover the depth information of the scene in image-based 3D object detection. According to the processing strategies adopted by different network models for the input images, we classify the image-based 3D object detection methods as: monocular image 3D object detection and binocular stereo vision 3D object detection.

Monocular images cannot provide accurate scene depth information, and how to overcome this limitation has become the primary task of 3D object detection algorithms based on monocular images. To this end, the CenterNet [22] model proposes a keypoint-based detection method, the structure of which is

shown in Fig. 3. This is an anchor-free object detection method that can extract image features within a single crop area with minimal computing cost. In the image, CenterNet uses an anchor-free 2D detector to identify the target object's category, center point, viewing angle, and relative position and depth offset between the image plane and the object's 3D center point. Combined with the internal and external parameters of the camera, the model further estimates the three-dimensional position and heading angle of the target object in the world coordinate system. In order to improve the detection accuracy, two custom modules are also designed: a cascaded corner cell and a center cell. These two modules not only enrich the feature information in the upper left and lower right corners, but also enhance the feature expression ability of the central area, so as to provide more recognizable information. The M3D-RPN [23] model introduces a monocular 3D detection method with anchor points for the first time. This method uses predefined 2D and 3D bounding box anchors to first locate the 2D bounding box of the target object in the monocular image by a 2D detector, and then infer the object's 3D bounding box by matching the 2D and 3D anchor points. For the first time, the ROI-10D [24] model achieved second-order monocular 3D detection. The model first uses Faster RCNN [25] to identify the 2D region of interest of the object on the image, and then predicts the 2D center point, observation angle, relative depth and 3D scale of the object. Eventually, this information is converted through the camera's KT matrix to recover the object's 3D bounding box in the world coordinate system.

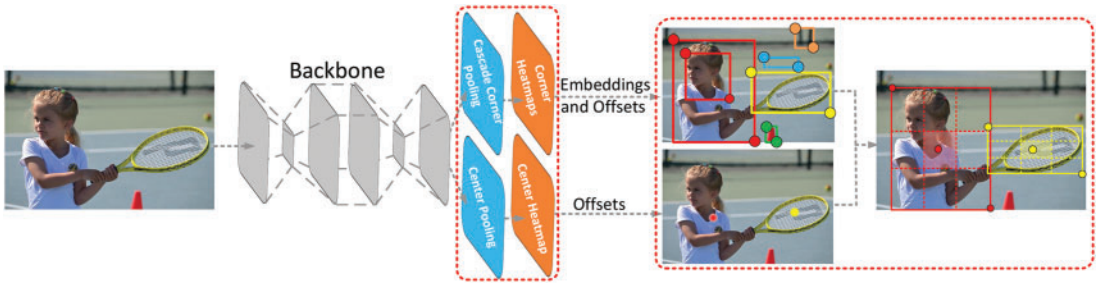

Figure 3: Object detection based on monocular vision: CenterNet

There are also some methods that use some prior knowledge of the image to make up for the disadvantage that the monocular image cannot provide scene depth information in the process of object detection, so as to optimize the detection effect of the model. Mono3D [26] proposed a monocular 3D object detection method with the help of prior information. In this method, the prior information such as object shape, pavement assumption, and object semantics are used to guide the detection process, and the difference between the 2D projected bounding box and the real 2D bounding box is combined to identify the optimal 3D candidate frame of the object, so as to optimize the 3D detection effect. On the basis of this method, models such as DeepMENTA [27], Mono3D++ [28], and 3D-RCNN [29] further propose monocular target detection methods guided by object contour information. These methods represent the outline of an object as a mesh connected by a fixed number of vertices and reconstruct the 3D bounding box of the target object by predicting the position of these mesh vertices.

Compared with monocular images, binocular images can obtain accurate scene depth information through stereo matching technology, which makes 3D object detection based on binocular stereo images an important research direction in the field of autonomous driving. The Disp R-CNN [30] model proposes an object detection method based on binocular instantiated parallax map, the structure of which is shown in Fig. 4. This instanced parallax map generation enables more accurate parallax estimation by predicting the parallax of pixels on an object of interest and learning the shapes of a particular class in advance. In order to solve the challenge of scarce parallax annotation in training, the method uses a statistical shape model

to generate dense parallax pseudo-ground true values, and still obtains the depth information of the scene without using lidar. In addition, with the support of the image instantiation segmentation network, Disp R-CNN can pre-exclude background pixels that are not related to the regression of the bounding box of the target object in the detection process, so as to achieve better detection results. The Stereo R-CNN [31] model designed a dual-network interconnected binocular 3D detection architecture, which extracted features from the left and right frames of the image through two 2D detectors, and generated the left and right regions of interest of the target object. By docking two regions and using the key constraints of the 2D-3D bounding box, the 3D bounding box of the target object is accurately returned. The TL-Net [32] model, which also adopts the dual-network interconnection mode, introduces the triangulation method in the region of interest alignment stage, which further optimizes the 3D positioning accuracy of the target object. The Sparse4D [33] model introduces three modules: time instance denoising, quality estimation, and decoupled attention. Time instance denoising effectively reduces noise in sparse instances by improving the attention module and replacing feature addition operations with concatenation; quality estimation introduces auxiliary training tasks to allow the model to better evaluate the accuracy of detection boxes, thereby improving detection indicators and accelerating model convergence; decoupled attention is proposed in the feature interaction module, which improves the perception effect without increasing the inference delay. The Far3D [34] model proposes an innovative long-range 3D object detection framework that aims to break through the bottleneck of long-distance detection. The method first generates 3D adaptive queries through high-quality 2D object priors to expand the perception range; then uses 3D spatial offset sampling and 3D-2D view transformation to aggregate multi-scale features and reduce the amount of computation; finally, a range-modulated 3D denoising method is used to solve the query error propagation problem and improve the model stability. These models improve the accuracy and efficiency of binocular 3D inspection through innovative network structures and alignment techniques.

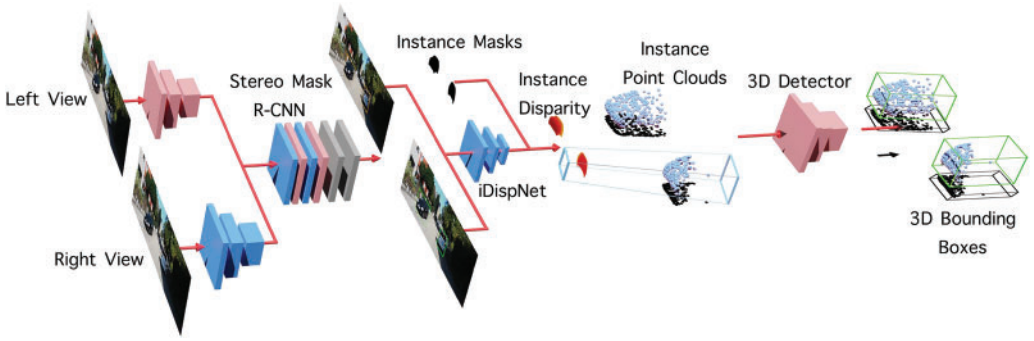

Figure 4: Object detection based on binocular vision: Disp R-CNN

In Table 2, we have sorted out the image-based object detection methods mentioned in this section in detail, systematically summarized the advantages and disadvantages of each method, and combined with actual application scenarios, deeply analyzed its scope of application and limitations, providing readers with a comprehensive and in-depth reference basis for better selecting detection technology suitable for specific needs.

Table 2: Advantages and disadvantages of each method and applicable scenarios

Method	Advantage	Disadvantage	Applicable scenarios
CenterNet [22]	Simple and efficient, no anchor frame design required, suitable for a variety of tasks	It requires high accuracy in center point positioning and is easily affected by occlusion or dense targets.	Scenarios with high real-time requirements, such as obstacle detection in autonomous driving.
M3D-RPN [23]	Combining 2D and 3D information, it performs better on monocular images	Depth estimation error affects 3D detection accuracy.	3D object detection in monocular camera scenes, such as vehicle detection in autonomous driving.
ROI-10D [24]	The 10-dimensional parameters represent the 3D box with high accuracy	The computational complexity is high and the real-time performance is poor.	Scenarios that require high-precision 3D detection, such as scene reconstruction for autonomous driving.
Mono3D [26]	Designed for monocular images, using geometric constraints and contextual information	It is highly dependent on depth estimation and easily affected by image quality.	3D object detection in monocular camera scenarios, such as autonomous driving.
DeepMENTA [27]	Multi-task learning, joint optimization of 2D and 3D detection tasks	The model is highly complex and takes a long time to train.	Scenarios that require simultaneous 2D and 3D detection, such as autonomous driving or robot navigation.
Mono3D++ [28]	Improved version of Mono3D, introducing more geometric constraints and context information	Still relies on depth estimation and is sensitive to image quality.	High-precision 3D detection in monocular camera scenes.
3D-RCNN [29]	Based on Faster R-CNN, it combines 2D and 3D information with high accuracy	The computational complexity is high and the real-time performance is poor.	Scenarios that require high-precision 3D detection, such as autonomous driving or augmented reality.
Disp R-CNN [30]	Combined with disparity information, suitable for binocular cameras, depth estimation is more accurate	Relying on binocular cameras, it is not applicable to monocular scenes.	3D object detection in binocular camera scenarios, such as autonomous driving or robot navigation.

(Continued)

Table 2 (continued)

Method	Advantage	Disadvantage	Applicable scenarios
Stereo R-CNN [31]	Designed for binocular images, combining left and right view information, high accuracy	Relying on binocular cameras, the computational complexity is high.	High-precision 3D detection in binocular camera scenarios, such as autonomous driving.
TL-Net [32]	Improve model performance and reduce training time and data requirements through transfer learning	The similarity between the source domain and the target domain is high, and the transfer effect may be unstable.	3D object detection in data-scarce or specific scenarios, such as industrial inspection.

3.2.2 Lidar Point Cloud-Based 3D Object Detection

Lidar depicts a scene by emitting a laser beam and receiving its reflection, and by measuring the time difference between the beam emission and reception, it can accurately obtain the three-dimensional spatial position of an object. However, the point cloud data generated by lidar is usually sparse, disordered, and irregular, which makes it difficult for traditional CNN models to process directly. At present, the commonly used lidar point cloud processing methods include projection method, voxel method, point cluster method, columnar method, and graph structure method. These methods are designed to make point cloud data suitable for deep learning models for effective 3D object detection and analysis.

Projection is a technique that converts a 3D lidar point cloud into a 2D image for processing using a traditional CNN model. The main projection methods include forward field of view (RV) projection and bird's-eye view (BEV) projection. For the first time, the VeloFCN [35] model uses RV projection to convert the point cloud into a 2D Point-map and then uses a fully convolutional neural network to detect it in 2D to obtain a 2D bounding box containing 3D shape information. For the first time, the BirdNet [36] model introduces a BEV projection based on lidar point cloud for 3D target detection, and its structure is shown in Fig. 5. The model converts the lidar point cloud into a 2D BEV-map containing altitude, reflection intensity, and density information, and then uses Faster RCNN to perform 2D object detection on the BEV-map. Finally, the 3D bounding box of the target object is recovered by combining the road surface information and the height of the object from the ground. This process enables the effective use of the 3D information extracted from the lidar data, improving the accuracy of the inspection. The BirdNet+ [37] model is optimized on the basis of BirdNet, transforming the regression process of the 3D bounding box into end-to-end processing and eliminating the time-consuming post-processing steps, thereby improving the detection accuracy and inference speed of the model. The MVF model [38] proposes a 3D detection method based on the fusion of lidar point cloud BEV projection and RV projection, which combines point cloud data from different perspectives and adopts dynamic voxel point cloud feature extraction technology, so as to stand out among many detection algorithms. The FocalFormer3D [39] model is based on point cloud projection and guides the model to focus on mining difficult instances through multi-stage instance detection. It uses multi-stage heat map prediction to generate queries for difficult instances and excludes simple positive samples by accumulating positive sample masks, allowing the model to focus on difficult-to-detect targets. In addition, the model uses a box-level Transformer decoder to efficiently distinguish real objects from a large number of candidate objects, thereby significantly improving detection accuracy.

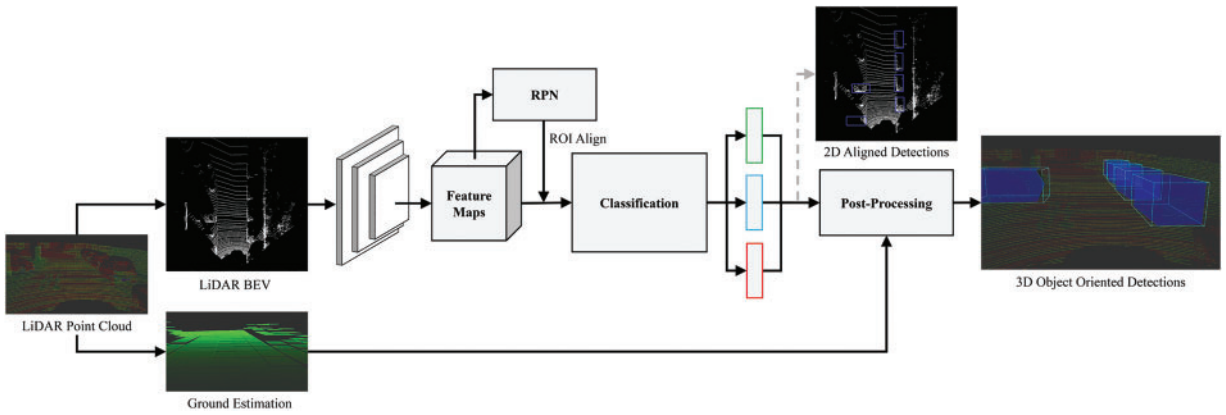

Figure 5: The architecture of BirdNet

The voxel method converts point cloud data into a continuous, regular three-dimensional grid structure, and the local features of each grid element are defined by the average or maximum feature value of all points within the cell. Then, 3D CNNs were used to extract and model the features of these voxelized data. The VoxelNet [40] model is the first to introduce a detection method based on lidar point cloud voxelization, and its structure is shown in Fig. 6. The model first divides the 3D point cloud into a fixed-size voxel mesh, and normalizes and fills the point cloud within each voxel to generate voxel-level input features. Then, the local and global features were extracted through the voxel feature encoding (VFE) module, and the sparse point cloud features were mapped into dense representations. Then, the 3D convolutional network was used to further extract the multi-scale spatial features and generate a structured feature map. Subsequently, the regional proposal network (RPN) performs a sliding window operation on the feature map, generates candidate boxes, and optimizes the detection results through classification and bounding box regression. Finally, the overlapping boxes were removed by non-maximum suppression (NMS) to obtain the final detection results, including the target class, 3D bounding box and confidence level. End-to-end learning is used to replace traditional manual feature engineering, which effectively improves the 3D object detection performance and robustness of point cloud data. The SA SSD [41] model further optimizes the voxel block division of point clouds, and proposes an efficient voxel block division method based on tensor calculation to solve the problem of time-consuming data traversal calculation in the original voxelization process, which significantly improves the computational efficiency of the model. The FSTR-L [42] model combines the latest sparse backbone network based on sparse voxelization, introduces dynamic query to provide the decoder with the prior position and context of the foreground, and discards high-confidence background tags to further reduce redundant calculations. Gaussian denoising query is proposed to accelerate decoder training and make it more adaptable to the distribution of sparse voxel features. The development of these models demonstrates the importance and potential for efficiency improvement of point cloud voxelization in 3D object detection.

The point clustering method refers to the independent processing of each data point, and the characteristics of the data points in the cluster are extracted through clustering and multilayer perceptron (MLP) layers. This allows points with similar features to converge, and point differences with different features to be amplified, so that different objects in the point cloud can be distinguished by discerning these differences. The PointRCNN model [43] is the first to propose a 3D detection mode based on lidar point cloud point cluster processing, which realizes high-precision target detection by directly processing sparse point clouds, and its structure is illustrated in Fig. 7. In the first stage, the features are extracted from the original point

cloud, the geometric and semantic features of each point are generated by using the PointNet++ [44] network and the 3D candidate box is predicted based on the features of the points. These candidate boxes generate a preliminary target position through regression, while the quality of the candidate boxes is evaluated by classification. In the second stage, the candidate boxes generated in the first stage are refined. The points in the candidate boxes are resampled, and local features are extracted to further optimize the size, position, and orientation of the 3D bounding box, as well as improve the accuracy of target classification. The entire process is trained in an end-to-end manner, learning directly from the original point cloud, avoiding the limitations of traditional manual feature engineering. The 3DSSD [45] model introduces a first-order lidar point cloud point cluster 3D detection method, which omits the step of generating object proposals in the traditional second-order detection method and directly detects the model, thereby improving the computational efficiency of the model. In addition, 3DSSD also proposes an innovative point cloud fusion sampling strategy, which can retain the feature information of the point cloud data better than the farthest point sampling method used in PointNet++, which is helpful to return to the 3D bounding box of the object more accurately.

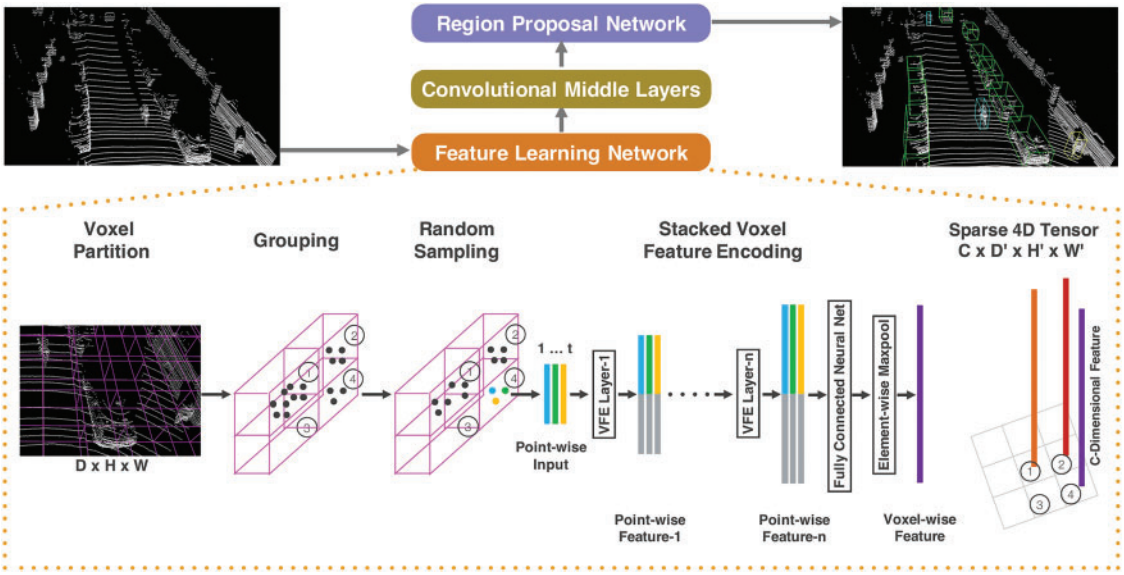

Figure 6: The architecture of VoxelNet

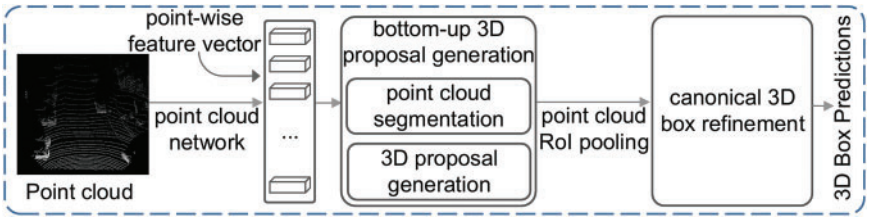

Figure 7: The architecture of PointRCNN

The columnar method refers to organizing the 3D point cloud data into a regular cylindrical structure, ignoring the height difference of the point cloud in the vertical direction, and compressing the point cloud

in the same area into a two-dimensional BEV feature map, so that 2D CNN can be used instead of 3D convolution, simplifying the calculation process. The PointPillar model [46] innovatively proposes a 3D detection mode based on lidar point cloud columnarization processing, and its structure is shown in Fig. 8. Firstly, the point cloud data is divided into a fixed grid structure according to pillars, and the point clouds in each cylinder are normalized and input into the PointNet [47] network, and the local point features are extracted and aggregated into a feature representation at the cylinder level. These beam features are then projected onto a BEV plane to form a dense 2D feature map, simplifying complex 3D data processing. Then, the 2D CNN extracts multi-scale features from the feature map to generate depth features suitable for object detection. Finally, the candidate target boxes were generated through the RPN, and the object detection was completed by classifying and regressing the bounding boxes. This method converts the 3D point cloud features into 2D, so that the processing process can use the conventional 2D CNN model without 3D convolution operation, which not only retains the sparse characteristics of the point cloud, but also significantly reduces the computational complexity, and realizes real-time and efficient 3D object detection. The Pillar-OD model [48] is an improved version of PointPillar, adding the RV projection to the columnar BEV projection. By stitching together the feature maps of both perspectives, it completes 3D target detection and improves performance.

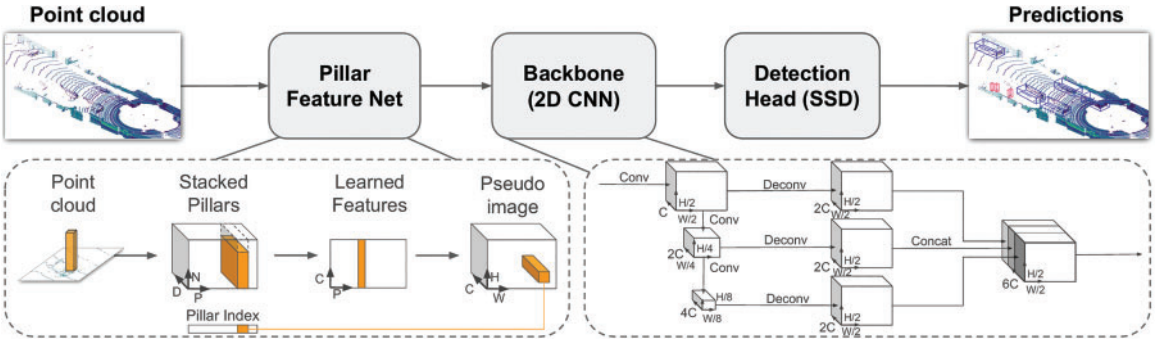

Figure 8: The architecture of PointPillar

The graph structure method is a method to construct a graph structure based on the topological connection relationship between point cloud data, which allows the application of graph data processing technology to model irregular lidar point clouds. This method is similar to point clustering, which can directly extract features from the point cloud without any pre-processing steps, so that the network model can be applied to the original point cloud data. For the first time, the Point-GNN [49] model introduces the graph structuring processing based on lidar point cloud for 3D detection, and its structure is shown in Fig. 9. The model first represents the point cloud as an undirected graph, with each point as a node, and the edges between the nodes are connected based on Euclidean distances or predefined rules. Secondly, the point cloud features are iteratively updated through the multilayer graph neural network, and the node features aggregate the surrounding information to learn the local structure features, while retaining the global context information. Subsequently, the edge connections and features are dynamically updated in each iteration to further optimize the graph structure and representation capabilities. Next, the model predicts the parameters of the target category, center point position, and 3D bounding box for each node in the point cloud. Finally, the redundant detection frame was removed by NMS to obtain the final target detection result. The core advantage of this method is that the point cloud data is directly modeled with a graph structure, which avoids the information loss caused by voxelization or projection, and significantly improves the 3D object detection performance of sparse point clouds. The SVGA-Net [50] and STA-GCN [51] models integrate the visual

attention mechanism into 3D object detection based on the point cloud map structure. The local attention mechanism is used to assign different weights to each node, and the importance of each data point in the point cloud in the convolution process is highlighted. The global attention mechanism assigns weights to different subgraphs to identify the importance of each local area in the point cloud map structure in object detection. These multi-level and multi-attribute attention mechanisms help the model extract point cloud features more efficiently, thereby improving the performance of 3D detection.

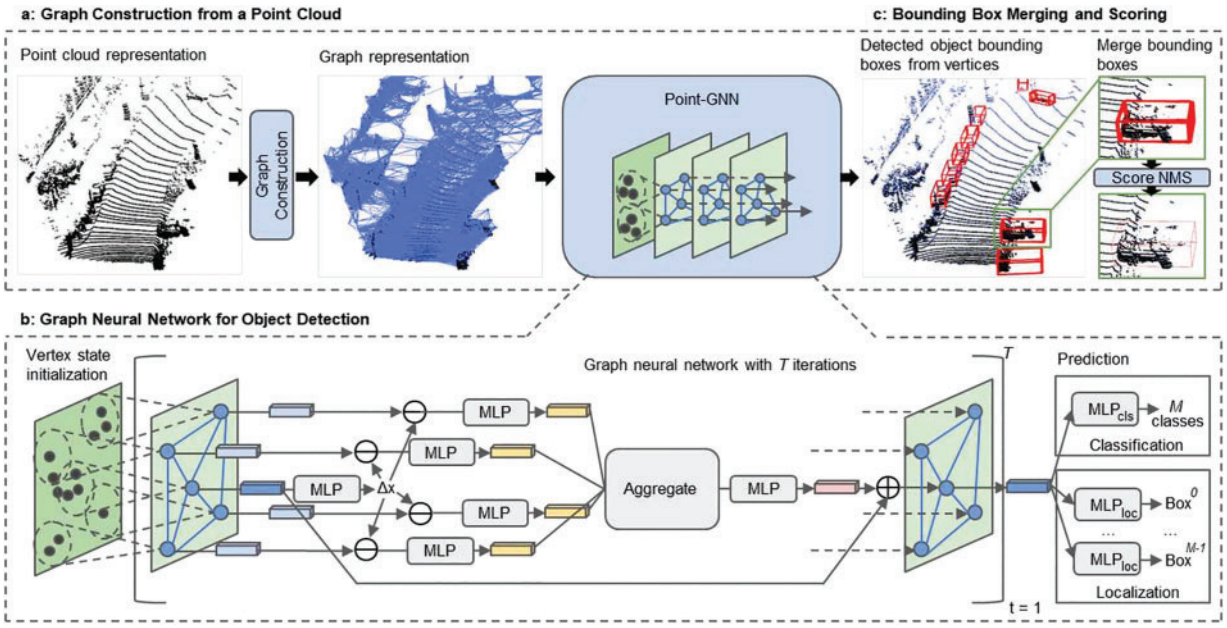

Figure 9: The architecture of Point-GNN

4 Multi-Modal 3D Object Detection

Image and point cloud have different data characteristics and advantages, and the fusion of these two data modalities can complement each other's advantages and disadvantages and improve the performance of 3D inspection. When constructing a neural network for multimodal fusion object detection, three core questions need to be considered: (1) What to fuse: determine which sensor data modalities should be fused, and how to correctly represent and process these data; (2) When to fuse: at which stage of the feature extraction of the neural network is the data fusion; (3) How to fuse: choose what operation to achieve data fusion. Sections 4 and 5 will summarize existing approaches to convergence around these three issues.

4.1 What to Fuse?

The inputs of the fusion module show the diversity of algorithms and represent the unique ideas of each algorithm, and this paper focuses on the fusion inputs of images and point clouds. Specifically, the sensor can collect image and point cloud data, but the input of the fusion module is generally processed data, including radar data, image feature maps, segmentation masks, or pseudo-lidar point clouds.

(1) Raw point cloud data: Lidar generates point clouds by scanning the environment, and each point cloud contains depth and reflection information, where the depth information is encoded by the attributes of the point, and the reflectivity information is reflected by the intensity value of the point. These point cloud points are typically represented as quadruples (x, y, z, r), where r stands for reflectivity, and different surface

textures produce different reflectances, providing additional information for a variety of tasks. Although point cloud data can be used directly, in order to improve efficiency and performance, researchers often convert point clouds into voxel or 2D projection before they are fed into downstream modules.

(2) BEV or RV projection: Commonly used point cloud inputs for the fusion module include BEV and RV projection. RV projection can retain the full resolution of lidar data and reduce the loss of spatial information, but it is susceptible to the change of object scale. In the fusion module, the RV projected data is usually processed by 2D CNN to extract the view features, and then the size is unified with the image features. In contrast, BEV projection provides accurate three-dimensional spatial information, such as the distance, height, and width of the target, making it suitable for complex environments and effective in low light or inclement weather. Due to the lack of occlusion between targets, the point clouds of each target can be identified and processed independently, showing strong 3D environment awareness.

(3) Image data: Deep neural networks can extract appearance and geometric feature maps from the original image, which can explore richer appearance cues and a larger receptive field than the original image, so as to achieve in-depth interaction between modalities. The Shift R-CNN [52] model combines deep learning and geometry to first predict the 2D bounding box, 3D dimensions, and orientation information of the target, and then use these prediction parameters and camera projection matrices to solve the closed solution of the 3D transformation.

(4) Another image processing method is to use semantic segmentation network to obtain pixel-by-pixel segmentation masks. These masks are a stand-alone result of image processing and are often used to fuse with other sensor data. Compared with feature maps, the use of segmentation masks as fusion inputs has two main advantages: first, the image mask is used as a compact outline feature of the image; Second, through point-to-pixel mapping, pixel-level image masks can be easily used to match lidar points.

(5) Another method is to convert the image into pseudo-point cloud data. The parallax map formed by stereo matching of binocular images can be converted into a scene depth map with the help of camera internal parameters, and then pseudo-point cloud data can be generated. Based on this idea, the Pseudo-LiDAR++ [53] model optimizes the process of generating disparity maps from binocular images. The CG-Stereo [54] model proposes an image processing method based on full-scene pseudo-point clouds, which predicts and generates a target object confidence distribution map in the process of converting binocular parallax maps into pseudo-point clouds, which guides the transformation process of pseudo-point clouds.

4.2 When to Fuse?

According to the time point at which fusion is performed in the data processing process, data fusion can be divided into three types: early fusion, middle fusion, and late fusion. The selection of the fusion period in multimodal object detection is mainly based on the following criteria:

1. Data characteristics: High correlation between modalities is suitable for early fusion, while large differences are suitable for mid-term or late fusion.

2. Task requirements: High-precision requirements tend to mid-term fusion, while real-time requirements tend to late fusion.

3. Computing resources: The computing cost of early fusion is high, while the efficiency of later fusion is high.

4. Complementarity between modalities: Mid-term fusion can exert its advantages more when the complementarity is strong.

The advantages, disadvantages, and applicable scenarios of the three fusion periods are shown in Table 3. Early fusion is suitable for scenarios where modal data is highly correlated and resources are sufficient; mid-term fusion is suitable for high-precision tasks that require the combination of complementary information between modalities; late fusion is suitable for scenarios with high real-time requirements or large differences in modal data. Based on specific application scenarios and requirements, selecting an appropriate fusion period can balance accuracy, efficiency, and robustness.

Table 3: Advantages and disadvantages of different fusion periods and applicable scenarios

Integration period	Advantage	Disadvantages	Applicable scenarios
Early fusion	Make full use of the original data correlation and the model is simple	High computational cost, high data alignment requirements, and sensitive to noise	High correlation between modal data (such as image + depth map), sufficient computing resources
Middle fusion	Combining complementary information between modalities, high flexibility	The model is complex and the computational cost is high	The information between modalities is highly complementary (such as image + point cloud), requiring high-precision detection
Late fusion	High computational efficiency, strong robustness, and low requirements for data alignment	Unable to fully utilize complementary information between modalities, and the fusion strategy design is complex	High real-time requirements and large data differences between modalities

(1) Early Fusion: Also known as data-level fusion, it is the combination of data from different modalities at the original data level to form a data modality containing richer feature information. The general flow of pre-fusion is shown in Fig. 10. This method can more completely retain the original data details of RGB images and lidar point clouds, provide more comprehensive multi-modal data for the 3D object detection model, and help improve the detection accuracy of target obstacles. The F-PointNet [55] model is the first to propose a 3D detection method based on image-point cloud pre-fusion fusion. In this method, a 2D detector is used to identify the 2D bounding boxes of the target object on the RGB image, and then these boxes are projected onto the lidar point cloud to filter out the point cloud data in the frame. The three-dimensional coordinates of these points are stitched with the corresponding pixel information in the image to form a 3D point cloud frustum space containing multimodal information. Next, F-PointNet utilizes two PointNet structures, one for splitting the data points of the target object from the frustum space, and the other for regressing the 3D bounding box from these split points. On the basis of F-PointNet, the F-ConvNet [56] model proposes to segment the 3D point cloud frustum space formed by fusion, and each segment uses an independent PointNet structure to extract features, and finally synthesizes the features of all segments to predict the 3D bounding box of the target object. The F-PointPillar [57] model has two improvements: one is to change the spatial segmentation process of the frustum based on PointNet to the point segmentation based on Gaussian distribution statistics; The second is to change the 3D bounding box regression process

from PointNet-based to Pillar-based method. These improvements are designed to improve the accuracy and efficiency of segmentation and regression.

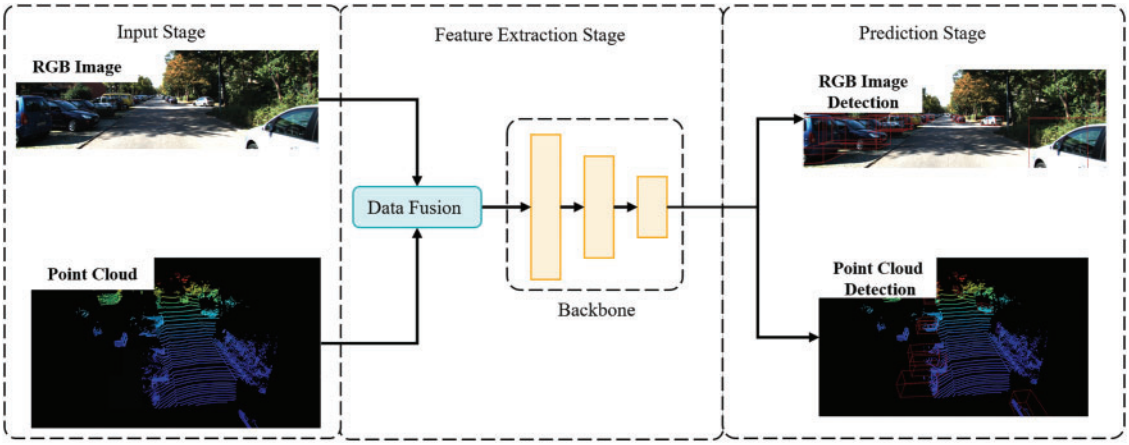

Figure 10: Early fusion framework

(2) Middle Fusion: Middle fusion refers to the combination of different forms of features or the generated target object proposal after the data of the two modalities are extracted to jointly complete the 3D object detection task. The general flow of mid-term fusion is shown in Fig. 11. This approach allows each modality to independently contribute its own unique information, which is then aggregated to improve the accuracy of detection. For the first time, the MV3D [58] model introduces a 3D detection mode based on image-point cloud mid-term fusion, which uses the fisheye view (FV) and BEV projection of the lidar point cloud and RGB images to generate a proposal of the target object from multiple perspectives. These proposals are clipped, aligned, and pooled together to complete the 3D bounding box regression task. The PI-RCNN [59] model further fuses the 2D proposal generated by 2D semantic segmentation of images with the 3D proposal generated by PointNet++. The SCANet [60] model uses a spatial attention mechanism to fuse RGB images and lidar point cloud BEV projection maps to generate 2D and BEV proposals. The FusionPainting [61] model fuses the 2D proposal generated by 2D semantic segmentation of the image with the 3D proposal generated by the 3D semantic segmentation of the lidar point cloud. The Graph-RCNN [62] model fuses the 2D proposal generated by image 2D object detection with the 3D proposal of lidar point cloud based on voxelization. The Fusion-RCNN [63] and TransFusion [64] models proposed an image-point cloud proposal fusion scheme based on Transformer [65], and the two proposals were fused through the Transformer module of the attention mechanism. These methods make use of the high-level semantic features extracted by deep learning, which makes the fusion process more efficient and more informative, and enhances the performance of object detection.

(3) Late Fusion: Late fusion is a method that integrates the decision-making results of different network modules independently on different data modes according to specific rules to produce a comprehensive and better detection result. The general flow of post-fusion is shown in Fig. 12. For the first time, the CLOCs [66] model proposes a 3D detection mode based on image-point cloud post-fusion, in which the 2D and 3D candidate boxes of the target object are generated through independent 2D and 3D detectors, and then these candidate boxes are fused according to the geometric relationship between the 2D–3D bounding boxes to improve the accuracy of the model output. Subsequently, the Fast-CLOCs [67] model improved the 2D and

3D detectors in CLOCs, and proposed a more efficient and lightweight post-fusion 3D detection mode, which further improved the detection accuracy and computing speed.

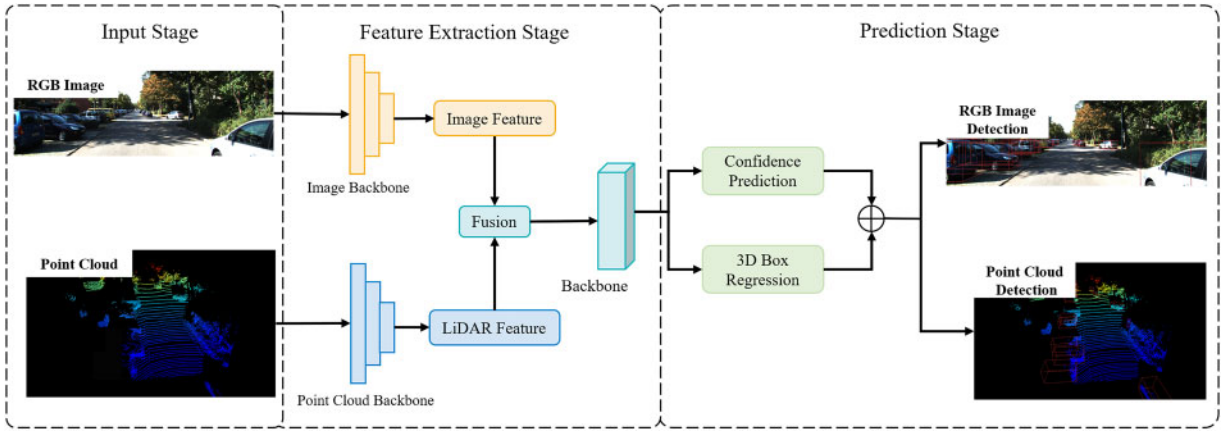

Figure 11: Middle fusion framework

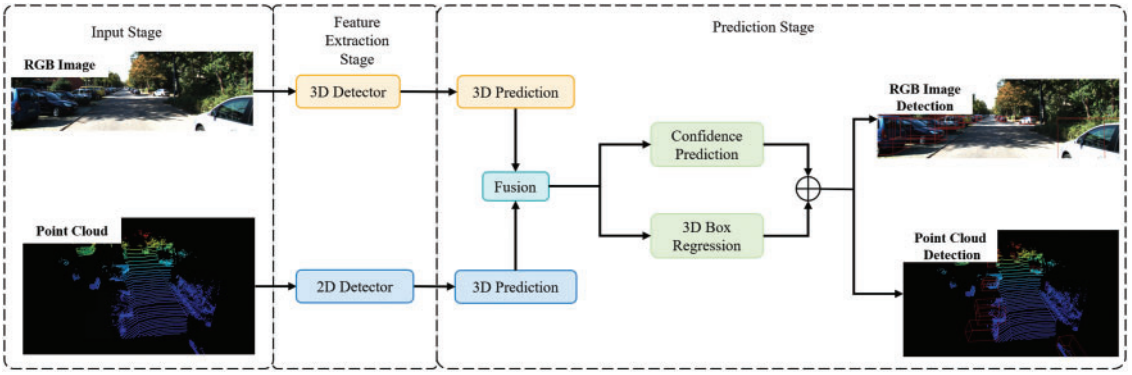

Figure 12: Late fusion framework

5 How to Fuse?

This section summarizes typical fusion operations in deep neural networks according to different types of sensor fusion strategies, including Camera-Lidar fusion, Camera-Radar fusion, Camera-Lidar-Radar fusion, and other sensor fusions.

5.1 Camera-LiDAR Fusion

The Camera-Lidar fusion technology combines the camera's high-resolution image information with the lidar's accurate distance measurement capabilities to provide more comprehensive environmental perception, enhance the robustness of the autonomous driving system under different lighting and weather conditions, and improve the accuracy of object detection and positioning, so as to achieve safer and more reliable decision-making and control in complex traffic scenarios. The existing image and point cloud fusion technologies are mainly divided into early fusion, feature-level fusion, decision-level fusion and BEV fusion.

5.1.1 Data-Level Fusion

Data-level fusion can make full use of the information of each sensor at the data level, so that the data of different sensors can complement each other and enhance the overall perception ability. This fusion strategy allows for the elimination of uncertainty and redundancy in the sensor data at an early stage of data processing, preserving more of the original information, thereby reducing the complexity of subsequent processing, helping to speed up the response of the system and improve the accuracy and reliability of the detection results.

Many of today's advanced 3D detection algorithms are dedicated to effectively fusing images and lidar point cloud data. Cao proposed a multi-view object detection method, MVFP [68], which effectively reduces the missed detection rate and improves the performance by adding a bird's-eye view (BEV) detection module on the basis of F-PointNet [55]. The network structure is shown in Fig. 13. Firstly, F-PointNet is used to generate preliminary 3D target detection results by combining RGB images and lidar point clouds. The lidar point cloud is then encoded into a BEV feature map, which is used to predict the 2D bounding box. By calculating the intersection union ratio, the detection results were matched with the BEV prediction results. For unmatched targets, their 2D bounding boxes are projected back into the F-PointNet network, where detection and fusion continue until all targets in the BEV feature map find a match in the detection results. The robustness of 3D detection is significantly improved by the joint optimization of multi-modal features.

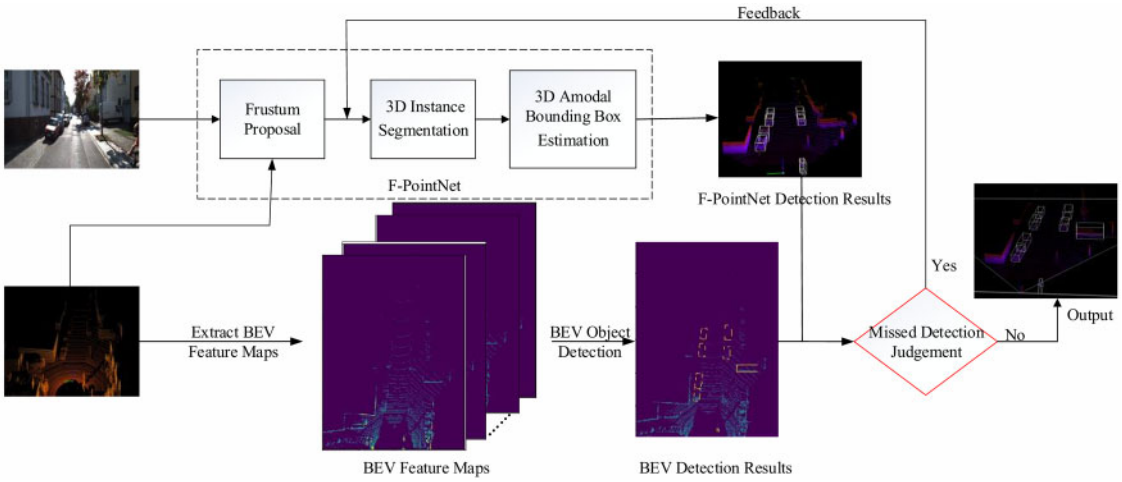

Figure 13: The architecture of MVFP

In order to better realize the fusion of different sensor data and improve the detection performance, Wei proposed a method called ConCs-Fusion [69], which uses a contextual clustering network to learn the multi-scale features of radar point clouds and images and performs upsampling and fusion of feature maps. The multi-layer perceptron is used to perform nonlinear representations of the fused features to reduce the feature dimension and improve the model inference speed. The context clustering network aggregates sensor feature points based on their similarity, fusing them into radar-camera feature fusion points for two-way cross-modal fusion. In order to reduce the feature dimension and improve the model inference speed, Ku proposed an aggregate view object detection network AVOD for autonomous driving scenarios [70]. The network uses lidar point clouds and RGB imagery to generate features shared by two sub-networks: the regional suggestion network (RPN) and the Stage 2 Detector Network. Among them, RPN can learn the high-resolution feature map after data fusion, and generate reliable suggested regions for different targets in the

scene. By learning from these regions, the second-stage detection network can achieve accurate 3D bounding boxes and category classification. In order to enhance environmental perception and improve robustness and accuracy, Liu et al. [71] proposed a large-scale 3D object detection algorithm that fuses 2D visual information and 3D lidar data. In this algorithm, a calibration module based on EPnP is proposed for data alignment to improve the performance of data association. A 2D bounding box of the object is generated through visual inspection, and the corresponding 3D point cloud is extracted. After ground fitting and point cloud clustering, an innovative scoring mechanism was introduced to identify the clusters belonging to the target. Finally, an accurate lidar point cloud is generated in the 3D frustum. This method does not require a large number of 3D point cloud datasets to be trained and is suitable for lidar with any number of lines to achieve real-time 3D object detection.

5.1.2 Feature-Level Fusion

Feature-level fusion is usually the integration of features extracted from the original data for subsequent 3D object detection. In the multimodal fusion model, different sensors usually have different resolutions, coordinate systems, and sampling densities, and how to achieve the matching and alignment of different modal data is crucial. Chen et al. proposed a feature fusion strategy called AutoAlign [72] to solve the challenge of RGB images and lidar point clouds complementing each other in autonomous driving, and its structure is shown in Fig. 14. Unlike the fixed mapping relationship based on the camera projection matrix, AutoAlign dynamically models the feature mapping between the image and the point cloud through a learnable alignment map. Through the cross-attention feature alignment module, the image and point cloud features can be adaptively aggregated, and then combined with the self-supervised feature interaction module, the semantic information is used to guide the process of feature alignment, which enhances the semantic consistency. This dynamic feature alignment enables more efficient cross-modal information fusion. AutoAlignV2 [73] uses a projection matrix to guide the automatic alignment of cross-modal features, and implements sparse sampling between modalities, so as to establish the association between image features and point cloud features for each voxel. Alaba et al. [74] proposed an end-to-end framework for sparse interaction fusion. In this method, the noise radar filter (NRF) module is introduced to extract the foreground features, the query semantic features are used to filter out the noise radar features from the image, and then the sparse cross attention encoder (SCAE) is used to fuse the foreground radar and the image, which solves the problem of position ambiguity at the sparse level. Xu et al. [75] proposed a Multi-Sem Fusion model. The model reprojects the 2D semantic information into a 3D point cloud with calibration parameters, introduces the AAF attention module to learn the fusion score to solve the problem of mismatch, and inputs the fused data to the 3D object detector to generate the final detection results.

There are also methods that focus on performing fusion operations on the features of the input. For example, FUTR 3D [55] uses a query-based feature sampler and a Transformer decoder, equipped with a set-to-set loss function, to achieve feature fusion and perform 3D object detection, showing strong flexibility. UVTR [76] transforms image features into a predefined space and constructs an image voxel space to achieve cross-modal interaction, enhance the fused modal information, and improve the accuracy of the detection results. Song et al. [77] proposed the VoxelNextFusion model, which projects point clouds onto images to obtain pixel and region-level features, and uses the self-attention mechanism to fuse these features, effectively correlating sparse point clouds with dense images, thereby reducing association errors and improving detection accuracy. Liu et al. [78] proposed the BAFusion module, which uses cross-attention to adaptively fuse lidar and camera information, optimizes computational complexity and facilitates high-level interaction of image and point cloud data, improving the flexibility and robustness of the algorithm. Deng et al. [79] proposed a multimodal 3D object detection framework, PoIFusion. This is different from the approach of

transforming multi-sensor data into a unified view or leveraging global attention mechanisms to facilitate fusion. The method follows the query-based object detection paradigm, where the target query is represented as a dynamic 3D box. For each query box, a set of points of interest (PoI) is generated and these points are projected into views of different modalities to sample the corresponding features. Through the dynamic fusion module, the multi-modal features of each group of PoI are integrated. Finally, all PoI features in the same query box are aggregated to update the query features.

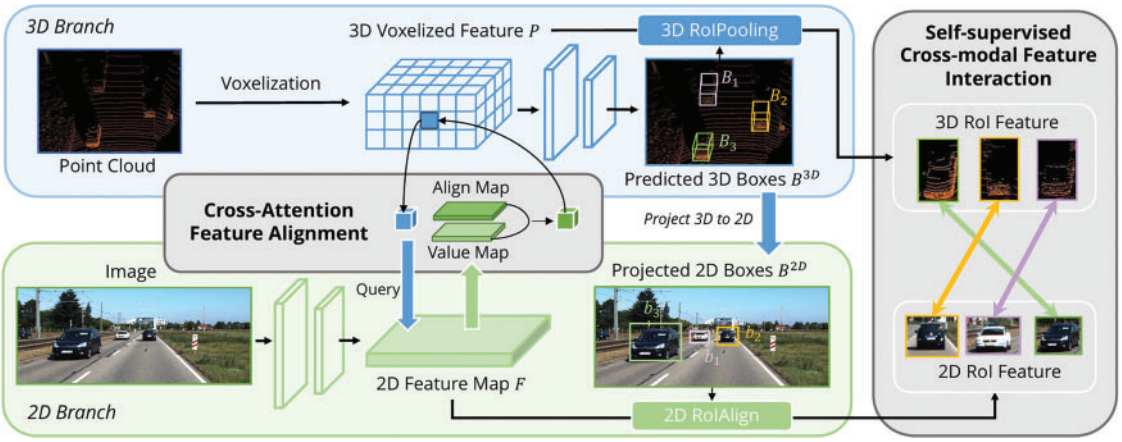

Figure 14: The architecture of AutoAlign

In order to improve the detection performance of fusion algorithms in complex environments and the ability to detect small targets, Wang proposed an embedded fusion three-dimensional object detection network based on deep learning (VoPiFNet) [80], which can receive data streams from lidar and camera sensors at the same time. The key module is the voxel binning (VPF) layer, which combines voxel and pixel features, uses appropriate mechanisms for fusion, and can be guided and enhanced by adjusting hyperparameters. Guo et al. [81] proposed a multi-layer fusion three-dimensional object detection network. In this method, the feature extractor with adaptive fusion module (AFM) is used to perform weighted fusion of different feature layers, which effectively reduces the interference of background on small targets. At the same time, AFM can filter information that is not relevant to the task, thereby improving the training efficiency of the network and speeding up the convergence speed.

Some other studies focus on learning better feature representations to improve detection performance. The Transformer architecture has been widely used in this field due to its powerful parallel computing and feature modeling capabilities. In the task of object detection, the Transformer can capture global features and long-distance dependencies through the self-attention mechanism, thereby achieving more accurate object positioning and classification. For example, DETR [82] (DEtection TRansformer) is the first algorithm to apply the Transformer to 2D object detection. It regards object detection as a set prediction problem and directly outputs the detection results in parallel through the Transformer encoder-decoder architecture. DETR simplifies the detection process and does not require traditional steps such as anchor generation or non-maximum suppression (NMS), but requires a large number of iterations to converge during training.

In the field of multimodal object detection, the Transformer architecture has also shown great potential. For example, Transformer-based frameworks such as CMT [83] can fuse multimodal data such as images and point clouds, significantly improving detection accuracy. In addition, multimodal pre-trained models learn universal representations through large-scale data, providing powerful feature extraction capabilities for

object detection. Transformers can also automatically learn the alignment relationship between modalities, reducing the complexity of data preprocessing. In terms of specific applications, the CAT-Det [84] network proposed by Zhang et al. uses the Transformer's ability to capture global features. It contains Transformer encoding branches for point clouds and images and a cross-modal Transformer module, and achieves effective feature learning through dual-branch encoding and cross-modal fusion. In addition, in order to solve the problem of information loss during modal conversion, Liu et al. [85] proposed a new framework for 3D object detection that iteratively updates radar and camera features through an interactive module. This module aggregates radar and image features with a set of sparse 3D object queries when sampling them, retaining the integrity of the original radar features to prevent information loss, thereby reducing association errors and improving detection accuracy.

In order to improve the robustness of fusion under inferior image conditions such as poor lighting, Wu et al. [86] generated pseudo-point clouds through depth completion, extracted their 2D image features and 3D geometric features, and adopted an effective ROI fusion strategy to obtain more accurate features for detection. Bai et al. proposed TransFusion based on Transformer structure [64], which uses a soft association mechanism to deal with the situation of inferior images, and adaptively determines the information that should be obtained from the images, so as to achieve better fusion results. In order to enhance the semantic information of the point cloud data, FusionPainting [61] uses a simple sequential multimodal fusion structure to project the point cloud onto the semantic segmented 2D image, and uses the existing 3D detector to realize the target detection.

5.1.3 Decision-Level Fusion

The advantages of decision-level fusion in multi-sensor information fusion are that it improves the fault tolerance and anti-interference of the system, enhances the integration ability of multi-source heterogeneous sensor data, reduces the data transmission and storage requirements, has good error correction ability, and has fast real-time response and adaptability. These features make decision-level convergence excellent in complex environments where fast and accurate decision-making is required.

The method of decision-level fusion is usually to perform 3D object detection on image data and point cloud data, and then merge the two results to obtain a better final output. Kim et al. [87] projected the sparse point cloud input into a dense front view, and with the help of Fast R-CNN [88], the area suggestion boxes were generated in the projection view of the image and the point cloud, respectively, and then the suggestion boxes on the two images were fused to form the final result. Huang et al. [89] proposed an Epnet network model, which converts the point cloud into a correlation depth map and a reflectivity density map, and uses YOLO to generate the detection results of the point cloud projection view and image, and then performs fusion operations. These methods reduce the dimensionality of point cloud data, resulting in information loss and unsatisfactory detection performance. Pang et al. proposed a low-complexity CLOCs [66] network, the structure of which is shown in Fig. 15. It combines the 2D and 3D suggestion boxes to extract features and learn their probabilistic correlations for fusion, which improves the detection performance under multi-modal data, but it needs to run 2D and 3D detectors at the same time, which is computationally expensive. In order to reduce the memory footprint, Pang et al. further proposed Fast-CLOCs [67] network, which removes the two-dimensional detector, extracts image features only with a lightweight network, projects the three-dimensional suggestion box into the two-dimensional and corrects it with image features, and finally fuses the two-dimensional and three-dimensional recommendation frames to achieve real-time detection. Guo et al. proposed the Liga-Stereo [90] model, which first extracted two-dimensional semantic features from binocular images to generate a suggestion box, and then projected them in three-dimensional space and combined with the point cloud three-dimensional detector to learn semantic features, realizing the

integration of semantic and geometric features, and improving the detection effect of large-scale and small-scale targets. Fan proposed Snow-CLOCs, a multimodal object detection algorithm specifically for snowy days [91]. The InceptionNeXt [92] network was used to enhance the image feature extraction ability of YOLOv5 [93], and the Wise IoU algorithm was used to reduce the dependence on high-quality data. On the basis of the SECOND [94] algorithm, the DROR filter is used to denoiser to improve the accuracy of lidar point cloud detection. Finally, the detection results of the camera and lidar are integrated into a detection set, and the sparse tensor representation and 2D convolutional neural network are used to extract features to achieve target detection and localization.

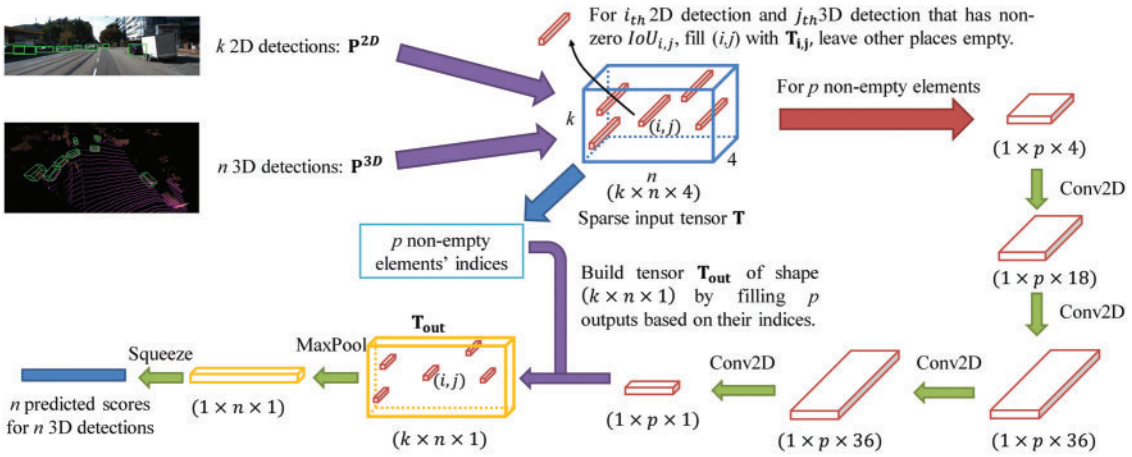

Figure 15: The architecture of CLOCs

5.1.4 BEV Fusion

BEV fusion technology provides an effective solution in the field of autonomous driving, simplifying the processing of three-dimensional space problems by converting multi-sensor data into a bird's-eye view, and its structure is illustrated in Fig. 16. This approach integrates information from different sensors, such as cameras, lidar, and radar, to provide a global view and enhance understanding of the vehicle's surroundings. BEV fusion simplifies object detection and segmentation tasks by representing three-dimensional spatial data on a two-dimensional plane, while improving the accuracy of positioning. In addition, it enhances the robustness of the system, as the fusion system provides reliable situational perception even when some sensor data is missing or inaccurate. BEVFusion [95] retains geometric and semantic information by stitching and fusing the shared image BEV features with the point cloud features, encoding the camera and lidar functions into the same BEV, ensuring system stability in the event of sensor failure. Fig. 16 shows the BEVFusion framework. The model performs single-modal feature extraction on lidar point cloud and camera images, respectively, and the network that extracts point cloud features generates sparse voxel representations, while the image feature extraction network encodes image features into multi-scale feature maps. Then, through geometric alignment, the image features are projected from a 2D plane into a BEV view, sharing a unified spatial representation with the point cloud features. In the BEV space, a fusion module is designed to integrate features from multiple modalities, and further improve the feature expression ability through cross-modal feature enhancement. The fused BEV features are used in the subsequent object detection head to generate a 3D bounding box and its associated properties (e.g., position, orientation, class, etc.).

In order to realize the joint capture of instance-level and scene-level context information, Yin et al. proposed a new multimodal fusion framework IS-FUSION by combining instance-level and scene-level

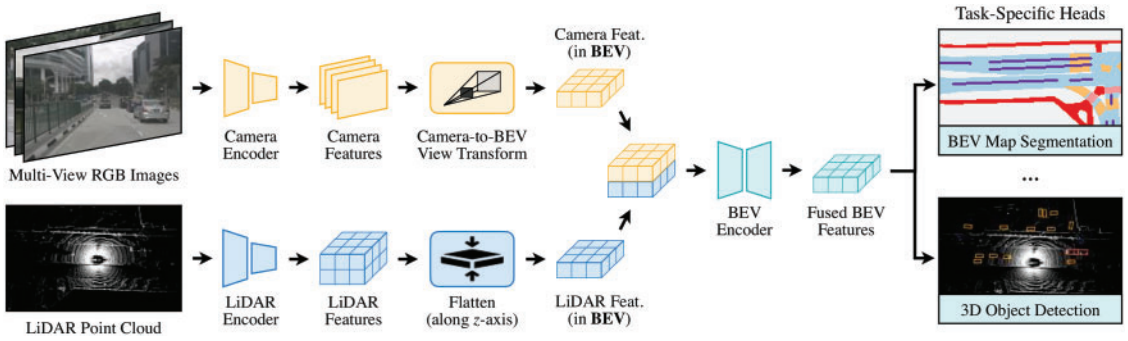

Figure 16: The architecture of BEVFusion

context information [96]. Fundamentally different from the existing BEV scene-level fusion methods, this method realizes BEV scene-level fusion by explicitly combining instance-level multimodal information, thereby simplifying the detection task of 3D objects. In order to enhance the interaction and feature guidance between lidar and camera, Li et al. [97] proposed a new multimodal 3D object detection method (GAFusion). By introducing sparse depth guidance (SDG) and lidar occupancy guidance (OHCHR), this method generates 3D features with rich depth information. At the same time, a lidar-guided adaptive fusion Transformer (LGAFT) was developed to adaptively enhance the interaction of BEV features of different modes from a global perspective, and realize the global interaction and adaptive feature fusion based on lidar. Zhao et al. proposed the Unibevfusion [98] model, which uses a shared module to extract BEV features across different modalities, integrates radar-specific data into the depth prediction process, and enhances the quality of visual bird's-eye view (BEV) features. Kim et al. proposed the RCM-Fusion [99] model, which attempts to fuse the two modes at the feature level and the instance level. By using a radar-guided BEV encoder, the camera features are converted into accurate BEV representations, and the combination of radar and camera BEV features is combined to reduce positioning errors. Zhao et al. proposed BEV-radar [100], a bidirectional fusion scheme for BEV radar that does not rely on the detection results of prior cameras. The method follows a BEV-based 3D detection approach, using a bidirectional converter to embed information from both modalities and enforce local spatial relationships based on subsequent convolutional blocks. After embedding the features, the BEV features are decoded by the 3D object prediction head. In order to overcome the difference between the foreground instance and the background area, and make full use of the depth information of the image to promote the precise alignment of the point cloud with the camera features, Hao et al. [101] proposed a coarse-to-fine image–point cloud fusion network for 3D object detection. In this method, a two-stage refinement strategy is used to process virtual point clouds, combined with dynamic 2D Gaussian distribution to achieve more accurate feature matching, and a dynamic density-aware RoI network is introduced for detailed feature extraction, so as to improve the performance and accuracy of object detection.

5.2 Camera-Radar Fusion

The Camera-Radar fusion technology combines the camera's high-resolution visual information with the all-weather and long-range detection capabilities of radar, providing a powerful solution for environmental perception. This convergence significantly enhances the performance of autonomous driving systems in adverse weather and complex lighting conditions, improving the accuracy of object detection, especially when dealing with occluded or long-range targets. Radar further improves the safety and reliability of the system due to its accurate speed and distance measurement capabilities and strong resistance to

electronic interference. This convergence approach is more cost-effective than lidar while maintaining superior performance, enabling autonomous vehicles to make safer and more reliable decisions in complex traffic environments.

However, due to the characteristics of radar, the point cloud information provided by it is relatively sparse, which limits the accuracy of target detection, and in order to overcome this limitation, Nabati proposed the CenterFusion [102] model, the structure of which is shown in Fig. 17. Firstly, the RGB image of the camera is used to generate the center point and its category prediction of the candidate target through the 2D object detector CenterNet [22], and the features of each center point are extracted. Subsequently, the radar point is projected onto the image plane, and according to the distance between its projection position and the detected 2D center point, the radar point is assigned to the corresponding candidate target area. Through the specially designed fusion module, the speed and range information of the radar are integrated with the center point feature of the image, so as to enhance the depth perception and motion information of the target. The fused features are used for subsequent 3D object detection to generate a 3D bounding box of the target, including properties such as position, orientation, and size. This method effectively realizes the deep fusion of radar and image information, significantly improves the detection performance in occlusion, long-distance and complex scenes, and has high real-time performance.

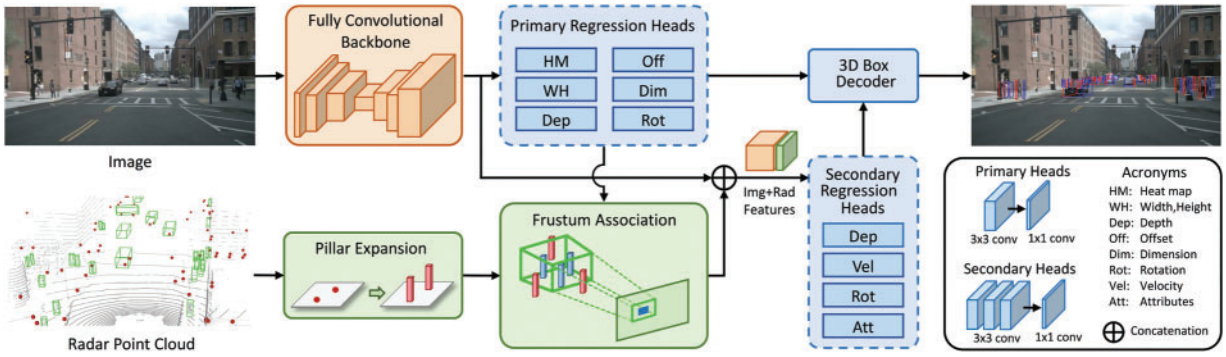

Figure 17: The architecture of CenterFusion

Based on the CenterFusion model, Shi proposed the CenRadfusion [103] network model. The model projects a radar point cloud onto the image plane and feeds it into the CenterNet inspection network as an additional channel to form a preliminary 3D inspection frame. Radar point clouds are processed through density-based clustering to improve data quality and eliminate irrelevant point clouds and noise, so as to enhance the reliability of target detection. Finally, the attention module of extrusion and excitation network is introduced to weight the feature channels and improve the influence of key features. This method not only ensures the integrity of the fusion architecture, but also enhances the network's ability to recognize important features through the attention mechanism, thereby improving the accuracy of 3D object detection.

In order to better fuse images and radar data to achieve more accurate object detection, Liu et al. [85] proposed a new framework to iteratively update the 3D object detection of radar and camera functions through interactive modules. The module samples radar and image features while aggregating them with a set of sparse 3D object queries, preserving the integrity of the original radar features to prevent information loss, thereby reducing correlation errors and improving detection accuracy. Kalgaonkar proposed an efficient camera-radar fusion network NeXtFusion [104], which uses the attention module method to enhance the representation of key features of target detection, while also minimizing the information loss from multi-modal data. Wang et al. proposed a network framework called MWRC3D [105], which improves the ability

to characterize image features by learning global associations and dependencies between individual pixels in an image through the attention-based Deep Aggregation (ADLA) module. The deformable convolutional network (DCN) is introduced to model the geometric transformation, and the data augmentation module is used to correct the 3D offset between the radar point cloud and the image center point. Finally, the image features are stitched and fused with the radar feature map as the input of the secondary return head to obtain an accurate 3D target detection frame.

In order to improve the correlation between images and point cloud data and make full use of the complementary characteristics of the two, Kong et al. [106] proposed a time-enhanced radar-camera fusion network model. In this model, a temporal fusion model is introduced to fuse radar features at different moments, thereby alleviating the problem of millimeter-wave sensor point target mismatch caused by object motion. In addition, this method proposes a new correlation-based fusion module, which uses mask cross-attention to more effectively fuse Radar and visual features. Lin et al. [107] proposed a high-precision millimeter-wave radar-camera fusion 3D perception network (RCBEVDet++). This method encodes sparse radar points into dense BEV features through a radar feature extractor, and introduces a deformation attention mechanism to align the BEV features of the radar and camera. Subsequently, multimodal features are further integrated through channel and spatial fusion layers. RCBEVDet++ not only supports 3D object detection, but also extends to BEV semantic segmentation and 3D multi-object tracking tasks, showing strong multi-task adaptability. Lei et al. [108] proposed the HVDet Fusion model. In order to fully utilize the advantages of radar signals, this method introduced prior information based on the positions of different objects to filter out false positive information in the original radar data. At the same time, the BEV features generated by the original camera data were supplemented and fused according to the positioning information and radial velocity information recorded by the radar sensor. During the fusion training process, these improvements further improved the detection effect.

In order to more intuitively compare the detection effects of different detection methods, we list the performance of these methods on the nuScenes dataset in Table 4. Among these detection methods, we found that the Lidar-based detection method FocalFormer3D achieved the highest mAP and the highest NDS. The detection methods based on cameras and millimeter-wave radars scored at lower levels in terms of mAP and NDS, mainly because the millimeter-wave radar has low accuracy and cannot generate accurate point cloud data.

Table 4: Detection effects of various detection methods in the nuScenes dataset

Method	Sparse 4D-v3 [33]	Far3D [34]	Focal Former 3D [39]	FSTR-L [42]	BEV Fusion [95]	Auto Align [72]	RCBEV Det++ [107]	HVDet Fusion [108]
Year	2023	2024	2023	2023	2022	2022	2024	2024
Modalities	Camera	Camera	Lidar	Lidar	Camera Lidar	Camera Lidar	Camera Radar	Camera Radar
mAP	0.668	0.635	0.705	0.702	0.702	0.684	0.673	0.609
mATE (m)	0.346	0.432	0.243	0.254	0.261	0.245	0.341	0.379
mASE (1-IOU)	0.234	0.237	0.238	0.266	0.239	0.233	0.234	0.243
mAOE (rad)	0.279	0.278	0.321	0.282	0.329	0.311	0.241	0.382
mAVE (m/s)	0.142	0.227	0.2	0.213	0.26	0.258	0.147	0.172
mAAE (1-acc)	0.145	0.130	0.130	0.131	0.134	0.113	0.13	0.132

(Continued)

Table 4 (continued)

Method	Sparse 4D-v3 [33]	Far3D [34]	Focal Former 3D [39]	FSTR-L [42]	BEV Fusion [95]	Auto Align [72]	RCBEV Det++ [107]	HVDet Fusion [108]
NDS	0.719	0.707	0.687	0.736	0.729	0.724	0.727	0.674

Note. mAP (mean Average Precision) represents the average detection accuracy of the model on different categories. mATE (mean Average Translation Error) measures the deviation between the center of the detection box and the center of the true box. mASE (mean Average Scale Error) is used to measure the scale difference between the detection box and the true box. mAOE (mean Average Orientation Error) is used to measure the orientation accuracy of the detection box. mAVE (mean Average Velocity Error) measures the accuracy of detecting the target speed. mAAE (mean Average Attribute Error) is used to measure the accuracy of target attribute prediction. NDS is the nuScenes Detection Score, which is used to comprehensively evaluate the performance of the detection model.

5.3 Camera-LiDAR-Radar Fusion

The Camera-Lidar-Radar fusion technology integrates the unique advantages of the three sensors to achieve a comprehensive improvement in environmental perception in the field of autonomous driving. The camera provides rich visual detail, the lidar accurately captures three-dimensional spatial information, and the radar maintains stable detection in bad weather, ensuring the system's all-weather capability. This fusion not only improves the accuracy and robustness of obstacle detection, but also optimizes the vehicle's decision-making and navigation capabilities. In addition, it enhances the system's performance in a wide range of lighting and weather conditions, improving safety, while reducing reliance on a single sensor failure, improving overall reliability and adaptability.

Camera-LiDAR-Radar fusion needs to process data from three different sensors at the same time, which puts forward higher requirements for the computing power of the processor and the real-time performance of the algorithm, especially in high-precision and low-latency application scenarios. In order to improve the accuracy and robustness of 3D object detection, Nobis et al. [109] proposed a 3D object detection network based on radar Voxel Fusion Net (RVF-Net), which is capable of fusing lidar, camera, and radar data. In this model, lidar points are first projected onto an image plane and combined with the camera image to create a simulated depth camera to generate 3D data points. Information from radar, lidar, and analog depth cameras is then combined into a discrete voxel framework. In order to effectively process the sparse voxel mesh input, RVF-Net adopts the sparse 3D convolution technology, which can improve the training efficiency of the network and accelerate the convergence speed. To solve the uncertainty problem in multi-target detection of autonomous vehicles, Ravindran et al. [110] proposed a Bayesian neural network (CLR-BNN) that fuses camera, lidar, and radar. The method first preprocesses the input data (such as camera images, lidar point clouds, and radar signals) to remove noise and extract key features. Then, the deep neural network is used to extract features from each modal data. In the Bayesian inference stage, the network weights are expressed in the form of probability distributions, and the weights are generated multiple times by variational inference or Monte Carlo sampling, and multiple prediction results and corresponding uncertainty estimates are output. The features of different modalities are integrated through the fusion module, so as to enhance the complementarity of information. The fused features are fed into the object detection head, which is used to predict the target class, the 3D bounding box, and the corresponding uncertainty. Through multiple forward propagations, BNN performs distribution statistics on the model output to generate the final detection results. Redundant test results were removed by non-maximum suppression and uncertainty estimation was further optimized using calibration curves. The final output includes the target class, 3D bounding box and its uncertainty, which can meet the needs of object detection and decision-making in autonomous driving scenarios.

In addition, the response time and safety of autonomous driving systems are also affected by the accuracy and reliability of multi-sensor fusion, which requires more efficient data fusion strategies. In order to better integrate the data of these three types of sensors, Li et al. [111] studied the advantages and disadvantages of each data mode, compared the performance of different fusion strategies, and proposed a simple and effective multimodal 3D object detection and tracking framework (EZFusion). This method pays special attention to the processing of radar data, fuses it with lidar and image data, and presents the fused data in the form of BEV for target detection and re-identification, which improves the detection performance and reduces the blind spot in the field of view. FishingNet [112] used a multilayer perceptron (MLP) for view transformation to handle the fusion of data from different sensors. Convert features from cameras, lidars, and radars into generic features and top-down semantic grid representations, which are then aggregated for downstream object detection tasks. The automatic labeling process is used to generate the training data through the 3D tracking bounding box and semantic labels, which improves the object detection ability of the algorithm. The FishingNet structure is shown in Fig. 18.

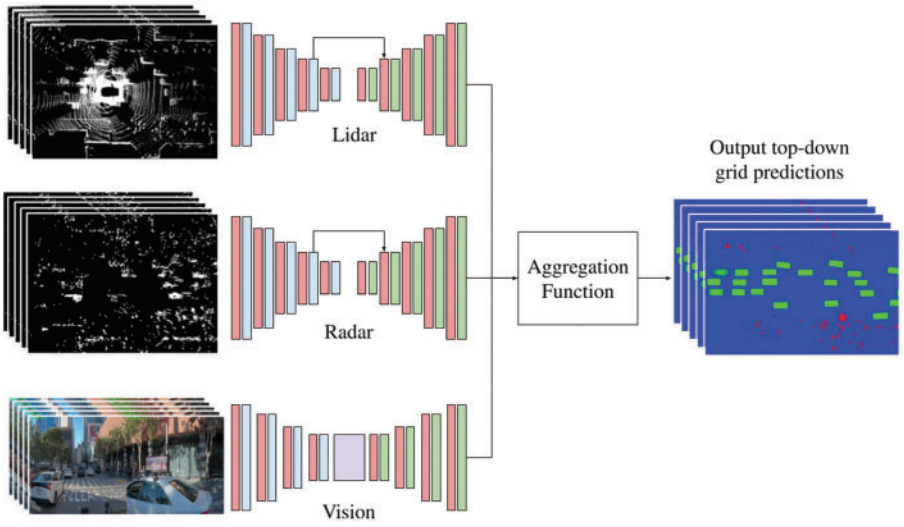

Figure 18: The architecture of fishing net

5.4 Other Sensor Fusion

In addition to the commonly used camera and radar fusion, there are also methods that focus on the fusion of other sensors such as drive test units, ultrasonic sensors, high-definition maps, etc. In order to solve the problem of insufficient single-sensor tracking accuracy caused by complex road environment and occlusion problems, Liu et al. [113] proposed a new multi-sensor data fusion method, which incorporates the drive test unit into the fusion strategy to improve the interactive perception level between road targets. This method introduces the image information of the drive test unit to solve the key problem of multi-sensor trajectory tracking data fusion. By analyzing the variation characteristics of the reflection intensity of the lidar point cloud and the high-precision detection ability of radar, the weight parameters of lidar and radar in the fusion process are dynamically determined. This method significantly improves the perception ability of the vehicle-road cooperation system, enhances the accuracy and reliability of target tracking, and provides important support for efficient object detection in complex traffic scenarios.

Fisheye cameras are commonly used for panoramic perception, but their performance is degraded in low light or strong sunlight. In contrast, ultrasonic sensors are stable under these conditions. Das et al. [114] proposed the first end-to-end multimodal fusion model, which combines a fisheye camera and an ultrasonic sensor to detect obstacles from a bird's-eye perspective. The model used ResNeXt-50 [115] to extract features, convert image features into BEV, and fuse them with ultrasonic features. Extended convolution is used to reduce sensor misalignment, and finally a two-level semantic occupancy decoder is used to generate grid prediction. The experimental results demonstrate the robustness and effectiveness of the method. Aniobi et al. [116] proposed a system based on sensor fusion that combines a camera with an ultrasonic sensor to detect an object and calculate its relative position relative to the vehicle. The system uses the YOLOv8 [117] object detection algorithm to detect and classify the objects in the camera's field of view, and combines the camera imaging formula with the bounding box information of YOLOv8 to design a positioning module for calculating the precise spatial coordinates of the objects. This method effectively improves the navigation accuracy and obstacle detection accuracy of unmanned vehicles, and provides support for the realization of safer and more reliable autonomous driving.

Infrared cameras are able to use the thermal radiation emitted by the object itself for imaging, which allows it to detect targets at night or in low visibility conditions, such as fog, rain and other bad weather. Infrared imaging technology is not limited by visible light conditions, so infrared cameras can provide richer target information when the performance of traditional optical sensors is limited, especially for detecting ambiguous or indistinguishable targets, especially small targets. In addition, the temperature information contained in the infrared image is abundant, which helps to highlight the target in complex backgrounds, improving the accuracy and robustness of detection. Therefore, infrared cameras play an irreplaceable role in object detection, especially in conditions where traditional vision systems are difficult to work. Zhong et al. [118] integrated infrared cameras into sensor suites for autonomous vehicles to cope with sensor limitations in adverse weather and low-visibility conditions such as nighttime, rain, snow, and haze, improving the ability to control vehicles and pedestrians. By combining infrared thermal imaging technology with the YOLOv5 [93] deep learning algorithm, a balance between detection accuracy and real-time performance is achieved, especially in challenging visibility conditions, which significantly improves target detection capabilities. Studies have shown that the integration of infrared thermography with YOLOv5 in ADAS can reduce the risk of accidents and improve road safety by providing more reliable scenario analysis at lower visibility.

High-definition maps can provide accurate geographic and environmental information in object detection, enhance feature extraction and sensor fusion, and improve the accuracy and robustness of detection. As prior knowledge, it assists the vehicle in maintaining performance in adverse weather and lighting conditions, reducing environmental impact. In addition, high-definition maps can reduce the search space, reduce the computational burden, improve the processing speed, and provide global positioning for the vehicle, support prediction and planning, and enhance scene understanding. Yang et al. [119] introduced high-definition map data (HD Map) into the object detection task and designed a single-stage detector that can extract geometric and semantic features from HD maps. In addition, a map prediction module is proposed, which can estimate the map in real time based on the original lidar data. This method proves that high-definition maps can provide powerful prior knowledge and improve the robustness of the network. Fang proposed the MapFusion [120] framework to integrate map information into the object detection network. The HD Map feature extraction, FeatureAgg fusion module, and mapseg detection module were designed, and the fusion of HD Map features, image and point cloud features was realized through the information conversion of different modules.

6 Summary and Outlook

Three-dimensional object detection is essential for autonomous driving technology, as it is not only the basis for perceiving the environment in autonomous driving systems, but also known as the eyes of autonomous vehicles. In this paper, we comprehensively review the 3D object detection methods based on multimodal fusion in recent years, and discuss the advantages and limitations of these methods. At the same time, this paper also introduces in detail the datasets and evaluation indexes used in the field of 3D object detection, discusses the shortcomings of the current field of 3D object detection in images, and puts forward the prospect of future research directions. With the continuous advancement of deep learning technology and autonomous driving technology, 3D object detection based on multimodal fusion has the potential to perform higher than single modal data due to its advantages of combining image and point cloud data. There is a growing need for this fusion approach in scenarios where detection performance is critical. We foresee that the research on multimodal fusion technology will continue to deepen, and the number of projects in its practical application will gradually increase, indicating that the future development of this field is full of hope.

Based on the analysis and summary of the relevant literature in recent years, this paper looks forward to the future research directions in the field of 3D object detection based on multimodal fusion, which can be divided into the following five directions:

(a) Convolutional operators suitable for non-Euclidean data feature extraction: Convolutional operators are the core of deep learning networks, and they perform well in the processing of regularized data such as images. However, the existing mainstream convolution operators are not suitable for non-Euclidean data, such as point clouds, which do not have a fixed grid structure in 3D space. At present, the research on feature extraction operators for non-Euclidean data is still in its infancy, but there are some preliminary results, such as models such as PointFormer [121] and GraphFormer [122]. These new operators are specifically designed to process non-Euclidean data such as point clouds, and can extract features directly from point clouds without having to convert them into regular data forms, thus preserving more spatial information. Therefore, the research and development of convolution operators suitable for non-Euclidean data will bring new breakthroughs in the field of 3D object detection and is an important way to improve detection performance.

(b) Exploration of semi-supervised learning and unsupervised learning: Compared with other visual fields, there are relatively few studies on semi-supervised and unsupervised learning in the field of 3D object detection, which limits the development of this field. Semi-supervised learning avoids the performance degradation of traditional supervised learning when the training samples are insufficient. In the field of 3D object detection, semi-supervised learning can use a small number of labeled samples and a large number of unlabeled samples to jointly train stronger models. Unsupervised learning does not rely on pre-labeled labels, it trains on unlabeled datasets to mine structures and patterns in the data to improve detection performance. Therefore, it is of great significance to explore the application of semi-supervised and unsupervised learning in 3D object detection.

(c) Multimodal object detection in polar coordinates: Traditional object detection uses the Cartesian coordinate system to characterize the target position, and although it excels in 2D object detection, it limits the potential of the sensor in 3D object detection. For lidar, the Cartesian coordinate system fails to make full use of the non-uniform spatial distribution of point clouds, resulting in low feature extraction efficiency. In the case of cameras, the physical world imaging of a surround camera has a wedge-shaped geometry and the axis is basically not perpendicular. In the Cartesian coordinate system, the camera image encoding ignores the symmetry of the target at different viewing angles and the loss of information when the convolutional kernel is downsampled, which is not conducive to feature fusion and accurate object detection. In contrast, polar coordinate systems have shown significant advantages in 3D object detection, but current research

has mainly focused on single-modal sensors. The camera-lidar fusion method in the polar coordinate system has not been widely studied, but theoretically, this method has obvious advantages. Therefore, the design of camera-lidar fusion method in polar coordinate system is expected to become a new trend in the development of fusion perception.

(d) Adaptive implicit multi-modal spatial fusion: Current multi-modal spatial fusion methods often project lidar point clouds to camera images, but this process is easy to cause geometric distortion, resulting in incorrect matching of 3D point cloud features to 2D images, resulting in semantic errors. The existing methods based on projection mapping are linear, explicit and absolute, but in fact, the correspondence between point clouds and pixels should be nonlinear, implicit and relative. This mandatory 3D-to-2D mapping can lead to position mapping errors and geometric distortions, affecting the quality of multimodal feature alignment and reducing detection accuracy. In the future, the lidar point cloud features and 3D position information can be used as a priori to allow the network to learn the coordinate correspondence between different modalities, realize adaptive feature fusion and alignment, and avoid the loss of feature information in the process of converting camera images into bird's-eye views. This method can improve the accuracy of multimodal fusion, thereby improving the accuracy of object detection.

(e) Efficient real-time detection: Efficient real-time detection is an important development direction of multimodal object detection, especially in scenarios that require low latency, such as autonomous driving and intelligent surveillance. With model lightweighting techniques such as quantization and model distillation, it is possible to reduce computational costs while maintaining accuracy. Efficient intermodal fusion strategies, such as dynamic weighted fusion and cross-modal attention mechanisms, can significantly reduce the complexity of multimodal processing. At the same time, optimizing network architectures (such as single-stage detectors and sparse networks) and data flows (such as block processing and real-time ROI extraction) can improve inference efficiency. In addition, hardware acceleration (GPU, FPGA, ASIC) and software optimization (model compilation, memory management) are combined to achieve software and hardware synergy to improve performance. In practical applications, real-time performance evaluations (such as frame rate testing, latency analysis) ensure that the system meets the requirements of real-time. Future development should continue to balance accuracy and efficiency, combined with emerging hardware and algorithms, to further improve the detection speed and robustness in complex scenarios.

Acknowledgement: Thank you to the editor and anonymous reviewer for their insightful comments, which have improved the quality of this publication.

Funding Statement: This research was funded by the Yangtze River Delta Science and Technology Innovation Community Joint Research Project (2023CSJGG1600), the Natural Science Foundation of Anhui Province (2208085MF173) and Wuhu "ChiZhu Light" Major Science and Technology Project (2023ZD01, 2023ZD03).

Author Contributions: Study conception and design: Peicheng Shi, Li Yang; data collection: Xinlong Dong, Heng Qi; analysis and interpretation of results: Peicheng Shi, Aixi Yang; draft manuscript preparation: Li Yang, Xinlong Dong. All authors reviewed the results and approved the final version of the manuscript.

Availability of Data and Materials: The authors confirm that the data supporting the findings of this study are available within the article.

Ethics Approval: Not applicable.

Conflicts of Interest: The authors declare no conflicts of interest to report regarding the present study.

References

1. Atakishiyev S, Salameh M, Yao H, Goebel R. Explainable artificial intelligence for autonomous driving: a comprehensive overview and field guide for future research directions. IEEE Access. 2024;12(3):101603–25. doi:10.1109/ACCESS.2024.3431437.
2. Wang Y, Mao Q, Zhu H, Deng J, Zhang Y, Ji J, et al. Multi-modal 3D object detection in autonomous driving: a survey. Int J Comput Vis. 2023;131(8):2122–52. doi:10.1007/s11263-023-01784-z.
3. Qi H, Shi P, Liu Z, Yang A. TSF: two-stage sequential fusion for 3D object detection. IEEE Sens J. 2022;22(12):12163–72. doi:10.1109/JSEN.2022.3175192.
4. Wang Z, Liu J. A review of object detection based on convolutional neural network. In: 2017 36th Chinese Control Conference (CCC); 2017 July 26–28; Dalian, China: IEEE; 2017. p. 11104–9.
5. Li Y, Miao N, Ma L, Shuang F, Huang X. Transformer for object detection: review and benchmark. Eng Appl Artif Intell. 2023;126(4):107021. doi:10.1016/j.engappai.2023.107021.
6. Zhao ZQ, Zheng P, Xu ST, Wu X. Object detection with deep learning: a review. IEEE Trans Neural Netw Learn Syst. 2019;30(11):3212–32. doi:10.1109/TNNLS.2018.2876865.
7. Geiger A, Lenz P, Urtasun R. Are we ready for autonomous driving? The KITTI vision benchmark suite. In: 2012 IEEE Conference on Computer Vision and Pattern Recognition; 2012 Jun 16–21; Providence, RI, USA: IEEE; 2012. p. 3354–61. doi:10.1109/CVPR.2012.6248074.
8. Caesar H, Bankiti V, Lang AH, Vora S, Liong VE, Xu Q, et al. nuScenes: a multimodal dataset for autonomous driving. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2020 Jun 13–19; Seattle, WA, USA. NY, USA: IEEE; 2020. p. 11618–28.
9. Sun P, Kretzschmar H, Dotiwalla X, Chouard A, Patnaik V, Tsui P, et al. Scalability in perception for autonomous driving: waymo open dataset. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2020 Jun 13–19; Seattle, WA, USA: IEEE; 2020. p. 2443–51. doi:10.1109/cvpr42600.2020.00252.
10. Mao J, Niu M, Jiang C, Liang H, Chen J, Liang X, et al. One million scenes for autonomous driving: once dataset. arXiv:2106.11037. 2021.
11. Chen L, Sima C, Li Y, Zheng Z, Xu J, Geng X, et al. PersFormer: 3D lane detection via perspective transformer and the OpenLane benchmark. In: Computer vision–ECCV 2022; Cham: Springer Nature Switzerland; 2022. p. 550–67.
12. Wang T, Kim S, Wenxuan J, Xie E, Ge C, Chen J, et al. A motion and accident prediction benchmark for V2X autonomous driving. In: Proceedings of the AAAI Conference on Artificial Intelligence; 2024 Feb 20–27; Washington, DC, USA. CA, USA: AAAI Press; 2024. p. 5599–606.
13. Huang X, Wang P, Cheng X, Zhou D, Geng Q, Yang R. The ApolloScape open dataset for autonomous driving and its application. IEEE Trans Pattern Anal Mach Intell. 2020;42(10):2702–19. doi:10.1109/TPAMI.2019.2926463.
14. Houston J, Zuidhof G, Bergamini L, Ye Y, Chen L, Jain A, et al. One thousand and one hours: self-driving motion prediction dataset. arXiv:2006.14480. 2020.
15. Pham QH, Sevestre P, Pahwa RS, Zhan H, Pang CH, Chen Y, et al. A*3D dataset: towards autonomous driving in challenging environments. In: 2020 IEEE International Conference on Robotics and Automation (ICRA); 2020 May 31–Aug 31; Paris, France: IEEE. p. 2267–73. doi:10.1109/ICRA40945.2020.9197385.
16. Patil A, Malla S, Gang H, Chen YT. The H3D dataset for full-surround 3D multi-object detection and tracking in crowded urban scenes. In: 2019 International Conference on Robotics and Automation (ICRA); 2019 May 20–24; Montreal, QC, Canada: IEEE. p. 9552–7. doi:10.1109/icra.2019.8793925.
17. Geyer J, Kassahun Y, Mahmudi M, Ricou X, Durgesh R, Chung AS, et al. A2D2: audi autonomous driving dataset. arXiv:2004.06320. 2020.
18. Gählert N, Jourdan N, Cordts M, Franke U, Denzler J. Cityscapes 3D: dataset and benchmark for 9 DoF vehicle detection. arXiv:2006.07864. 2020.
19. Chang MF, Ramanan D, Hays J, Lambert J, Sangkloy P, Singh J, et al. Argoverse: 3D tracking and forecasting with rich maps. In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2019 Jun 15–20; Long Beach, CA, USA: IEEE; 2019. doi:10.1109/cvpr.2019.00895.
20. Weng X, Man Y, Cheng D, Yuan Y, O'Toole M, Kitani K. All-in-one drive: a large-scale comprehensive perception dataset with high-density long-range point clouds; 2020. doi:10.13140/RG.2.2.21621.81122.

21. Liao Y, Xie J, Geiger A. KITTI-360: a novel dataset and benchmarks for urban scene understanding in 2D and 3D. IEEE Trans Pattern Anal Mach Intell. 2023;45(3):3292–310. doi:10.1109/TPAMI.2022.3179507.
22. Duan K, Bai S, Xie L, Qi H, Huang Q, Tian Q. CenterNet: keypoint triplets for object detection. In: Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV); 2019 Oct 27–Nov 02; Seoul, Republic of Korea. NY, USA: IEEE; 2019. p. 6569–78.
23. Brazil G, Liu X. M3D-RPN: monocular 3D region proposal network for object detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV); 2019 Oct 27–Nov 02; Seoul, Republic of Korea. NY, USA: IEEE; 2019. p. 9287–96.
24. Manhardt F, Kehl W, Gaidon A. ROI-10D: monocular lifting of 2D detection to 6D pose and metric shape. In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2019 Jun 15–20; Long Beach, CA, USA: IEEE; 2019. p. 2064–73. doi:10.1109/cvpr.2019.00217.
25. Sun X, Wu P, Hoi SCH. Face detection using deep learning: an improved faster RCNN approach. Neurocomputing. 2018;299(2):42–50. doi:10.1016/j.neucom.2018.03.030.
26. Chen X, Kundu K, Zhang Z, Ma H, Fidler S, Urtasun R. Monocular 3D object detection for autonomous driving. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 2016 Jun 27–30; Las Vegas, NV, USA: IEEE; 2016. p. 2147–56. doi:10.1109/CVPR.2016.236.
27. Chabot F, Chaouch M, Rabarisoa J, Teulière C, Chateau T. Deep MANTA: a coarse-to-fine many-task network for joint 2D and 3D vehicle analysis from monocular image. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 2017 Jul 21–26; Honolulu, HI, USA: IEEE; 2017. p. 1827–36. doi:10.1109/CVPR.2017.198.
28. He T, Soatto S. Mono3D++: monocular 3D vehicle detection with two-scale 3D hypotheses and task priors. Proc AAAI Conf Artif Intell. 2019;33(1):8409–16. doi:10.1609/aaai.v33i01.33018409.
29. Kundu A, Li Y, Rehg JM. 3D-RCNN: instance-level 3D object reconstruction via render-and-compare. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2018 Jun 18–23; Salt Lake City, UT, USA: IEEE; 2018. p. 3559–68. doi:10.1109/CVPR.2018.00375.
30. Sun J, Chen L, Xie Y, Zhang S, Jiang Q, Zhou X, et al. Disp R-CNN: stereo 3D object detection via shape prior guided instance disparity estimation. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2020 Jun 13–19; Seattle, WA, USA: IEEE; 2020. p. 10545–54. doi:10.1109/cvpr42600.2020.01056.
31. Li P, Chen X, Shen S. Stereo R-CNN based 3D object detection for autonomous driving. In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2019 Jun 15–20; Long Beach, CA, USA: IEEE; 2019. p. 7636–44. doi:10.1109/cvpr.2019.00783.
32. Qin Z, Wang J, Lu Y. Triangulation learning network: from monocular to stereo 3D object detection. In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2019 Jun 15–20; Long Beach, CA, USA: IEEE. p. 7607–15. doi:10.1109/cvpr.2019.00780.
33. Lin X, Pei Z, Lin T, Huang L, Su Z. Sparse4D v3: advancing end-to-end 3D detection and tracking. arXiv:2311.11722. 2023.
34. Jiang X, Li S, Liu Y, Wang S, Jia F, Wang T, et al. Far3D: expanding the horizon for surround-view 3D object detection. Proc AAAI Conf Artif Intell. 2024;38(3):2561–9. doi:10.1609/aaai.v38i3.28033.
35. Li B, Zhang T, Xia T. Vehicle detection from 3D lidar using fully convolutional network. arXiv:1608.07916. 2016.
36. Beltrán J, Guindel C, Moreno FM, Cruzado D, García F, De La Escalera A. BirdNet: a 3D object detection framework from LiDAR information. In: 2018 21st International Conference on Intelligent Transportation Systems (ITSC); 2018 Nov 4–7; Maui, HI, USA: IEEE; 2018. p. 3517–23. doi:10.1109/ITSC.2018.8569311.
37. Barrera A, Guindel C, Beltran J, Garcia F. BirdNet+: end-to-end 3D object detection in LiDAR bird's eye view. In: 2020 IEEE 23rd International Conference on Intelligent Transportation Systems (ITSC); 2020 Sep 20–23; Greece: Rhodes; 2020. p. 1–6. doi:10.1109/itsc45102.2020.9294293.
38. Zhou Y, Sun P, Zhang Y, Anguelov D, Gao J, Ouyang TY, et al. End-to-end multi-view fusion for 3D object detection in LiDAR point clouds. In: Proceedings of the Conference on Robot Learning; 2020 Oct 30–Nov 01; Osaka, Japan. NY, USA: PMLR; 2020. p. 923–32.

39. Chen Y, Yu Z, Chen Y, Lan S, Anandkumar A, Jia J, et al. FocalFormer3D: focusing on hard instance for 3D object detection. In: 2023 IEEE/CVF International Conference on Computer Vision (ICCV); 2023 Oct 1–6; Paris, France: IEEE; 2023. p. 8360–71. doi:10.1109/ICCV51070.2023.00771.
40. Zhou Y, Tuzel O. VoxelNet: end-to-end learning for point cloud based 3D object detection. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2018 Jun 18–23; Salt Lake City, UT, USA: IEEE; 2018. p. 4490–9. doi:10.1109/CVPR.2018.00472.
41. He C, Zeng H, Huang J, Hua XS, Zhang L. Structure aware single-stage 3D object detection from point cloud. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2020 Jun 13–19; Seattle, WA, USA: IEEE; 2020. p. 11870–79. doi:10.1109/cvpr42600.2020.01189.
42. Zhang D, Zheng Z, Niu H, Wang X, Liu X. Fully sparse transformer 3-D detector for LiDAR point cloud. IEEE Trans Geosci Remote Sens. 2023;61:5705212. doi:10.1109/TGRS.2023.3328929.
43. Shi S, Wang X, Li H. PointRCNN: 3d object proposal generation and detection from point cloud. In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2019 Jun 15–20; Long Beach, CA, USA: IEEE. p. 770–9. doi:10.1109/cvpr.2019.00086.
44. Qi CR, Yi L, Su H, Guibas LJ. PointNet++: deep hierarchical feature learning on point sets in a metric space. arXiv:1706.02413. 2017.
45. Yang Z, Sun Y, Liu S, Jia J. 3DSSD: point-based 3D single stage object detector. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2020 Jun 13–19; Seattle, WA, USA: IEEE; 2020. p. 11037–45. doi:10.1109/cvpr42600.2020.01105.
46. Lang AH, Vora S, Caesar H, Zhou L, Yang J, Beijbom O. PointPillars: fast encoders for object detection from point clouds. In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2019 Jun 15–20; Long Beach, CA, USA: IEEE; 2019. p. 12689–97. doi:10.1109/CVPR.2019.01298.
47. Charles RQ, Hao S, Mo K, Guibas LJ. PointNet: deep learning on point sets for 3D classification and segmentation. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 2017 Jul 21–26; Honolulu, HI, USA: IEEE; 2017. p. 77–85. doi:10.1109/CVPR.2017.16.
48. Wang Y, Fathi A, Kundu A, Ross DA, Pantofaru C, Funkhouser T, et al. Pillar-based object detection for autonomous driving. In: Proceedings of the 16th European Conference on Computer Vision; 2020; Glasgow: Springer. p. 18–34.
49. Shi W, Rajkumar R. Point-GNN: graph neural network for 3D object detection in a point cloud. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2020 Jun 13–19; Seattle, WA, USA: IEEE; 2020. p. 1708–16. doi:10.1109/cvpr42600.2020.00178.
50. He Q, Wang Z, Zeng H, Zeng Y, Liu Y. SVGA-net: sparse voxel-graph attention network for 3D object detection from point clouds. Proc AAAI Conf Artif Intell. 2022;36(1):870–8. doi:10.1609/aaai.v36i1.19969.
51. Wang L, Song Z, Zhang X, Wang C, Zhang G, Zhu L, et al. SAT-GCN: self-attention graph convolutional network-based 3D object detection for autonomous driving. Knowl Based Syst. 2023;259(99):110080. doi:10.1016/j.knosys.2022.110080.
52. Naiden A, Paunescu V, Kim G, Jeon B, Leordeanu M. Shift R-CNN: deep monocular 3D object detection with closed-form geometric constraints. In: IEEE International Conference on Image Processing (ICIP); 2019 Sep 22–25; Taipei, China: IEEE; 2019. p. 61–5. doi:10.1109/icip.2019.8803397.
53. You YR, Wang Y, Chao WL, Garg D, Pleiss G, Hariharan B, et al. Pseudo- LiDAR++: accurate depth for 3D object detection in autonomous driving. In: Proceedings of the 8th International Conference on Learning Representations; 2020; Addis Ababa: ICLR.
54. Li C, Ku J, Waslander SL. Confidence guided stereo 3D object detection with split depth estimation. In: 2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS); 2021 Oct 24 2020–Jan 24; Las Vegas, NV, USA: IEEE; 2020. p. 5776–83. doi:10.1109/iros45743.2020.9341188.
55. Qi CR, Liu W, Wu C, Su H, Guibas LJ. Frustum PointNets for 3D object detection from RGB-D data. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2018 Jun 18–23; Salt Lake City, UT, USA: IEEE; 2018. p. 918–27. doi:10.1109/CVPR.2018.00102.

56. Wang Z, Jia K. Frustum ConvNet: sliding Frustums to aggregate local point-wise features for amodal 3D object detection. In: 2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS); 2019 Nov 3–8; Macau, China: IEEE; 2019. p. 1742–49. doi:10.1109/iros40897.2019.8968513.
57. Paigwar A, Sierra-Gonzalez D, Erkent Ö., Laugier C. Frustum-PointPillars: a multi-stage approach for 3D object detection using RGB camera and LiDAR. In: 2021 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW); 2021 Oct 11–17; Montreal, BC, Canada: IEEE; 2021. p. 2926–33.
58. Chen X, Ma H, Wan J, Li B, Xia T. Multi-view 3D object detection network for autonomous driving. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 2017 Jul 21–26; Honolulu, HI, USA: IEEE; 2017. p. 6526–34. doi:10.1109/CVPR.2017.691.
59. Xie L, Xiang C, Yu Z, Xu G, Yang Z, Cai D, et al. PI-RCNN: an efficient multi-sensor 3D object detector with point-based attentive cont-conv fusion module. Proc AAAI Conf Artif Intell. 2020;34(7):12460–7. doi:10.1609/aaai.v34i07.6933.
60. Lu H, Chen X, Zhang G, Zhou Q, Ma Y, Zhao Y. Scanet: spatial-channel attention network for 3D object detection. In: ICASSP 2019–2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP); 2019 May 12–17; Brighton, UK: IEEE; 2019. p. 1992–6. doi:10.1109/icassp.2019.8682746.
61. Xu S, Zhou D, Fang J, Yin J, Bin Z, Zhang L. FusionPainting: multimodal fusion with adaptive attention for 3D object detection. In: 2021 IEEE International Intelligent Transportation Systems Conference (ITSC); 2021 Sep 19–22; Indianapolis, IN, USA: IEEE. p. 3047–54. doi:10.1109/itsc48978.2021.9564951.
62. Yang J, Lu J, Lee S, Batra D, Parikh D. Graph R-CNN for scene graph generation. In: Proceedings of the European Conference on Computer Vision (ECCV2018); 2018 Sep 8–14; Munich, Germany. Berlin, Germany: Springer. 2018. p. 670–85.
63. Xu X, Dong S, Xu T, Ding L, Wang J, Jiang P, et al. FusionRCNN: lidar-camera fusion for two-stage 3D object detection. Remote Sens. 2023;15(7):1839. doi:10.3390/rs15071839.
64. Bai X, Hu Z, Zhu X, Huang Q, Chen Y, Fu H, et al. TransFusion: robust LiDAR-camera fusion for 3D object detection with transformers. In: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2022 Jun 18–24; New Orleans, LA, USA: IEEE; 2022. p. 1080–9. doi:10.1109/CVPR52688.2022.00116.
65. Vaswanic A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention is all you need. In: Proceedings of the 31st Conference on Neural Information Processing Systems (NIPS 2017); 2017 Dec 4–9; Long Beach, CA, USA. NY, USA: Curran Associates Inc.; 2017. p. 6000–10.
66. Pang S, Morris D, Radha H. CLOCs: camera-LiDAR object candidates fusion for 3D object detection. In: 2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS); 2021 Oct 24 2020–Jan 24; Las Vegas, NV, USA: IEEE; 2020. p. 10386–93. doi:10.1109/iros45743.2020.9341791.
67. Pang S, Morris D, Radha H. Fast-CLOCs: fast camera-LiDAR object candidates fusion for 3D object detection. In: 2022 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV); 2022 Jan 3–8; Waikoloa, HI, USA: IEEE; 2022. p. 3747–56. doi:10.1109/WACV51458.2022.00380.
68. Cao P, Chen H, Zhang Y, Wang G. Multi-view frustum pointnet for object detection in autonomous driving. In: IEEE International Conference on Image Processing (ICIP); 2019 Sep 22–25; Taipei, China: IEEE; 2019. p. 3896–9. doi:10.1109/icip.2019.8803572.
69. He W, Deng Z, Ye Y, Pan P. ConCs-fusion: a context clustering-based radar and camera fusion for three-dimensional object detection. Remote Sens. 2023;15(21):5130. doi:10.3390/rs15215130.
70. Ku J, Mozifian M, Lee J, Harakeh A, Waslander SL. Joint 3D proposal generation and object detection from view aggregation. In: 2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS); 2018 Oct 1–5; Madrid, Spain: IEEE; 2018. p. 1–8. doi:10.1109/IROS.2018.8594049.
71. Liu Y, Suo C, Liu Z, Liu YH. A multi-sensor fusion based 2D-driven 3D object detection approach for large scene applications. In: 2019 IEEE International Conference on Robotics and Biomimetics (ROBIO); 2019 Dec 6–8; Dali, China: IEEE; 2019. p. 2181–8. doi:10.1109/robio49542.2019.8961637.
72. Chen Z, Li Z, Zhang S, Fang L, Jiang Q, Zhao F, et al. AutoAlign: pixel-instance feature aggregation for multi-modal 3D object detection. arXiv:2201.06493. 2022.

73. Chen Z, Li Z, Zhang S, Fang L, Jiang Q, Zhao F. AutoAlignV2: deformable feature aggregation for dynamic multi-modal 3D object detection. arXiv:2207.10316. 2022.
74. Alaba SY, Gurbuz AC, Ball JE. Emerging trends in autonomous vehicle perception: multimodal fusion for 3D object detection. World Electr Veh J. 2024;15(1):20. doi:10.3390/wevj15010020.
75. Xu S, Li F, Song Z, Fang J, Wang S, Yang ZX. Multi-sem fusion: multimodal semantic fusion for 3-D object detection. IEEE Trans Geosci Remote Sens. 2024;62(1):5703114. doi:10.1109/TGRS.2024.3387732.
76. Li Y, Chen Y, Qi X, Li Z, Sun J, Jia J. Unifying voxel-based representation with transformer for 3d object detection. Adv Neural Inf Process Syst. 2022;35:18442–55.
77. Song Z, Zhang G, Xie J, Liu L, Jia C, Xu S, et al. VoxelNextFusion: a simple, unified and effective voxel fusion framework for multi-modal 3D object detection. arXiv:2401.02702. 2024.
78. Liu M, Jia Y, Lyu Y, Dong Q, Yang Y. BAFusion: bidirectional attention fusion for 3D object detection based on LiDAR and camera. Sensors. 2024;24(14):4718. doi:10.3390/s24144718.
79. Deng J, Zhang S, Dayoub F, Ouyang W, Zhang Y, Reid I. PoIFusion: multi-modal 3D object detection via fusion at points of interest. arXiv:2403.09212. 2024.
80. Wang CH, Chen HW, Chen Y, Hsiao PY, Fu LC. VoPiFNet: voxel-pixel fusion network for multi-class 3D object detection. IEEE Trans Intell Transp Syst. 2024;25(8):8527–37. doi:10.1109/TITS.2024.3392783.
81. Guo Y, Hu H. Multi-layer fusion 3D object detection via lidar point cloud and camera image. Appl Sci. 2024;14(4):1348. doi:10.3390/app14041348.
82. Zhu X, Su W, Lu L, Li B, Wang X, Dai J, et al. Deformable DETR: deformable transformers for end-to-end object detection. arXiv:2010.04159. 2020.
83. Yan J, Liu Y, Sun J, Jia F, Li S, Wang T, et al. Cross modal transformer: towards fast and robust 3D object detection. In: 2023 IEEE/CVF International Conference on Computer Vision (ICCV); 2023 Oct 1–6; Paris, France: IEEE; 2023. p. 18222–32. doi:10.1109/ICCV51070.2023.01675.
84. Zhang Y, Chen J, Huang D. CAT-det: contrastively augmented transformer for multimodal 3D object detection. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2022 Jun 18–24; New Orleans, LA, USA: IEEE; 2022. p. 898–907. doi:10.1109/CVPR52688.2022.00098.
85. Liu X, Li Z, Zhou Y, Peng Y, Luo J. Camera-radar fusion with modality interaction and radar Gaussian expansion for 3D object detection. Cyborg Bionic Syst. 2024;5(11):79. doi:10.34133/cbsystems.0079.
86. Wu X, Peng L, Yang H, Xie L, Huang C, Deng C, et al. Sparse fuse dense: towards high quality 3D detection with depth completion. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2022 Jun 18–24; New Orleans, LA, USA. NY, USA: IEEE; 2022. p. 5418–27.
87. Kim T, Ghosh J. Robust detection of non-motorized road users using deep learning on optical and LIDAR data. In: 2016 IEEE 19th International Conference on Intelligent Transportation Systems (ITSC); 2016 Nov 1–4. p. 271–6. doi:10.1109/ITSC.2016.7795566.
88. Ren S, He K, Girshick R, Sun J. Faster R-CNN: towards real-time object detection with region proposal networks. IEEE Trans Pattern Anal Mach Intell. 2017;39(6):1137–49. doi:10.1109/TPAMI.2016.2577031.
89. Huang T, Liu Z, Chen X, Bai X. EPNet: enhancing point features with image semantics for 3D object detection. In: Computer vision–ECCV 2020. Cham: Springer International Publishing; 2020. p. 35–52. doi:10.1007/978-3-030-58555-6_3.
90. Guo X, Shi S, Wang X, Li H. LIGA-stereo: learning LiDAR geometry aware representations for stereo-based 3D detector. In: 2021 IEEE/CVF International Conference on Computer Vision (ICCV); 2021 Oct 10–17; Montreal, QC, Canada: IEEE; 2021. p. 3133–43. doi:10.1109/ICCV48922.2021.00314.
91. Fan X, Xiao D, Li Q, Gong R. Snow-CLOCs: camera-LiDAR object candidate fusion for 3D object detection in snowy conditions. Sensors. 2024;24(13):4158. doi:10.3390/s24134158.
92. Yu W, Zhou P, Yan S, Wang X. InceptionNeXt: when inception meets ConvNeXt. In: 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2024 Jun 16–22; Seattle, WA, USA: IEEE; 2024. p. 5672–83. doi:10.1109/CVPR52733.2024.00542.
93. Jocher G, Chaurasia A, Stoken A, Borovec J, Kwon Y, Michael K, et al. ultralytics/yolov5: v7. 0-yolov5 sota realtime instance segmentation. Zenodo. 2022. doi:10.5281/zenodo.3908559.

94. Yan Y, Mao Y, Li B. SECOND: sparsely embedded convolutional detection. Sensors. 2018;18(10):3337. doi:10.3390/s18103337.
95. Liu Z, Tang H, Amini A, Yang X, Mao H, Rus DL, et al. BEVFusion: multi-task multi-sensor fusion with unified bird's-eye view representation. In: 2023 IEEE International Conference on Robotics and Automation (ICRA); 2023 May 29–June 2; London, UK: IEEE; 2023. p. 2774–81. doi:10.1109/ICRA48891.2023.10160968.
96. Yin J, Shen J, Chen R, Li W, Yang R, Frossard P, et al. IS-fusion: instance-scene collaborative fusion for multimodal 3D object detection. In: 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2024 Jun 16–22; Seattle, WA, USA: IEEE; 2024. p. 14905–15. doi:10.1109/CVPR52733.2024.01412.
97. Li X, Fan B, Tian J, Fan H. GAFusion: adaptive fusing LiDAR and camera with multiple guidance for 3D object detection. In: 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2024 Jun 16–22; Seattle, WA, USA: IEEE; 2024. p. 21209–18. doi:10.1109/CVPR52733.2024.02004.
98. Zhao H, Guan R, Wu T, Man KL, Yu L, Yue Y. UniBEVFusion: unified radar-vision BEVFusion for 3D object detection. arXiv:2409.14751. 2024.
99. Kim J, Seong M, Bang G, Kum D, Choi JW. RCM-fusion: radar-camera multi-level fusion for 3D object detection. In: 2024 IEEE International Conference on Robotics and Automation (ICRA); 2024 May 13–17; Yokohama, Japan: IEEE; 2024. p. 18236–42. doi:10.1109/ICRA57147.2024.10611449.
100. Zhao Y, Zhang L, Deng J, Zhang Y. BEV-radar: bidirectional radar-camera fusion for 3D object detection. Justc. 2024;54(1):101. doi:10.52396/JUSTC-2023-0006.
101. Hao M, Zhang Z, Li L, Dong K, Cheng L, Tiwari P, et al. Coarse to fine-based image-point cloud fusion network for 3D object detection. Inf Fusion. 2024;112(7):102551. doi:10.1016/j.inffus.2024.102551.
102. Nabati R, Qi H. CenterFusion: center-based radar and camera fusion for 3D object detection. In: 2021 IEEE Winter Conference on Applications of Computer Vision (WACV); 2021 Jan 3–8; Waikoloa, HI, USA: IEEE; 2021. p. 1526–35. doi:10.1109/WACV48630.2021.00157.
103. Shi P, Jiang T, Yang A, Liu Z. CenRadfusion: fusing image center detection and millimeter wave radar for 3D object detection. Signal Image Video Process. 2024;18(8):5811–21. doi:10.1007/s11760-024-03273-3.
104. Kalgaonkar P, El-Sharkawy M. NeXtFusion: attention-based camera-radar fusion network for improved three-dimensional object detection and tracking. Future Internet. 2024;16(4):114. doi:10.3390/fi16040114.
105. Wang R, Lu N. MWRC3D: 3d object detection with millimeter-wave radar and camera fusion. In: 2024 7th International Conference on Advanced Algorithms and Control Engineering (ICAACE); 2024 Mar 1–3; Shanghai, China: IEEE; 2024. p. 384–9. doi:10.1109/ICAACE61206.2024.10549596.
106. Kong L, Wang Y, Chang D, Zhao Y. Temporal-enhanced radar and camera fusion for object detection. ACM Trans Multimedia Comput Commun Appl. 2025;21(1):1–16. doi:10.1145/3700442.
107. Lin Z, Liu Z, Wang Y, Zhang L, Zhu C. RCBEVDet++: toward high-accuracy radar-camera fusion 3D perception network. arXiv:2409.04979. 2024.
108. Lei K, Chen Z, Jia S, Zhang X. HVDetFusion: a simple and robust camera-radar fusion framework. arXiv:2307.11323. 2023.
109. Nobis F, Shafiei E, Karle P, Betz J, Lienkamp M. Radar voxel fusion for 3D object detection. Appl Sci. 2021;11(12):5598. doi:10.3390/app11125598.
110. Ravindran R, Santora MJ, Jamali MM. Camera, LiDAR, and radar sensor fusion based on Bayesian neural network (CLR-BNN). IEEE Sens J. 2022;22(7):6964–74. doi:10.1109/JSEN.2022.3154980.
111. Li Y, Deng J, Zhang Y, Ji J, Li H, Zhang Y. EZFusion: a close look at the integration of LiDAR, millimeter-wave radar, and camera for accurate 3D object detection and tracking. IEEE Robot Autom Lett. 2022;7(4):11182–9. doi:10.1109/LRA.2022.3193465.
112. Hendy N, Sloan C, Tian F, Duan P, Charchut N, Xie Y, et al. FISHING net: future inference of semantic heatmaps in grids. arXiv:2006.09917. 2020.
113. Liu S, Wu J, Lv B, Pan X, Wang X. Data fusion of roadside camera, LiDAR, and millimeter-wave radar. IEEE Sens J. 2024;24(20):32630–40. doi:10.1109/JSEN.2024.3448428.
114. Das A, Paul S, Scholz N, Malviya AK, Sistu G, Bhattacharya U, et al. Fisheye camera and ultrasonic sensor fusion for near-field obstacle perception in bird's-eye-view. arXiv:2402.00637. 2024.

115. Xie S, Girshick R, Dollár P, Tu Z, He K. Aggregated residual transformations for deep neural networks. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 2017 Jul 21–26; Honolulu, HI, USA: IEEE; 2017. p. 5987–95. doi:10.1109/CVPR.2017.634.
116. Aniobi A. Sensor fusion for real-time object detection and spatial positioning in unmanned vehicles using YOLOv8 and ESP32-Cam; 2024. doi:10.20944/preprints202411.0611.v1.
117. Terven J, Córdova-Esparza DM, Romero-González JA. A comprehensive review of YOLO architectures in computer vision: from YOLOv1 to YOLOv8 and YOLO-NAS. Mach Learn Knowl Extr. 2023;5(4):1680–716. doi:10.3390/make5040083.
118. Zhong M, Jiang B. Enhancing target detection and recognition in advanced driver assistance systems using infrared thermal imaging and the YOLOv5 algorithm. Int J Heat Technol. 2024;42(5):1761–8. doi:10.18280/ijht.420530.
119. Yang B, Liang M, Urtasun R. HDNET: exploiting hd maps for 3D object detection. Proceedings of the Conference on Robot Learning; 2018 Oct 29–31; Zurich, Switzerland. NY, USA: PMLR; 2018. p. 146–155.
120. Fang J, Zhou D, Song X, Zhang L. MapFusion: a general framework for 3D object detection with HDMaps. In: 2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS); 2021 Sep 27–Oct 1; Prague, Czech Republic: IEEE; 2021. p. 3406–13. doi:10.1109/IROS51168.2021.9636724.
121. Chen Y, Yang Z, Zheng X, Chang Y, Li X. PointFormer: a dual perception attention-based network for point cloud classification. In: Proceedings of the Asian Conference on Computer Vision; 2022 Dec 4–6; Macao, China. Berlin, Germany: Springer; 2022. p. 3291–307.
122. Cai D, Lam W. Graph transformer for graph-to-sequence learning. Proc AAAI Conf Artif Intell. 2020;34(5):7464–71. doi:10.1609/aaai.v34i05.6243.