### **ORIGINAL ARTICLE**



# A survey of 3D object detection algorithms for intelligent vehicles development

Zhen Li<sup>1</sup> · Yuren Du<sup>2</sup> · Miaomiao Zhu<sup>1</sup> · Shi Zhou<sup>1</sup> · Lifeng Zhang<sup>1</sup>

Received: 14 April 2021 / Accepted: 17 October 2021 / Published online: 1 November 2021 © International Society of Artificial Life and Robotics (ISAROB) 2021

#### **Abstract**

With the rapid development of Artificial Intelligent algorithms on Computer Vision, 2D object detection has greatly succeeded and been applied in various industrial products. In the past several years, the accuracy of 2D object detection has been dramatically improved, even beyond the human eyes detection ability. However, there is still a limitation of 2D object detection for the applications of Intelligent Driving. A safe and reliable self-driving car needs to detect a 3D model of the around objects so that an intelligent driving car has a perception ability to real driving situations. This paper systematically surveys the development of 3D object detection methods applied to intelligent driving technology. This paper also analyzes the shortcomings of the existing 3D detection algorithms and the future development directions of 3D detection algorithms on intelligent driving.

**Keywords** Computer vision · 3D object detection · Intelligent driving · Dataset

# 1 Introduction

Every day, almost 3700 people are killed globally in crashes involving cars, buses, motorcycles or pedestrians. More than half of those killed are pedestrians, motorcyclists, or cyclists. Crash injuries are estimated to be the eighth leading cause of death globally for all age groups and the first leading cause of death for children and young people. As reported by the U.S. Department of Transportation, more than 90% of car crashes are attributed to drivers errors. The adoption of autonomous vehicles is expected to improve driving safety and traffic efficiency. Therefore, an accurate environment perception system is necessary to reduce the traffic accidents.

The intelligent driving is one of the most important developing direction of smart cars and intelligent transportation systems. A safe and reliable self-driving car needs to detect

This work was presented in part at the 26th International Symposium on Artificial Life and Robotics (Online, January 21–23, 2021).

- Kyushu Institute of Technology, Kitakyushu, Japan
- <sup>2</sup> Yangzhou University, Yangzhou, China

a 3D model of the around objects so that an intelligent driving car should have a perception ability to real driving situations. An intelligent driving car equipped with monocular cameras, stereo cameras, or laser radars can acquire various types of information. The rapid development of these sensors contributes to intelligent driving greatly. Besides, based on 2D object detection algorithms, 3D object detection algorithms have been developing rapidly. Especially in these 2 years, some outstanding 3D detection works that are not inspired by traditional algorithm structures also achieved good performance. The improvement of 3D detection algorithms combined with the intelligent upgrading of traditional cars makes Intelligent driving closer to human lives. The comparison between 2D and 3D object detection is shown in Fig. 1.

This paper is structured as follows. Section 2 describes commonly used sensors and datasets for perception works in Intelligent Driving. Section 3 lists representative frameworks for 3D object detection in autonomous driving. The results of different 3D object detection algorithms are compared and summarized in Sect. 4. Section 5 provides a brief conclusion of 3D object detection algorithms.



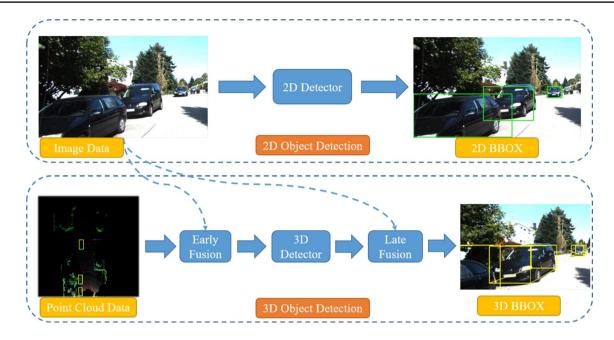


Fig. 1 The comparison between 2D detection and 3D detection

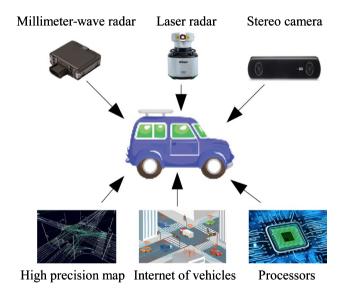
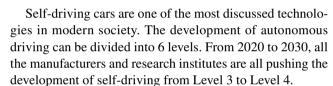


Fig. 2 The technologies demanded by intelligent vehicles

# 2 The hardware preparations for intelligent driving

# 2.1 The technologies demanded by intelligent driving

Self-driving technology involves a lot of technical areas. A mature self-driving car needs different kinds of sensors, high-precision maps, Internet of Vehicles (IoV) and high-performance chips. As is shown in Fig. 2.



Level 0: The driver does all the driving: steering, brakes, and power, etc.

Level 1: These cars can handle one task at a time, like automatic braking.

Level 2: These cars would have at least two automated functions.

Level 3: Drivers are still indispensable to intervene if necessary, but are able to shift all the functions to the vehicle.

Level 4: These cars are officially driver-free in certain environments.

Level 5: A fully-autonomous system is expected to function as good as a human driver, coping with various unconstrained driving scenarios.

# 2.2 The current manufacturers of intelligent vehicles

Autonomous driving has the most important place for artificial intelligence technology being applied to real human life. Especially in the Novel Coronavirus period, non-contact robot applications become particularly important. Many car manufacturers have made good achievements in autonomous driving technology.

In 2020, Waymo made fully driverless taxis commercially available in Arizona, US. For now, Waymo is the



only platform to transport passengers by fully autonomous vehicles, and it is seen as an industry leader.

Tesla's self-driving car has ditched expensive LiDAR to develop Tesla Vision using a combination of cameras. This makes self-driving cars cheaper and more consumers able to experience the latest self-driving technology.

Honda Research Institute plans to launch an L3 autonomous vehicle in 2022 and an L4 autonomous vehicle in 2025.

BMW plans to produce L4 autonomous vehicles by 2024, capable of driving themselves in certain geographical locations, such as highways or dual carriageways.

# 2.3 3D object detection sensors

In autonomous driving, detection sensors equipped in smart cars are the most important devices that can provide accurate and real-time information for Intelligent Driving systems. Many categories of detection sensors can be used in 3D object detection algorithms, and the most commonly used detection sensors have been listed in Fig. 3. Sensors can be generally divided into three main categories by different kinds of information detected.

First, visual sensors, such as monocular cameras and stereo cameras, can detect color information and pixel-level images. The depth information can also be calculated by the triangulation method. Second, LiDARs and radars can directly measure the depth information and acquire point

clouds maps of the 3D scenes. But the point clouds in maps are irregularly arranged and very sparse, which could bring some difficulties in later netural network calculations. The third category of sensors is the fusion of the above two categories. The fused sensors can provide more comprehensive data, including pixel-level images and point cloud maps. Theoretically, the more information provided by sensors, the higher detection accuracy could achieve by detectors.

# 2.4 3D object detection datasets

Since 2012, an increasing number of datasets has been created to train and evaluate the object detection frameworks. In 3D object detection research field, datasets can be divided into indoors and outdoors according to the application scenes. This paper surveyed the 3D object detection frameworks mainly used under Intelligent Driving scenes, so this section highlights several widely used outdoor datasets, especially autonomous driving datasets, as is shown in Table 1.

These well-known 3D object detection datasets listed here tend to use RGB cameras and LiDARs to detect different kinds of data. These fusion datasets can not only provide colorful and pixel-level images, but also provide point cloud maps with depth information. And all the datasets contain many different self-driving scenes, multiple object classifications and a mass of 3D annotation bounding boxes, which could ensure adequate training and evaluation data for 3D

	Laser radar	Millimeter-wave radar	Camera
Detecting distance	200-300m	250-1000m	<50m
Detecting precision	In the range of 2cm	In the range of 2cm	In relation to the number of cameras
Detecting angle	In vertical 360°; In horizontal 40°	In the range from $10^{\circ}$ to $70^{\circ}$	In the range of $30^{\circ}$
Detecting resolution ratio	In vertical 0.1°; In horizontal $2^{\circ}$	3°-5°	<0.1°
Wave length	905nm	1mm-10mm	-

Fig. 3 The comparison of different sensors

**Table 1** A summary of 3D object detection datasets for intelligent driving

Datasets	Scenes	Classes	Annotated frames	3D boxes	Year	Sensors
KITTI	22	8	15k	200K	2012	RGB + LiDAR
nuScenes	1k	23	40k	1.4M	2020	RGB + LiDAR
H3D	160	8	27k	1.1M	2019	RGB + LiDAR
Waymo	1k	4	200k	12M	2020	RGB + LiDAR
Lyft Level 5	366	9	46k	1.3M	2019	RGB + LiDAR
A*3D	_	7	39k	230K	2019	RGB + LiDAR
ApolloScape	-	35	140K	70K	2019	RGB + LiDAR



object detectors. A few of the most commonly used datasets are introduced as below.

### 2.4.1 KITTI

The KITTI dataset is the largest computer vision algorithm evaluation dataset in the world for autonomous driving scenes. The dataset is used to evaluate the performance of computer vision technologies such as optical flow, visual odometry, 3D object detection, and 3D tracking in autonomous driving environments. KITTI contains real-world image data collected from scenes such as urban, rural and highways. One image in KITTI could obtain up to 15 cars and 30 pedestrians. The original datasets are classified as Road, City, Residential, Campus and Person. For 3D object detection, labels are subdivided into car, van, truck, pedestrian, cyclist, tram and so on.

### 2.4.2 ApolloScape

ApolloScape dataset was provided by Baidu Inc. The 3D LiDAR object detection and tracking dataset consist of point clouds and high-quality labels. It is collected under various lighting conditions and traffic situations in Beijing. More specifically, ApolloScape dataset contains very complex traffic flows mixed with vehicles, cyclists and pedestrians.

### 2.4.3 H3D

H3D is a point cloud dataset for self-driving scenes provided by Honda. H3D dataset contains HDD dataset, which is also a large-scale natural driving dataset collected in the San Francisco Bay Area. H3D has a complete 360 degree lidar dataset with 3D bounding box labels. The dataset also contains vedio data, manually annotated from 2 to 10 Hz linearly.

# 2.4.4 Waymo

Waymo is a self-driving car company owned by Google Inc. Waymo company announced Waymo Open Dataset, a bonus-based benchmark compared to previous academic benchmarks. In addition, Waymo contains 3000 driving records, 600,000 frames, with approximately 25 million 3D bounding boxes and 22 million 2D bounding boxes. Waymo has a huge amount of data in various autonomous driving scenes.

### 2.4.5 Lyft Level 5

Lyft Level 5 is also well known like KITTI dataset. Lyft Level 5 dataset is acquired by 64-wire radars and multiple cameras. The Level 5 dataset includes over 55,000

human-labeled 3D annotated frames, surface maps, and an underlying HD spatial semantic maps.

# 3 The classification of 3D detection frameworks

We divide 3D object detection methods into three categories: vision based, point clouds based and multi-sensor fusion based methods. An overview of methodologies, advantages and limitations for these methods are demonstrated below. The following subsections address each category individually, which are shown in Fig. 4.

### 3.1 Vision based 3D detection framework

Vision sensors, especially monocular cameras, are the most important devices for 2D object detection. Meanwhile, stereo cameras and depth cameras are more suitable to detect a 3D object in real world for their ability of measuring distance. Depth information is the fourth image channel that contains information about the distance from the surface of the object to the viewpoint. Depth Map is similar to a gray image, but each of its pixel values is the actual distance the sensor is from the object. Usually, the RGB image and the

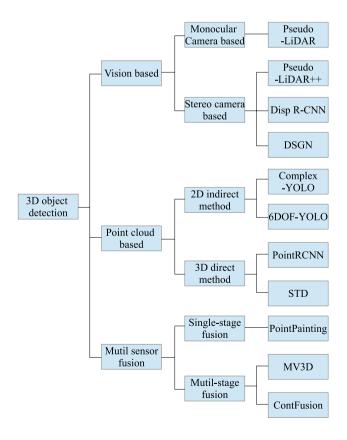


Fig. 4 The classification of the 3D object detection frameworks



Depth image have a correspondence relationship from pixels to pixels.

This subsection focuses on the frameworks that could estimate 3D bounding boxes based on RGB images and depth information. And the frameworks are mainly divided into 2 parts based on the different cameras.

### 3.1.1 Monocular camera frameworks

3D object detection from a single image without lidar is a challenging task due to the lack of accurate depth information. Traditional 2D convolution frameworks are unsuitable for this task because they cannot capture tridimensional objects and their scale information, which are important for 3D object detection. To better represent 3D structure, Ding [1] transformed depth maps estimated from 2D images into a pseudo-Lidar (PL) representation. And then use the existing 3D object detectors to detect objects. This method can predict point clouds information from a normal image pictured by monocular cameras. But there is definitely an error existing between predicted point clouds and real point clouds.

#### 3.1.2 Stereo camera frameworks

The 3D object detection using LiDAR is always much more accurate than those using cameras. While PL has led to a drastic reduction in the accuracy gap between methods based on LiDAR sensors and those based on cheap cameras. PL combined advanced deep neural networks for 3D depth estimation with those for 3D object detection together, and performed well.

However, PL's two networks need to be trained one by one. To obtain an end-to-end PL framework, Qian [2] introduced a new framework based on differentiable Change of Representation (CoR) modules that can allow PL to be trained end-to-end. The final framework achieves better results than PL in combination with Point R-CNN. This work gets the highest entry on the KITTI image-based 3D object detection leaderboard at the submission time in 2020. Although, stereo cameras could measure the depth information by disparity map. While many recent works still try to recovery point clouds with disparity estimation and then apply a 3D detector based on LiDAR sensors.

In 2020, Sun [3] proposed a new framework called Disp R-CNN for 3D object detection using stereo images. To address the challenge from scarcity of disparity annotation in training, Sun J. proposed to use a statistical truth without the need for LiDAR sensors, which reduced much of the cost in Intelligent Driving. The disparity map is computed for the entire image, which is computationally costly. So Sun J. designed an instance disparity estimation network (iDispNet) that predicts disparity only for pixels on objects of interest and learns a category-specific shape prior for

more accurate disparity estimation. Disp R-CNN achieves competitive performance and outperforms the average precision of the lasted advanced methods by 20%.

In the past several years, as is known to all, LiDAR-based 3D detection frameworks perform much better than visualbased methods. But 3D detection frameworks using LiDAR sensors could not extract semantic information without the help of images. So, Chen [4] proposed a method called Deep Stereo Geometry Network (DSGN). Significantly reduces this gap by detecting 3D objects on a differentiable volumetric representation: 3D geometric volume, which effectively encodes 3D geometric structure for 3D regular space. With this method, depth information and semantic information could be acquired simultaneously. Chen Y. provides a simple and effective one-stage stereo-based 3D detection pipeline that jointly estimates the depth and detects 3D objects in an end-to-end learning manner. DSGN approach outperforms previous stereo-based 3D detectors and even achieves comparable performance with several LiDAR-based methods on the KITTI 3D object detection leaderboard.

### 3.2 LiDAR based 3D detection framework

Currently, the point clouds-based methods achieve state-of-the-art performance in 3D detection methods. Due to the accurate depth measurement of point clouds dataset, many frameworks, which used 3D-based methods to regress 3D bounding boxes, with high detecting accuracy. But 3D based methods also will bring a lot of computation burdens. While the shortcoming of the great amount of computations becomes a challenge to real-time self-driving applications. So many researchers tend to use the 2D-based detection algorithms to realize the object detection in 3 dimensions. This subsection divided the LiDAR-based frameworks into categories: 2D indirect method and 3D direct method.

### 3.2.1 2D indirect method frameworks

Martin et al. [5] used the YOLOv2 detection network to realized a 3D object detection framework in 2018, called Complex-YOLO. And this work achieved an excellent detection speed to meet the real-time 3D detection in self-driving. This work transformed the 3D point clouds into BEV (Bird's Eye View) images, including the height, density and intensity messages. The main idea of this work is to transform the 3D point clouds data into 2D image shapes, and then use the 2D detectors to complete the 3D object detection.

Li et al. [6] proposed a 2.5D object detection framework called 6DOF-YOLO in 2021. This work also transformed the 3D point clouds into BEV (Bird's Eye View) images. But the difference from the Complex-YOLO is that Li's work did a 2.5D detection, which could regress the height of vehicles. While Complex-YOLO actually did a 2D detection and then



used point clouds message to calculate 3D bounding boxes. Li's work also used the state-of-the-art YOLOv4 [7], getting a great balance between the detecting efficiency and the accuracy.

#### 3.2.2 3D direct method frameworks

In 2018, Shi et al. [8] proposed a 3D detector called PointRCNN, which is a two-stage 3D point cloud target detection framework. By dividing the point cloud of the whole scene into foreground and background, the stage-one network directly generates a small number of high-quality 3D proposals from the point cloud. The stage-two network refines the proposals in the canonical coordinate by combining semantic features and local spatial features. PointRCNN actually used PointNet [9] to complete a 3D convolution, which has a great improvement in accuracy, but the efficiency is not mentioned.

In 2019, Yang et al. [10] present a new two-stage 3D object detection framework, named sparse-to-dense 3D Object Detector (STD). This two-stage network used Point-Net++ network as the backbone for 3D detection. Novelties are a new spherical anchor and 3D intersection-over-union (IoU), which achieves a high recall and the localization accuracy. This method outperforms other state-of-the-arts, especially on the hard dataset, but with an inference speed of only around 10 FPS.

# 3.3 Sensors fusion framework

To integrate all the advantages of different sensors, fusion technology plays an important role to balance the different characters from different sensors. Up to now, multi-sensor fusion 3D detection methods are classified into two kinds: Single-stage Fusion and Multi-stage Fusion.

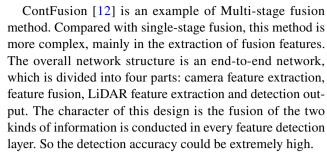
## 3.3.1 Single-stage fusion

Single-stage fusion means that different features from different sensors will be fused only one time at the feature extraction layer or at the data preprocessing stage.

One of the representing works is PointPainting [11]. Instead of end-to-end training in the design of fusion network structure, it is divided into two stages: the first stage is semantic segmentation of camera data; the second stage is 3D target detection by combining LiDAR point cloud with semantic information. The two stages are connected by feature projection.

### 3.3.2 Multi-stage fusion

Multi-stage fusion means that different features from different sensors will be fused many times at different feature extraction layers during the whole fusion algorithm process.



Chen [13] proposed a two-stage network of 3D detection called MV3D in 2017. MV3D uses point clouds and images as input. The whole network can be divided into two stages: the first is a 3D Proposal Network to generate the Region of Interest (ROI). The second is a Region-based Fusion Network to fuse the LiDAR Bird View, LiDAR Front View and images data. MV3D comprehensively analysed the impact of different fusion stages on the detection results. And MV3D finally used the multi-stages fusion in the different layers of the Fusion Network.

# 4 Summary

As we can see from Table 2, the results of the frameworks mentioned are evaluated using a new evaluation metric on the KITTI leaderboard. Several methods undergoing old evaluation are not available on the leaderboard. First, generally comparing, all the Bird's Eye View (BEV) detections have a better performance than the 3D Detection in Average Precision (AP). Secondly, as we can see from this table, frameworks using different sensors also have a huge gap in AP. Visual sensors-based frameworks always get the lowest accuracy, while the multi-sensor fusion frameworks achieved state-of-the-art performances. But when we take the sensors' prices into account, multi-sensor fusion frameworks would cost the most in these three kinds of works, and the camera sensors are the least expensive. The most excellent visual 3D detection frameworks have achieved results that are greatly close to the best performance. So Intelligent Driving systems using visual sensors will have a great prospect for future development.

In February 2021, the California DMV (Department of Motor Vehicles) released new autonomous driving data for the year 2020. The California DMV annual Miles Per Intervention (MPI) is one of the key measurements of autonomous driving, reflecting the average miles of each intervention annually. It is widely regarded by the industry as a more objective, quantitative and accurate measure than driving experience test. As the analyze of the different intelligent vehicles manufacturers, and along with the excellent research and development institutions, the best testing results until now are listed in Table 3. As the Table shows, the top 10 testing results are listed here, and the intelligent



**Table 2** 3D object detection benchmark on KITTI test set

Methods	Sensors	BEV det	BEV detection AP (%)			3D detection AP (%)		
		easy	mod	hard	easy	mod	hard	
Pseudo-LiDAR	Camera	60.70	39.20	37.00	40.00	27.40	25.30	
Pseudo-LiDAR++		82.00	64.00	57.30	67.90	50.10	45.30	
Disp R-CNN		77.63	64.38	50.68	64.29	47.73	40.11	
DSGN		72.31	54.27	47.71	83.24	63.91	57.83	
Complex-YOLO	LiDAR	44.11	39.03	37.50	34.48	29.39	27.26	
6DOF-YOLO		_	_	_	92.35	33.10	72.14	
STD		77.23	68.15	64.09	72.86	61.46	57.93	
PointRCNN		_	_	_	88.88	78.63	77.38	
PointPainting	Fusion	_	75.80	_	_	_	_	
MV3D		86.55	78.10	76.67	71.29	62.68	56.56	
ContFusion		88.81	85.83	77.33	82.54	66.22	64.04	

Table 3 The testing results of California self-driving car disengagements

Manufacturer	Km	Disengage	ement per Km
Waymo	1,006,142	21	47911.50
Cruise	1,232,079	27	45632.55
AutoX	65,174	2	32,587.20
Pony.AI	360,794	21	17180.65
Argo.AI	33,659	2	16829.74
WeRide	20,822	2	10411.20
DiDi	16,642	2	8321.19
Nuro	88,592	11	8053.79
Deeproute.AI	16,029	3	5342.93
Zoox	164,034	63	2603.71

driving technology of Waymo company possesses the best performance.

This paper also summarized some shortcomings of the 3D object detection systems applied in autonomous driving. First, in the practical application of driverless vehicles, obstructing and surrounding environment confusion will lead detectors to miss the target object in the traffic. Second, there will be a great amount of data to process in the traffic, while most algorithms cannot meet the requirements of real-time in autonomous driving. Third, in unmanned driving, the real-time performance of 3D target detection is difficult to be guaranteed and the accuracy of 3D target detection is not high enough for safe self-driving.

The final goal of Intelligent Driving is that the self-driving skills need to exceed the average driving skill of human. So, Intelligent Driving needs to use more accurate 3D object detection algorithms and more faster processing chips to make the self-driving cars more security and more fluency. Frameworks using vision cameras are the most competitive methods because of the low cost and potential for higher detecting accuracy. It is believed that self-driving

cars using cameras only have the capacity to ensure security because human beings use their eyes only. The most important thing for self-driving is to possess a smart brain—a smart 3D detector.

### 5 Conclusion

This paper systematically surveys the development of 3D object detection methods applied to intelligent driving technology. This paper also discusses the different types of information acquired by different sensors, such as RGB images, point cloud maps, and multi-source information fusion. After analyzing the effective and representative 3D detection works in details, this paper summarized these works into different types according to the different realizing algorithms. The experiment data of different kinds of 3D detection frameworks are discussed and compared on the KITTI Benchmark. And the practical application of these detectors in real traffic environments is also analyzed. The relationship of the detecting accuracy and the demanding accuracy for safe driving is also researched. This paper finally talks about the shortcomings of the existing 3D object detection frameworks and the future development directions on Intelligent Driving.

# References

- Ding M, Huo Y, Yi H et al (2020) Learning Depth-Guided Convolutions for Monocular 3D Object Detection. 2020 IEEE/CVF conference on computer vision and pattern recognition (CVPR), pp. 11669–11678
- Qian R, Garg D, Wang Y et al (2020) End-to-end pseudo-LiDAR for image-based 3D object detection, 2020 IEEE/CVF conference on computer vision and pattern recognition (CVPR), pp. 5880–5889
- 3. Sun J, Chen L, Xie Y et al (2020) Disp R-CNN: stereo 3D object detection via shape prior guided instance disparity estimation,



- 2020 IEEE/CVF conference on computer vision and pattern recognition (CVPR), pp. 10545–10554
- Chen Y, Liu S, Shen X et al (2020) DSGN: deep stereo geometry network for 3D object detection, 2020 IEEE/CVF conference on computer vision and pattern recognition (CVPR), pp. 12533–12542
- Martin S, Stefan M, Karl A et al (2018) Complex-yolo: real-time 3d objectdetection on point clouds. Computer vision and pattern recognition
- Zhen L, Yuren D, Qingqing H et al (2021) A 3D object detection framework for self-driving using YOLOv4, Proceedings of the 8th IIAE international conference on intelligent systems and image processings, pp. 9–15
- Alexey B, Chien-Yao W, Hong-Yuan L et al (2020) YOLOv4: optimal speed and accuracy of object detection. Computer vision and pattern recognition
- Shi S, Wang X, Li H et al (2019) PointRCNN: 3D object proposal generation and detection from point cloud, IEEE/CVF conference on computer vision and pattern recognition (CVPR), pp. 770–779
- Charles R. Q, Hao S, Kaichun M et al (2017) PointNet: deep learning on point sets for 3D classification and segmentation. Computer vision and pattern recognition

- Zetong Y, Yanan S, Liu S et al (2019) Std: sparse-to-dense 3d object detector for point cloud, IEEE/CVF conference on computer vision and pattern recognition, pp. 1951–1960
- Vora S, Lang A H, Helou B et al (2020) PointPainting: sequential fusion for 3D object detection, 2020 IEEE/CVF conference on computer vision and pattern recognition (CVPR), pp. 4603–4611
- 12. Liang M, Yang B, Wang S et al (2018) Deep continuous fusion for multi-sensor 3d object detection, Proceedings of the European conference on computer vision (ECCV), pp. 641–656
- Chen X, Ma H, Wan J et al (2017) Multi-view 3D object detection network for autonomous driving, 2017 IEEE conference on computer vision and pattern recognition (CVPR), pp. 6526–6534

**Publisher's Note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

