International Journal of Computer Vision (2023) 131:1909–1963
https://doi.org/10.1007/s11263-023-01790-1

3D Object Detection for Autonomous Driving: A Comprehensive Survey

Jiageng Mao[1] · Shaoshuai Shi[2] · Xiaogang Wang[1,3] · Hongsheng Li[1,3,4]

Received: 17 June 2022 / Accepted: 11 March 2023 / Published online: 27 April 2023

Abstract
Autonomous driving, in recent years, has been receiving increasing attention for its potential to relieve drivers' burdens and improve the safety of driving. In modern autonomous driving pipelines, the perception system is an indispensable component, aiming to accurately estimate the status of surrounding environments and provide reliable observations for prediction and planning. 3D object detection, which aims to predict the locations, sizes, and categories of the 3D objects near an autonomous vehicle, is an important part of a perception system. This paper reviews the advances in 3D object detection for autonomous driving. First, we introduce the background of 3D object detection and discuss the challenges in this task. Second, we conduct a comprehensive survey of the progress in 3D object detection from the aspects of models and sensory inputs, including LiDAR-based, camera-based, and multi-modal detection approaches. We also provide an in-depth analysis of the potentials and challenges in each category of methods. Additionally, we systematically investigate the applications of 3D object detection in driving systems. Finally, we conduct a performance analysis of the 3D object detection approaches, and we further summarize the research trends over the years and prospect the future directions of this area.

Keywords 3D object detection · Perception · Autonomous driving · Deep learning · Computer vision · Robotics

1 Introduction

Autonomous driving, which aims to enable vehicles to perceive the surrounding environments intelligently and move safely with little or no human effort, has attained rapid progress in recent years. Autonomous driving techniques have been broadly applied in many scenarios, including self-driving trucks, robotaxis, delivery robots, etc., and are capable of reducing human error and enhancing road safety. As a core component of autonomous driving systems, automotive perception helps autonomous vehicles understand the surrounding environments with sensory input. Perception systems generally take multi-modality data (images from cameras, point clouds from LiDAR scanners, high-definition maps etc.) as input, and predict the geometric and semantic information of critical elements on a road. High-quality perception results serve as reliable observations for the following steps such as object tracking, trajectory prediction, and path planning.

To obtain a comprehensive understanding of driving environments, many vision tasks can be involved in a perception system, e.g. object detection and tracking, lane detection, and semantic and instance segmentation. Among these perception tasks, 3D object detection is one of the most indispensable tasks in an automotive perception system. 3D object detection aims to predict the locations, sizes, and classes of critical objects, e.g. cars, pedestrians, cyclists, in the 3D space. In contrast to 2D object detection which only generates 2D bounding boxes on images and ignores the actual distance information of objects from the ego-vehicle, 3D object detection focuses on the localization and recognition of objects in the real-world 3D coordinate system. The

Communicated by Vittorio Ferrari.

✉ Hongsheng Li
hsli@ee.cuhk.edu.hk

Jiageng Mao
maojiageng@gmail.com

Shaoshuai Shi
shaoshuaics@gmail.com

Xiaogang Wang
xgwang@ee.cuhk.edu.hk

[1] The Chinese University of Hong Kong, Hong Kong, China

[2] Max Planck Institute for Informatics, Saarbrücken, Germany

[3] Centre for Perceptual and Interactive Intelligence, Hong Kong, China

[4] Shanghai AI Laboratory, Shanghai, China

geometric information predicted by 3D object detection in real-world coordinates can be directly utilized to measure the distances between the ego-vehicle and critical objects, and to further help plan driving routes and avoid collisions.

3D object detection methods have evolved rapidly with the advances of deep learning techniques in computer vision and robotics. These methods have been trying to address the 3D object detection problem from a particular aspect, e.g. detection from a particular sensory type or data representation, and lack a systematic comparison with the methods of other categories. Hence a comprehensive analysis of the strengths and weaknesses of all types of 3D object detection methods is desirable and can provide some valuable findings to the research community.

To this end, we propose to comprehensively review the 3D object detection methods for autonomous driving applications and provide in-depth analysis and a systematic comparison on different categories of approaches. Compared to the existing surveys (Arnold et al., 2019; Liang et al., 2021b; Qian et al., 2021b), our paper broadly covers the recent advances in this area, e.g. 3D object detection from range images, self-/semi-/weakly-supervised 3D object detection, 3D detection in end-to-end driving systems. In contrast to the previous surveys that only focus on detection from point cloud (Guo et al., 2020; Fernandes et al., 2021; Zamanakos et al., 2021), from monocular images (Wu et al., 2020a; Ma et al., 2022), and from multi-modal inputs (Wang et al., 2021h), our paper systematically investigate the 3D object detection methods from all sensory types and in most application scenarios. The major contributions of this work can be summarized as follows:

- We provide a comprehensive review of the 3D object detection methods from different perspectives, including detection from different sensory inputs (LiDAR-based, camera-based, and multi-modal detection), detection from temporal sequences, label-efficient detection, as well as the applications of 3D object detection in driving systems.
- We summarize 3D object detection approaches structurally and hierarchically, conduct a systematic analysis of these methods, and provide valuable insights for the potentials and challenges of different categories of methods.
- We conduct a comprehensive performance and speed analysis on the 3D object detection approaches, identify the research trends over years, and provide insightful views on the future directions of 3D object detection.

The structure of this paper is organized as follows. First, we introduce the problem definition, datasets, and evaluation metrics of 3D object detection in Sect. 2. Then, we review and analyze the 3D object detection methods based on LiDAR sensors (Sect. 3), cameras (Sect. 4), multi-sensor fusion (Sect. 5), and Transformer-based architectures (Sect. 6). Next, we introduce the detection methods that leverage temporal data in Sect. 7 and utilize fewer labels in Sect. 8. We subsequently discuss some critical problems of 3D object detection in driving systems in Sect. 9. Finally, we conduct a speed and performance analysis, investigate the research trends, and prospect the future directions of 3D object detection in Sect. 10. A hierarchically-structured taxonomy is shown in Fig. 1. We also provide a constantly updated project page here.

2 Background

2.1 What is 3D Object Detection?

Problem definition 3D object detection aims to predict bounding boxes of 3D objects in driving scenarios from sensory inputs. A general formula of 3D object detection can be represented as

$$\mathcal{B} = f_{det}(\mathcal{I}_{sensor}), \tag{1}$$

where $\mathcal{B} = \{B_1, \cdots, B_N\}$ is a set of N 3D objects in a scene, f_{det} is a 3D object detection model, and $\mathcal{I}_{sensor}$ is one or more sensory inputs. How to represent a 3D object B_i is a crucial problem in this task, since it determines what 3D information should be provided for the following prediction and planning steps. In most cases, a 3D object is represented as a 3D cuboid that includes this object, that is

$$B = [x_c, y_c, z_c, l, w, h, \theta, class], \tag{2}$$

where (x_c, y_c, z_c) is the 3D center coordinate of a cuboid, l, w, h is the length, width, and height of a cuboid respectively, θ is the heading angle, i.e. the yaw angle, of a cuboid on the ground plane, and $class$ denotes the category of a 3D object, e.g. cars, trucks, pedestrians, cyclists. In Caesar et al. (2020), additional parameters v_x and v_y that describe the speed of a 3D object along x and y axes on the ground are employed.

Sensory inputs There are many types of sensors that can provide raw data for 3D object detection. Among the sensors, radars, cameras, and LiDAR (Light Detection And Ranging) sensors are the three most widely adopted sensory types. Radars have long detection range and are robust to different weather conditions. Due to the Doppler effect, radars could provide additional velocity measurements. Cameras are cheap and easily accessible, and can be crucial for understanding semantics, e.g. the type of traffic sign. Cameras produce images $\mathcal{I}_{cam} \in R^{W \times H \times 3}$ for 3D object detection, where W, H are the width and height of an image, and each

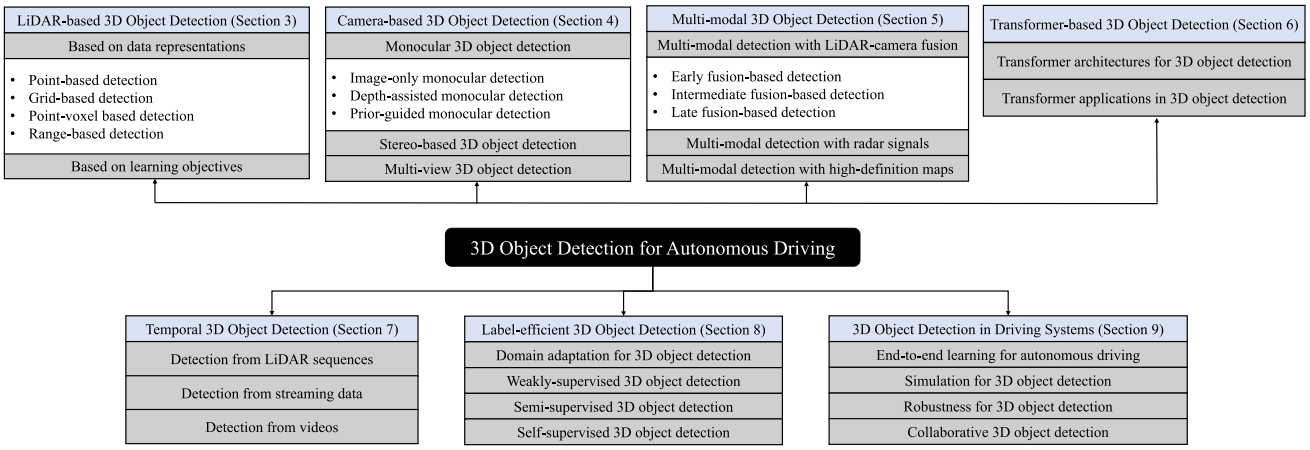

Fig. 1 Hierarchically-structured taxonomy of 3D object detection for autonomous driving

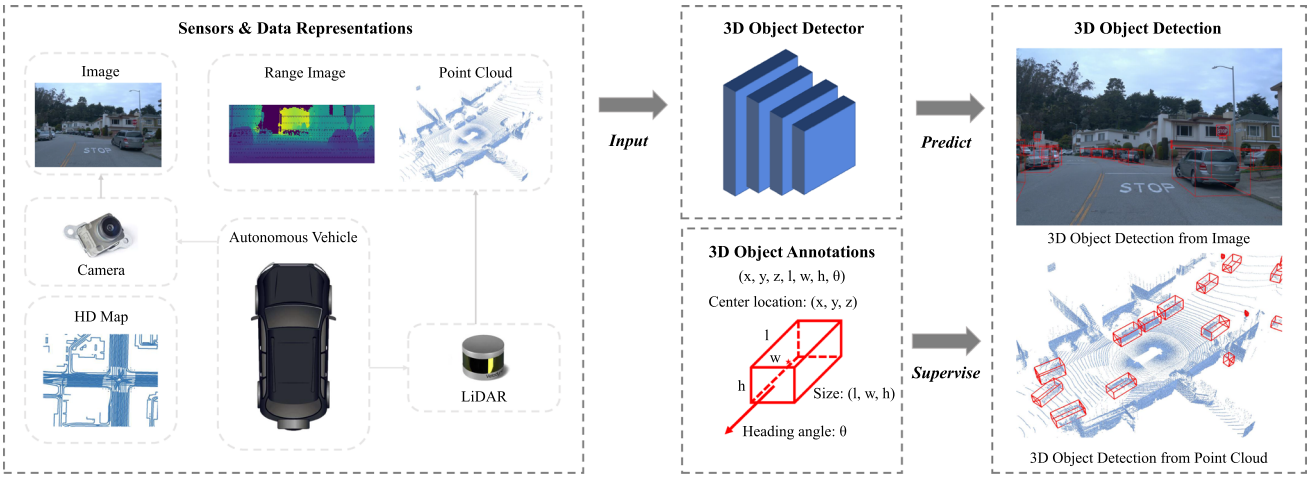

Fig. 2 An illustration of 3D object detection in autonomous driving scenarios

pixel has 3 RGB channels. Albeit cheap, cameras have intrinsic limitations to be utilized for 3D object detection. First, cameras only capture appearance information, and are not capable of directly obtaining 3D structural information about a scene. On the other hand, 3D object detection normally requires accurate localization in the 3D space, while the 3D information, e.g. depth, estimated from images normally has large errors. In addition, detection from images is generally vulnerable to extreme weather and time conditions. Detecting objects from images at night or on foggy days is much harder than detection on sunny days, which leads to the challenge of attaining sufficient robustness for autonomous driving.

As an alternative solution, LiDAR sensors can obtain fine-grained 3D structures of a scene by emitting laser beams and then measuring their reflective information. A LiDAR sensor that emits m beams and conducts measurements for n times in one scan cycle can produce a range image $\mathcal{I}_{range} \in R^{m \times n \times 3}$, where each pixel of a range image contains range r, azimuth α, and inclination ϕ in the spherical coordinate system as well as the reflective intensity. Range images are the raw data format obtained by LiDAR sensors, and can be further converted into point clouds by transforming spherical coordinates into Cartesian coordinates. A point cloud can be represented as $\mathcal{I}_{point} \in R^{N \times 3}$, where N denotes the number of points in a scene, and each point has 3 channels of xyz coordinates. Both range images and point clouds contain accurate 3D information directly acquired by LiDAR sensors. Hence in contrast to cameras, LiDAR sensors are more suitable for detecting objects in the 3D space, and LiDAR sensors are also less vulnerable to time and weather changes. However, LiDAR sensors are much more expensive than cameras, which may limit the applications in driving scenarios. An illustration of 3D object detection is shown in Fig. 2.

Analysis: comparisons with 2D object detection 2D object detection, which aims to generate 2D bounding boxes on images, is a fundamental problem in computer vision. 3D object detection methods have borrowed many design paradigms from the 2D counterparts: proposals generation and refinement, anchors, non maximum suppression, etc. However, from many aspects, 3D object detection is not a

naive adaptation of 2D object detection methods to the 3D space. (1) 3D object detection methods have to deal with heterogeneous data representations. Detection from point clouds requires novel operators and networks to handle irregular point data, and detection from both point clouds and images needs special fusion mechanisms. (2) 3D object detection methods normally leverage distinct projected views to generate object predictions. As opposed to 2D object detection methods that detect objects from the perspective view, 3D methods have to consider different views to detect 3D objects, e.g. from the bird's-eye view, point view, and cylindrical view. (3) 3D object detection has a high demand for accurate localization of objects in the 3D space. A decimeter-level localization error can lead to a detection failure of small objects such as pedestrians and cyclists, while in 2D object detection, a localization error of several pixels may still maintain a high Intersection over Union (IoU) between predicted and ground truth bounding boxes. Hence accurate 3D geometric information is indispensable for 3D object detection from either point clouds or images.

Analysis: comparisons with indoor 3D object detection There is also a branch of works (Qi et al., 2018, 2019, 2020; Liu et al., 2021d) on 3D object detection in indoor scenarios. Indoor datasets, e.g. ScanNet (Dai et al., 2017), SUN RGB-D (Song et al., 2015), provide 3D structures of rooms reconstructed from RGB-D sensors and 3D annotations including doors, windows, beds, chairs, etc. 3D object detection in indoor scenes is also based on point clouds or images. However, compared to indoor 3D object detection, there are unique challenges of detection in driving scenarios. (1) Point cloud distributions from LiDAR and RGB-D sensors are different. In indoor scenes, points are relatively uniformly distributed on the scanned surfaces and most 3D objects receive a sufficient number of points on their surfaces. However, in driving scenes most points fall in a near neighborhood of the LiDAR sensor, and those 3D objects that are far away from the sensor will receive only a few points. Thus methods in driving scenarios are specially required to handle various point cloud densities of 3D objects and accurately detect those faraway and sparse objects. (2) Detection in driving scenarios has a special demand for inference latency. Perception in driving scenes has to be real-time to avoid accidents. Hence those methods are required to be computationally efficient, otherwise they will not be applied in real-world applications.

2.2 Datasets

A large number of driving datasets have been built to provide multi-modal sensory data and 3D annotations for 3D object detection. Table 1 lists the datasets that collect data in driving scenarios and provide 3D cuboid annotations. KITTI (Geiger et al., 2012) is a pioneering work that proposes a standard data collection and annotation paradigm: equipping a vehicle with cameras and LiDAR sensors, driving the vehicle on roads for data collection, and annotating 3D objects from the collected data. The following works made improvements mainly from the 4 aspects. (1) Increasing the scale of data. Compared to Geiger et al. (2012), the recent large-scale datasets (Sun et al., 2020c; Caesar et al., 2020; Mao et al., 2021b) have more than 10x point clouds, images and annotations. (2) Improving the diversity of data. Geiger et al. (2012) only contains driving data obtained in the daytime and in good weather, while recent datasets (Choi et al., 2018; Chang et al., 2019; Pham et al., 2020; Caesar et al., 2020; Sun et al., 2020c; Xiao et al., 2021; Mao et al., 2021b; Wilson et al., 2021) provide data captured at night or in rainy days. (3) Providing more annotated categories. Some datasets (Liao et al., 2021; Xiao et al., 2021; Geyer et al., 2020; Wilson et al., 2021; Caesar et al., 2020) can provide more fine-grained object classes, including animals, barriers, traffic cones, etc. They also provide fine-grained sub-categories of existing classes, e.g. the adult and child category of the existing pedestrian class in Caesar et al. (2020). (4) Providing data of more modalities. In addition to images and point clouds, recent datasets provide more data types, including high-definition maps (Kesten et al., 2019; Chang et al., 2019; Sun et al., 2020c; Wilson et al., 2021), radar data (Caesar et al., 2020), long-range LiDAR data (Weng et al., 2020; Wang et al., 2021j), thermal images (Choi et al., 2018).

Analysis: future prospects of driving datasets The research community has witnessed an explosion of datasets for 3D object detection in autonomous driving scenarios. A subsequent question may be asked: what will the next-generation autonomous driving datasets look like? Considering the fact that 3D object detection is not an independent task but a component in driving systems, we propose that future datasets will include all important tasks in autonomous driving: perception, prediction, planning, and mapping, as a whole and in an end-to-end manner, so that the development and evaluation of 3D object detection methods will be considered from an overall and systematic view. There are some datasets (Sun et al., 2020c; Caesar et al., 2020; Yogamani et al., 2019) working towards this goal.

2.3 Evaluation Metrics

Various evaluation metrics have been proposed to measure the performance of 3D object detection methods. Those evaluation metrics can be divided into two categories. The first category tries to extend the Average Precision (AP) metric (Lin et al., 2014) in 2D object detection to the 3D space:

$$AP = \int_0^1 max\{p(r'|r' \geq r)\}dr, \tag{3}$$

Table 1 Datasets for 3D object detection in driving scenarios

Dataset	Year	Size (h)	Real-world	LiDAR scans	Images	3D annotations	Classes	Night/rain	Locations	Other data
KITTI Geiger et al. (2012, 2013)	2012	1.5	Yes	15k	15k	200k	8	No/No	Germany	–
KAIST Choi et al. (2018)	2018	–	Yes	8.9k	8.9k	Yes	3	Yes/No	Korea	thermal images
ApolloScape Huang et al. (2019); Ma et al. (2019b)	2019	100	Yes	20k	144k	475k	6	–/–	China	–
H3D Patil et al. (2019)	2019	0.77	Yes	27k	83k	1.1 M	8	No/No	USA	–
Lyft L5 Kesten et al. (2019)	2019	2.5	Yes	46k	323k	1.3 M	9	No/No	USA	maps
Argoverse Chang et al. (2019)	2019	0.6	Yes	44k	490k	993k	15	Yes/Yes	USA	Maps
WoodScape Yogamani et al. (2019)	2019	–	Yes	10k	10k	–	3	Yes/Yes	–	Fish-eye camera
AIODrive Weng et al. (2020)	2020	6.9	No	250k	250k	26 M	–	Yes/Yes	–	Long-range data
A*3D Pham et al. (2020)	2020	55	Yes	39k	39k	230k	7	Yes/Yes	SG	–
A2D2 Geyer et al. (2020)	2020	–	Yes	12.5k	41.3k	–	14	–/–	Germany	–
Cityscapes 3D Gählert et al. (2020)	2020	–	Yes	0	5k	–	8	No/No	Germany	–
nuScenes Caesar et al. (2020)	2020	5.5	Yes	400k	1.4 M	1.4 M	23	Yes/Yes	SG, USA	Maps, radar data
Waymo Open Sun et al. (2020c)	2020	6.4	Yes	230k	1 M	12 M	4	Yes/Yes	USA	Maps
Cirrus Wang et al. (2021j)	2021	–	Yes	6.2k	6.2k	–	8	–/–	USA	Long-range data
PandaSet Xiao et al. (2021)	2021	0.22	Yes	8.2k	49k	1.3 M	28	Yes/Yes	USA	–
KITTI-360 Liao et al. (2021)	2021	–	Yes	80k	300k	68k	37	–/–	Germany	–
Argoverse v2 Wilson et al. (2021)	2021	–	Yes	–	–	–	30	Yes/Yes	USA	Maps
ONCE Mao et al. (2021b)	2021	144	Yes	1 M	7 M	417K	5	Yes/Yes	China	–

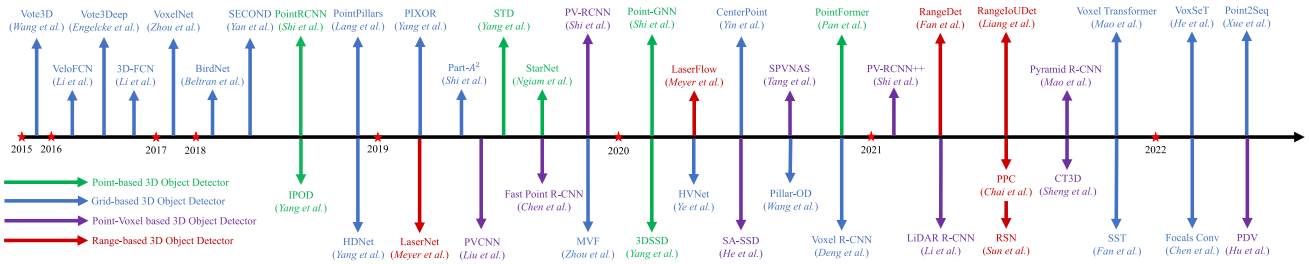

Fig. 3 Chronological overview of the LiDAR-based 3D object detection methods

where $p(r)$ is the precision-recall curve same as (Lin et al., 2014). The major difference with the 2D AP metric lies in the matching criterion between ground truths and predictions when calculating precision and recall. KITTI Geiger et al. (2012) proposes two widely-used AP metrics: AP_{3D} and AP_{BEV}, where AP_{3D} matches the predicted objects to the respective ground truths if the 3D Intersection over Union (3D IoU) of two cuboids is above a certain threshold, and AP_{BEV} is based on the IoU of two cuboids from the bird's-eye view (BEV IoU). NuScenes Caesar et al. (2020) proposes AP_{center} where a predicted object is matched to a ground truth object if the distance of their center locations is below a certain threshold, and NuScenes Detection Score (NDS) is further proposed to take both AP_{center} and the error of other parameters, i.e. size, heading, velocity, into consideration. Waymo Sun et al. (2020c) proposes $AP_{hungarian}$ that applies the Hungarian algorithm to match the ground truths and predictions, and AP weighted by Heading (APH) is proposed to incorporate heading errors as a coefficient into the AP calculation.

The other category of approaches tries to resolve the evaluation problem from a more practical perspective. The idea is that the quality of 3D object detection should be relevant to the downstream task, i.e. motion planning, so that the best detection methods should be most helpful to planners to ensure the safety of driving in practical applications. Toward this goal, PKL (Philion et al., 2020) measures the detection quality using the KL-divergence of the ego vehicle's future planned states based on the predicted and ground truth detections respectively. SDE Deng et al. (2021a) leverages the minimal distance from the object boundary to the ego vehicle as the support distance and measures the support distance error.

Analysis: pros and cons of different evaluation metrics AP-based evaluation metrics (Geiger et al., 2012; Caesar et al., 2020; Sun et al., 2020c) can naturally inherit the advantages from 2D detection. However, those metrics overlook the influence of detection on safety issues, which are also critical in real-world applications. For instance, a misdetection of an object near the ego vehicle and far away from the ego vehicle may receive a similar level of punishment in AP calculation, but a misdetection of nearby objects is substantially more dangerous than a misdetection of faraway objects in practical applications. Thus AP-based metrics may not be the optimal solution from the perspective of safe driving. PKL Philion et al. (2020) and SDE Deng et al. (2021a) partly resolve the problem by considering the effects of detection in downstream tasks, but additional challenges will be introduced when modeling those effects. PKL Philion et al. (2020) requires a pre-trained motion planner for evaluating the detection performance, but a pre-trained planner also has innate errors that could make the evaluation process inaccurate. SDE Deng et al. (2021a) requires reconstructing object boundaries which is generally complicated and challenging.

3 LiDAR-Based 3D Object Detection

In this section, we introduce the 3D object detection methods based on LiDAR data, i.e. point clouds or range images. In Sect. 3.1, we review and analyze the LiDAR-based 3D object detection models based on different data representations, including the point-based, grid-based, point-voxel based, and range-based methods. In Sect. 3.2, we investigate the learning objectives for 3D object detectors, including the anchor-based and anchor-free frameworks, as well as the auxiliary tasks adopted in LiDAR-based 3D object detection. A chronological overview of the LiDAR-based 3D detection methods is shown in Fig. 3.

3.1 Data Representations for 3D Object Detection

Problem and Challenge In contrast to images where pixels are regularly distributed on an image plane, point cloud is a sparse and irregular 3D representation that requires specially designed models for feature extraction. Range image is a dense and compact representation, but range pixels contain 3D information instead of RGB values. Hence directly applying conventional convolutional networks on range images may not be an optimal solution. On the other hand, detection in autonomous driving scenarios generally has a requirement for real-time inference. Therefore, how to develop a model that could effectively handle point cloud or range image data

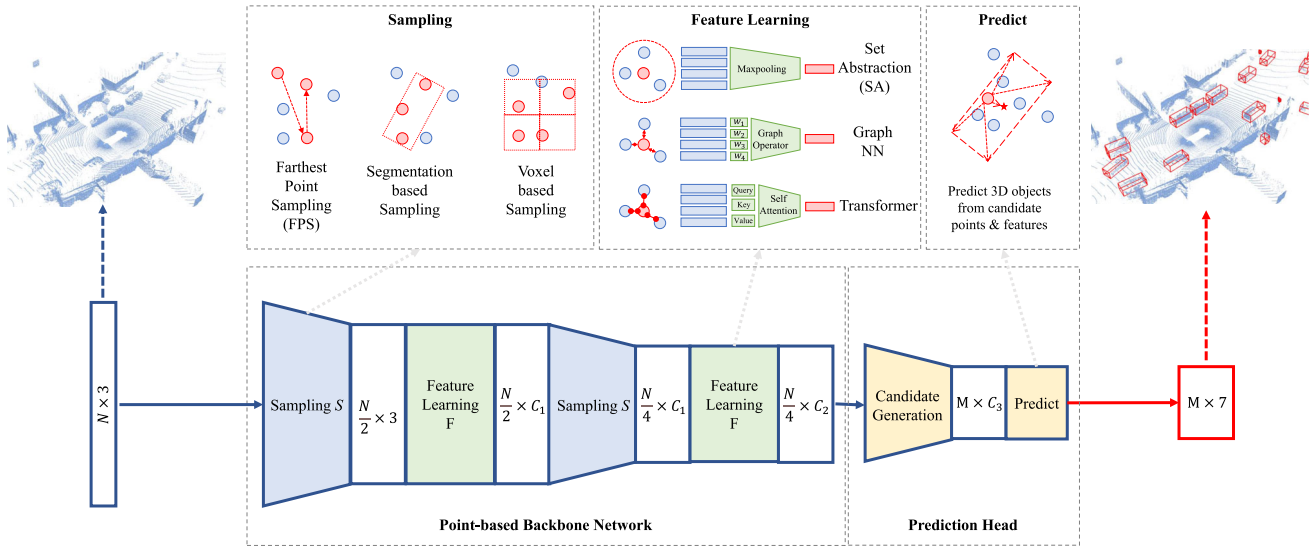

Fig. 4 An illustration of point-based 3D object detection methods

while maintaining a high efficiency remains an open challenge to the research community.

3.1.1 Point-Based 3D Object Detection

General Framework Point-based 3D object detection methods generally inherit the success of deep learning techniques on point cloud (Qi et al., 2017a, b; Wang et al., 2019c; Mao et al., 2019) and propose diverse architectures to detect 3D objects directly from raw points. Point clouds are first passed through a point-based backbone network, in which the points are gradually sampled and features are learned by point cloud operators. 3D bounding boxes are then predicted based on the downsampled points and features. A general point-based detection framework is shown in Fig. 4 and a taxonomy of point-based detectors is in Table 2. There are two basic components of a point-based 3D object detector: point cloud sampling and feature learning.

Point Cloud Sampling Farthest Point Sampling (FPS) in PointNet++ Qi et al. (2017b) has been broadly adopted in point-based detectors, in which the farthest points are sequentially selected from the original point set. PointRCNN Shi et al. (2019) is a pioneering work that adopts FPS to progressively downsample input point cloud and generate 3D proposals from the downsampled points. Similar design paradigm has also been adopted in many following works with improvements like segmentation guided filtering (Yang et al., 2018c), feature space sampling (Yang et al., 2020c), random sampling (Ngiam et al., 2019), voxel-based sampling (Shi & Rajkumar, 2020), and coordinate refinement (Pan et al., 2021).

Point Cloud Feature Learning A series of works (Shi et al., 2019; Yang et al., 2019; Zhou et al., 2020a; Wang et al., 2021f) leverage set abstraction in Qi et al. (2017a) to learn features from point cloud. Specifically, context points are first collected within a pre-defined radius by ball query. Then, the context points and features are aggregated through multi-layer perceptrons and maxpooling to obtain the new features. There are also other works resorting to different point cloud operators, including graph operators (Shi & Rajkumar, 2020; Zarzar et al., 2019; Ngiam et al., 2019; Feng et al., 2020; He et al., 2020c), attentional operators (Paigwar et al., 2019), and Transformer (Pan et al., 2021).

Analysis: potentials and challenges on point cloud feature learning and sampling The representation power of point-based detectors is mainly restricted by two factors: the number of context points and the context radius adopted in feature learning. Increasing the number of context points will gain more representation power but at the cost of increasing much memory consumption. Suitable context radius in ball query is also an important factor: the context information may be insufficient if the radius is too small and the fine-grained 3D information may lose if the radius is too large. These two factors have to be determined carefully to balance the efficacy and efficiency of detection models.

Point cloud sampling is a bottleneck in inference time for most point-based methods. Random uniform sampling can be conducted in parallel with high efficiency. However, considering points in LiDAR sweeps are not uniformly distributed, random uniform sampling may tend to over-sample those regions of high point cloud density while under-sample those sparse regions, which normally leads to poor performance compared to farthest point sampling. Farthest point sampling and its variants can attain a more uniform sampling result by sequentially selecting the farthest point from the existing point set. Nevertheless, farthest point sampling

Table 2 A taxonomy of point-based detection methods based on point cloud sampling and feature learning

Method	Context Ω	Sampling S	Feature F
PointRCNN Shi et al. (2019)	Ball Query	FPS	Set Abstraction
IPOD Yang et al. (2018c)	Ball Query	Seg.	Set Abstraction
STD Yang et al. (2019)	Ball Query	FPS	Set Abstraction
3DSSD Yang et al. (2020c)	Ball Query	Fusion-FPS	Set Abstraction
Point-GNN Shi and Rajkumar (2020)	Ball Query	Voxel	Graph
StarNet Ngiam et al. (2019)	Ball Query	Targeted-FPS	Graph
Pointformer Pan et al. (2021)	Ball Query	FPS + Refine	Transformer

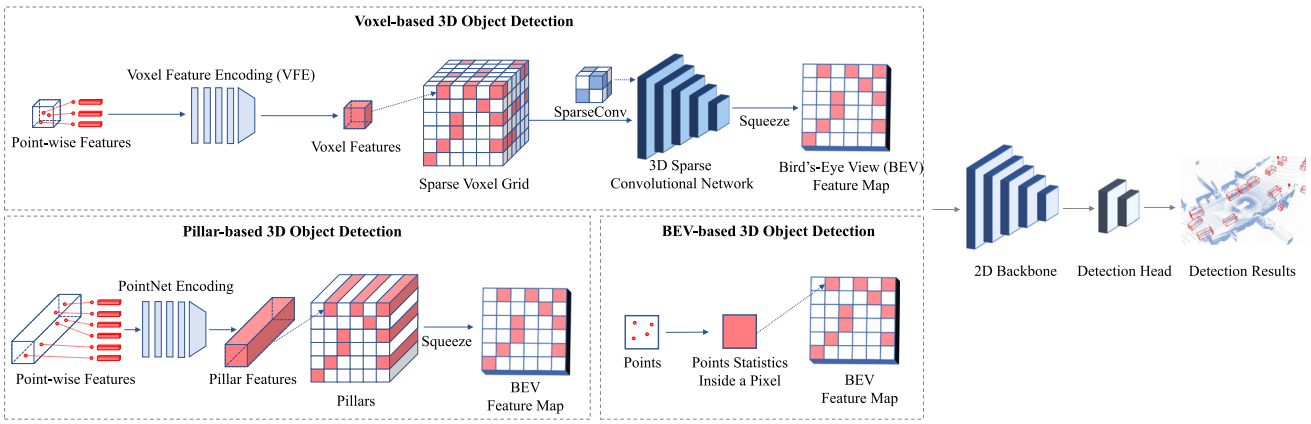

Fig. 5 An illustration of grid-based 3D object detection methods

is intrinsically a sequential algorithm and can not become highly parallel. Thus farthest point sampling is normally time-consuming and not ready for real-time detection.

3.1.2 Grid-Based 3D Object Detection

General Framework Grid-based 3D object detectors first rasterize point clouds into discrete grid representations, i.e. voxels, pillars, and bird's-eye view (BEV) feature maps. Then they apply conventional 2D convolutional neural networks or 3D sparse neural networks to extract features from the grids. Finally, 3D objects can be detected from the BEV grid cells. An illustration of grid-based 3D object detection is shown in Fig. 5 and a taxonomy of grid-based detectors is in Table 3. There are two basic components in grid-based detectors: grid-based representations and grid-based neural networks.

Grid-based representations There are 3 major types of grid representations: voxels, pillars, and BEV feature maps.

Voxels If we rasterize the detection space into a regular 3D grid, voxels are the grid cells. A voxel can be non-empty if point clouds fall into this grid cell. Since point clouds are sparsely distributed, most voxel cells in the 3D space are empty and contain no point. In practical applications, only those non-empty voxels are stored and utilized for feature extraction. VoxelNet Zhou and Tuzel (2018) is a pioneering work that utilizes sparse voxel grids and proposes a novel voxel feature encoding (VFE) layer to extract features from the points inside a voxel cell. A similar voxel encoding strategy has been adopted by a series of following works (Zhu et al., 2019; Ge et al., 2020; Yin et al., 2021a; Wang & Solomon, 2021; Zheng et al., 2021a; Li et al., 2021a; Deng et al., 2021b; Shi et al., 2020b). In addition, there are two categories of approaches trying to improve the voxel representation for 3D object detection: (1) Multi-view voxels. Some methods propose a dynamic voxelization and fusion scheme from diverse views, e.g. from both the bird's-eye view and the perspective view (Zhou et al., 2020c), from the cylindrical and spherical view (Chen et al., 2020a), from the range view (Deng et al., 2022). (2) Multi-scale voxels. Some papers generate voxels of different scales (Ye et al., 2020a) or use reconfigurable voxels (Wang et al., 2020b).

Pillars Pillars can be viewed as special voxels in which the voxel size is unlimited in the vertical direction. Pillar features can be aggregated from points through a PointNet (Qi et al., 2017a) and then scattered back to construct a 2D BEV image for feature extraction. PointPillars Lang et al. (2019) is a seminal work that introduces the pillar representation and is followed by Wang et al. (2020f), Fan et al. (2022).

BEV feature maps Bird's-eye view feature map is a dense 2D representation, where each pixel corresponds to a specific region and encodes the points information in this region. BEV

Table 3 A taxonomy of grid-based detection methods based on models and data representations

Method	Representation			Encoder			Neural Networks			
	Voxels	BEV maps	Pillars	Voxelization	Projection	PointNet	3D CNN	2D CNN	Head	Transformer
Vote3D Wang and Posner (2015)	✓			✓			✓			
Vote3Deep Engelcke et al. (2017)	✓			✓			✓			
3D-FCN Li (2017)	✓			✓			✓			
VeloFCN Li et al. (2016)		✓			✓			✓		
BirdNet Beltrán et al. (2018)		✓			✓			✓		
PIXOR Yang et al. (2018b)		✓			✓			✓		
HDNet Yang et al. (2018a)		✓			✓			✓		
VoxelNet Zhou and Tuzel (2018)	✓	✓		✓		✓	✓	✓		
SECOND Yan et al. (2018)	✓	✓		✓			✓	✓		
MVF Zhou et al. (2020c)	✓	✓		✓				✓		
PointPillars Lang et al. (2019)		✓	✓			✓		✓		
Pillar-OD Wang et al. (2020f)		✓	✓			✓		✓		
Part-A^2 Net Shi et al. (2020b)	✓	✓		✓			✓	✓	✓	
Voxel R-CNN Deng et al. (2021b)	✓	✓		✓			✓	✓	✓	
CenterPoint Yin et al. (2021a)	✓	✓		✓			✓	✓		
Voxel Transformer Mao et al. (2021c)	✓	✓		✓				✓		✓
SST Fan et al. (2022)	✓	✓		✓				✓		✓
SWFormer Sun et al. (2022)	✓	✓		✓						✓

feature maps can be obtained from voxels and pillars by projecting the 3D features into the bird's-eye view, or they can be directly obtained from raw point clouds by summarizing points statistics within the pixel region. The commonly-used statistics include binary occupancy (Yang et al., 2018b, a; Aghdam et al., 2021) and the height and density of local point cloud (Chen et al., 2017b; Beltrán et al., 2018; Zeng et al., 2018; Ali et al., 2018; Simony et al., 2018; Zhang et al., 2019; Barrera et al., 2020; Li et al., 2016).

Grid-based neural networks There are 2 major types of grid-based networks: 2D convolutional neural networks for BEV feature maps and pillars, and 3D sparse neural networks for voxels.

2D convolutional neural networks Conventional 2D convolutional neural networks can be applied to the BEV feature map to detect 3D objects from the bird's-eye view. In most works, the 2D network architectures are generally adapted from those successful designs in 2D object detection, e.g. ResNet (He et al., 2016) adopted in Yang et al. (2018b), Region Proposal Network (RPN) (Ren et al., 2015a) and Feature Pyramid Network (FPN) (Lin et al., 2017a) in Beltrán et al. (2018), Barrera et al. (2020), Lang et al. (2019), Kuang et al. (2020), Simony et al. (2018), Li et al. (2021a), and spatial attention in Li et al. (2021b), Liu et al. (2020d), Ye et al. (2020c).

3D sparse neural networks 3D sparse convolutional neural networks are based on two specialized 3D convolutional operators: sparse convolutions and submanifold convolutions (Graham et al., 2018), which can efficiently conduct 3D convolutions only on those non-empty voxels. Compared to Wang and Posner (2015), Engelcke et al. (2017), Najibi et al. (2020), Li (2017) that perform standard 3D convolutions on the whole voxel space, sparse convolutional operators are highly efficient and can obtain a real-time inference speed. SECOND Yan et al. (2018) is a seminal work that implements these two sparse operators with GPU-based hash tables and builds a sparse convolutional network to extract 3D voxel features. This network architecture has been applied in numerous works (Zhu et al., 2019; Yin et al., 2021a; Yi et al., 2020; Wang & Solomon, 2021; Ge et al., 2020; Chen et al., 2020b; Zhu et al., 2020; Yan et al., 2018; Deng et al., 2021b; Zheng et al., 2021a) and becomes the most widely-used backbone network in voxel-based detectors. There is also a series of works trying to improve the sparse operators (Chen et al., 2022b), extend (Yan et al., 2018) into a two-stage detector (Shi et al., 2020b; Deng et al., 2021b), and introduce the Transformer (Vaswani et al., 2017) architecture into voxel-based detection (Mao et al., 2021c; Fan et al., 2022).

Analysis: pros and cons of different grid representations In contrast to the 2D representations like BEV feature maps and pillars, voxels contain more structured 3D information. In addition, deep voxel features can be learned through a 3D sparse network. However, a 3D neural network brings additional time and memory costs. BEV feature map is the most efficient grid representation that directly projects point cloud into a 2D pseudo image without specialized 3D operators like sparse convolutions or pillar encoding. 2D detection techniques can also be seamlessly applied to BEV feature maps without much modification. BEV-based detection methods generally can obtain high efficiency and a real-time inference speed. However, simply summarizing points statistics inside pixel regions loses too much 3D information, which leads to less accurate detection results compared to voxel-based detection. Pillar-based detection approaches leverage PointNet to encode 3D points information inside a pillar cell, and the features are then scattered back into a 2D pseudo image for efficient detection, which balances the effectiveness and efficiency of 3D object detection.

Analysis: challenges of the grid-based detection methods A critical problem that all grid-based methods have to face is choosing the proper size of grid cells. Grid representations are essentially discrete formats of point clouds by converting the continuous point coordinates into discrete grid indices. The quantization process inevitably loses some 3D information and its efficacy largely depends on the size of grid cells: smaller grid size yields high resolution grids, and hence maintains more fine-grained details that are crucial to accurate 3D object detection. Nevertheless, reducing the size of grid cells leads to a quadratic increase in memory consumption for the 2D grid representations like BEV feature maps or pillars. As for the 3D grid representation like voxels, the problem can become more severe. Therefore, how to balance the efficacy brought by smaller grid sizes and the efficiency influenced by the memory increase remains an open challenge to all grid-based 3D object detection methods.

3.1.3 Point-Voxel Based 3D Object Detection

Point-voxel based approaches resort to a hybrid architecture that leverages both points and voxels for 3D object detection. Those methods can be divided into two categories: the single-stage and two-stage detection frameworks. An illustration of the two categories is shown in Fig. 6 and a taxonomy is in Table 4.

Single-stage point-voxel detection frameworks Single-stage point-voxel based 3D object detectors try to bridge the features of points and voxels with the point-to-voxel and voxel-to-point transform in the backbone networks. Points contain fine-grained geometric information and voxels are efficient for computation, and combining them together in the feature extraction stage naturally benefits from both two representations. The idea that leverages point-voxel feature fusion in backbones has been explored by many works, with the contributions like point-voxel convolutions (Liu et al., 2019b; Tang et al., 2020), auxiliary point-based networks (He

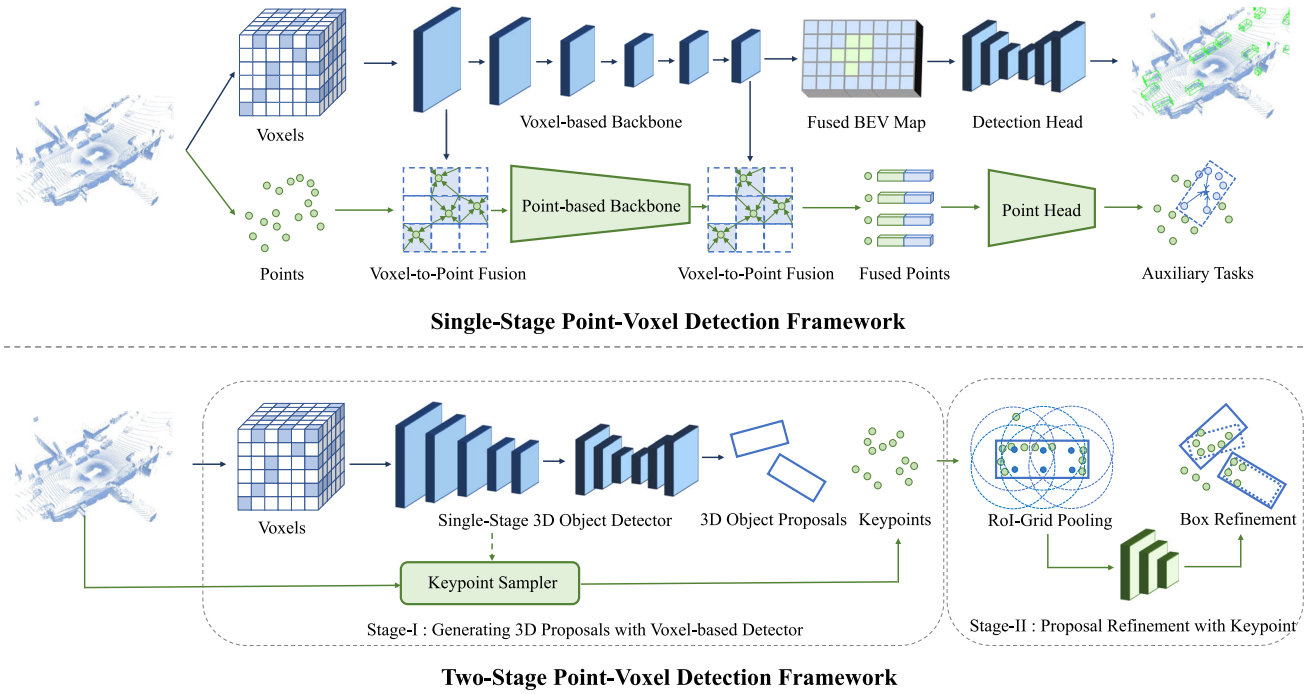

Fig. 6 An illustration of point-voxel based 3D object detection methods

et al., 2020a; Li et al., 2021c; Deng et al., 2021c), and multi-scale feature fusion (Miao et al., 2021; Noh et al., 2021; Guan et al., 2022).

Two-stage point-voxel detection frameworks Two-stage point-voxel based 3D object detectors resort to different data representations for different detection stages. Specifically, at the first stage, they employ a voxel-based detection framework to generate a set of 3D object proposals. In the second stage, keypoints are first sampled from the input point cloud, and then the 3D proposals are further refined from the keypoints through novel point operators. PV-RCNN (Shi et al., 2020a) is a seminal work that adopts (Yan et al., 2018) as the first-stage detector, and the RoI-grid pooling operator is proposed for the second-stage refinement. The following works try to improve the second-stage head with novel modules and operators, e.g. RefinerNet (Chen et al., 2019c), VectorPool (Shi et al., 2021a), point-wise attention (Wang et al., 2020a), scale-aware pooling (Li et al., 2021g), RoI-grid attention (Mao et al., 2021a), channel-wise Transformer (Sheng et al., 2021), and point density-aware refinement module (Hu et al., 2022).

Analysis: potentials and challenges of the point-voxel based methods The point-voxel based methods can naturally benefit from both the fine-grained 3D shape and structure information obtained from points and the computational efficiency brought by voxels. However, some challenges still exist in these methods. For the hybrid point-voxel backbones, the fusion of point and voxel features generally relies on the voxel-to-point and point-to-voxel transform mechanisms, which can bring non-negligible time costs. For the two-stage point-voxel detection frameworks, a critical challenge is how to efficiently aggregate point features for 3D proposals, as the existing modules and operators are generally time-consuming. In conclusion, compared to the pure voxel-based detection approaches, the point-voxel based detection

Table 4 A taxonomy of point-voxel based detection methods

Method	Contribution
Single-Stage Detection Framework	
PVCNN Liu et al. (2019b)	Point-Voxel Convolution
SPVNAS Tang et al. (2020)	Sparse Point-Voxel Convolution
SA-SSD He et al. (2020a)	Auxiliary Point Network
PVGNet Miao et al. (2021)	Point-Voxel-Grid Fusion
Two-Stage Detection Framework	
Fast Point R-CNN Chen et al. (2019c)	RefinerNet
PV-RCNN Shi et al. (2020a)	RoI-grid Pooling
PV-RCNN++ Shi et al. (2021a)	VectorPool
Pyramid R-CNN Mao et al. (2021a)	RoI-grid Attention
LiDAR R-CNN Li et al. (2021g)	Scale-aware Pooling
CT3D Sheng et al. (2021)	Channel-wise Transformer

methods can obtain a better detection accuracy while at the cost of increasing the inference time.

3.1.4 Range-Based 3D Object Detection

Range image is a dense and compact 2D representation in which each pixel contains 3D distance information instead of RGB values. Range-based methods address the detection problem from two aspects: designing new models and operators that are tailored for range images, and selecting suitable views for detection. An illustration of the range-based 3D object detection methods is shown in Fig. 7 and a taxonomy is in Table 5.

Range-based detection models Since range images are 2D representations like RGB images, range-based 3D object detectors can naturally borrow the models in 2D object detection to handle range images. LaserNet Meyer et al. (2019b) is a seminal work that leverages the deep layer aggregation network (DLA-Net) (Yu et al., 2018) to obtain multi-scale features and detect 3D objects from range images. Some papers also adopt other 2D object detection architectures, e.g. U-Net (Ronneberger et al., 2015) is applied in Meyer et al. (2020), Liang et al. (2020b), Sun et al. (2021), RPN (Ren et al., 2015a) and R-CNN (Ren et al., 2015b) are employed in Liang et al. (2020b), Bewley et al. (2020), FCN (Long et al., 2015) is used in Liang et al. (2021c), and FPN (Lin et al., 2017a) is leveraged in Fan et al. (2021).

Range-based operators Pixels of range images contain 3D distance information instead of color values, so the standard convolutional operator in conventional 2D network architectures is not optimal for range-based detection, as the pixels in a sliding window may be far away from each other in the 3D space. Some works resort to novel operators to effectively extract features from range pixels, including range dilated convolutions (Bewley et al., 2020), graph operators (Chai et al., 2021), and meta-kernel convolutions (Fan et al., 2021).

Views for range-based detection Range images are captured from the range view (RV), and ideally, the range view is a spherical projection of a point cloud. It has been a natural solution for many range-based approaches (Meyer et al., 2019b; Bewley et al., 2020; Fan et al., 2021; Chai et al., 2021) to detect 3D objects directly from the range view. Nevertheless, detection from the range view will inevitably suffer from the occlusion and scale-variation issues brought by the spherical projection. To circumvent these issues, many methods have been working on leveraging other views for predicting 3D objects, e.g. the cylindrical view (CYV) leveraged in Rapoport-Lavie and Raviv (2021), a combination of the range-view, bird's-eye view (BEV), and/or point-view (PV) adopted in Liang et al. (2021c), Sun et al. (2021), Meyer et al. (2020), Liang et al. (2020b).

Analysis: potentials and challenges of the range-based methods Range image is a dense and compact 2D representation, so the conventional or specialized 2D convolutions can be seamlessly applied on range images, which makes the feature extraction process quite efficient. Nevertheless, compared to bird's-eye view detection, detection from the range view is vulnerable to occlusion and scale variation. Hence, feature extraction from the range view and object detection from the bird's eye view becomes the most practical solution to range-based 3D object detection.

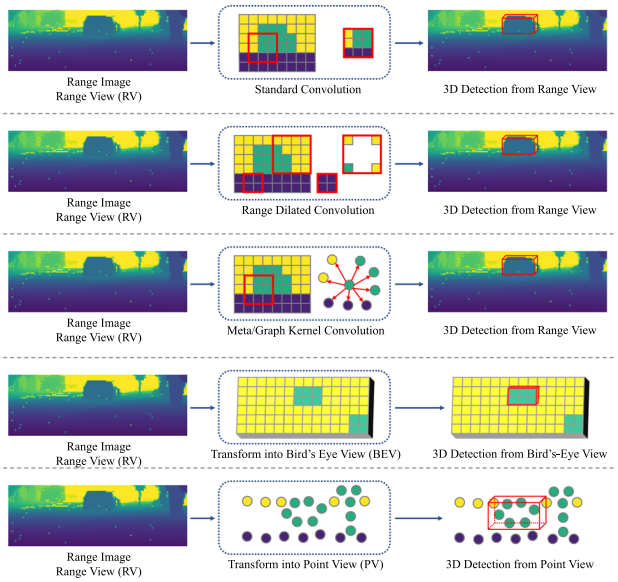

Fig. 7 An illustration of range-based 3D object detection

3.2 Learning Objectives for 3D Object Detection

Problem and Challenge Learning objectives are critical in object detection. Since 3D objects are quite small relative to the whole detection range, special mechanisms to enhance the localization of small objects are strongly required in 3D detection. On the other hand, considering point cloud is sparse and objects normally have incomplete shapes, accurately estimating the centers and sizes of 3D objects is a long-standing challenge.

3.2.1 Anchor-Based 3D Object Detection

Anchors are pre-defined cuboids with fixed shapes that can be placed in the 3D space. 3D objects can be predicted based on the positive anchors that have a high intersection over union (IoU) with ground truth. We will introduce the anchor-based 3D object detection methods from the aspect of anchor configurations and loss functions. An illustration of anchor-based learning objectives is shown in Fig. 8 and a taxonomy is in Table 6.

Prerequisites The ground truth 3D objects can be represented as $[x^g, y^g, z^g, l^g, w^g, h^g, \theta^g]$ with the class cls^g. The

Table 5 A taxonomy of range-based detection methods based on views, models, and operators

Method	View	Operator	Model	Note
LaserNet Meyer et al. (2019b)	RV	Convolution	DLA-Net	–
Rapoport-Lavie et al. Rapoport-Lavie and Raviv (2021)	RV, CYV	Convolution	Range-Guided Net	–
LaserFlow Meyer et al. (2020)	RV, BEV	Convolution	U-Net	multi-sweep fusion
RangeRCNN Liang et al. (2020b)	RV, PV, BEV	Dilated Convolution	U-Net, RPN, RCNN	–
RangeIoUDet Liang et al. (2021c)	RV, PV, BEV	Convolution	FCN, PointNet	Point-wise segmentation
RCD Bewley et al. (2020)	RV	Conditioned Dilated Convolution	RPN, RCNN	–
RangeDet Fan et al. (2021)	RV	Meta Kernel Convolution	FPN	–
PPC Chai et al. (2021)	RV	Graph Kernel Convolution	DLA-Net	–
RSN Sun et al. (2021)	RV, BEV	Convolution	U-Net, VoxelNet	Range-based segmentation

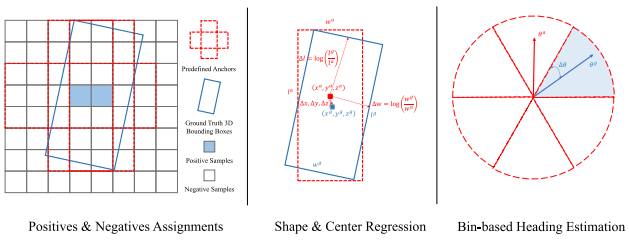

Fig. 8 An illustration of anchor-based learning objectives

anchors $[x^a, y^a, z^a, l^a, w^a, h^a, \theta^a]$ are used to generate predicted 3D objects $[x, y, z, l, w, h, \theta]$ with a predicted class probability p.

Anchor configurations Anchor-based 3D object detection approaches generally detect 3D objects from the bird's-eye view, in which 3D anchor boxes are placed at each grid cell of a BEV feature map. 3D anchors normally have a fixed size for each category, since objects of the same category have similar sizes.

Loss functions The anchor-based methods employ the classification loss L_{cls} to learn the positive and negative anchors, and the regression loss L_{reg} is utilized to learn the size and location of an object based on a positive anchor. Additionally, L_θ is applied to learn the object's heading angle. The loss function is

$$L_{det} = L_{cls} + L_{reg} + L_\theta. \tag{4}$$

VoxelNet Zhou and Tuzel (2018) is a seminal work that leverages the anchors that have a high IoU with the ground truth 3D objects as positive anchors, and the other anchors are treated as negatives. To accurately classify those positive and negative anchors, for each category, the binary cross entropy loss can be applied to each anchor on the BEV feature map, which can be formulated as

$$L_{cls}^{bce} = -[q \cdot \log(p) + (1-q) \cdot \log(1-p)], \tag{5}$$

where p is the predicted probability for each anchor and the target q is 1 if the anchor is positive and 0 otherwise. In addition to the binary cross entropy loss, the focal loss (Lin et al., 2017b; Yun et al., 2019) has also been employed to enhance the localization ability:

$$L_{cls}^{focal} = -\alpha(1-p)^\gamma \log(p), \tag{6}$$

where $\alpha = 0.25$ and $\gamma = 2$ are adopted in most works.

The regression targets can be further applied to those positive anchors to learn the sizes and locations of 3D objects:

$$\begin{aligned} &\Delta x = \frac{x^g - x^a}{d^a}, \Delta y = \frac{y^g - y^a}{d^a}, \Delta z = \frac{z^g - z^a}{h^a}, \\ &\Delta l = \log(\frac{l^g}{l^a}), \Delta w = \log(\frac{w^g}{w^a}), \Delta h = \log(\frac{h^g}{h^a}), \end{aligned} \tag{7}$$

Table 6 A taxonomy of anchor-based methods based on loss functions

Loss Function	Methods
L_{cls}^{bce} (Eq. 5)	(Simony et al., 2018; Ali et al., 2018; Beltrán et al., 2018; Zeng et al., 2018; Zhou & Tuzel, 2018; Barrera et al., 2020; Yang et al., 2019, 2018c; Chen et al., 2019c; He et al., 2020c)
L_{cls}^{focal} (Eq. 6)	(Yan et al., 2018; Lang et al., 2019; Zhou et al., 2020c, 2019a; Yun et al., 2019; Ye et al., 2020a; Du et al., 2020; Zhu et al., 2020; Wang et al., 2020b; Yi et al., 2020; Mao et al., 2021c; Deng et al., 2021b; Zheng et al., 2021a; Du et al., 2021; Liang et al., 2020b; Ngiam et al., 2019; Shi et al., 2020a; He et al., 2020a; Wang et al., 2020a; Miao et al., 2021; Noh et al., 2021; Mao et al., 2021a; Shi et al., 2021a)
L_{reg} (Eq. 8)	(Simony et al., 2018; Ali et al., 2018; Beltrán et al., 2018; Zeng et al., 2018; Yan et al., 2018; Zhou & Tuzel, 2018; Lang et al., 2019; Zhou et al., 2020c; Yun et al., 2019; Ye et al., 2020a; Du et al., 2020; Zhu et al., 2020; Wang et al., 2020b; Yi et al., 2020; Barrera et al., 2020; Mao et al., 2021c; Deng et al., 2021b; Zheng et al., 2021a; Du et al., 2021; Liang et al., 2020b; Yang et al., 2019; Ngiam et al., 2019; Yang et al., 2018c; Chen et al., 2019c; Shi et al., 2020a; He et al., 2020a; Wang et al., 2020a; He et al., 2020c; Miao et al., 2021; Noh et al., 2021; Mao et al., 2021a; Shi et al., 2021a)
L_{θ}^{1} (Eq. 9)	(Ali et al., 2018; Zhou & Tuzel, 2018; Ye et al., 2020a; Du et al., 2020, 2021; Chen et al., 2019c)
L_{θ}^{2} (Eq. 10)	(Beltrán et al., 2018; Wang et al., 2020b; Yi et al., 2020; Barrera et al., 2020; Mao et al., 2021c; Zheng et al., 2021a; Liang et al., 2020b; Yang et al., 2019, 2018c; Shi et al., 2020a; He et al., 2020a; Wang et al., 2020a; Mao et al., 2021a; Shi et al., 2021a)
L_{θ}^{3} (Eq. 11)	(Simony et al., 2018; Zeng et al., 2018; Yan et al., 2018; Lang et al., 2019; Zhou et al., 2020c; Zhu et al., 2020; Deng et al., 2021b; Ngiam et al., 2019; He et al., 2020c; Miao et al., 2021; Noh et al., 2021)
L_{IoU} (Eq. 12)	(Zhou et al., 2019a)
L_{corner} (Eq. 13)	(Ye et al., 2020a; Liang et al., 2020b; Yang et al., 2019, 2018c)

where $d^a = \sqrt{(l^a)^2 + (w^a)^2}$ is the diagonal length of an anchor from the bird's-eye view. Then the SmoothL1 loss (Ren et al., 2015b) is adopted to regress the targets, which is represented as

$$L_{reg} = \sum_{\substack{u \in \{x,y,z,l,w,h\}, \\ v \in \{\Delta x, \Delta y, \Delta z, \Delta l, \Delta w, \Delta h\}}} \text{SmoothL1}(u - v). \tag{8}$$

To learn the heading angle θ, the radian orientation offset can be directly regressed with the SmoothL1 loss:

$$\begin{aligned} \Delta\theta &= \theta^g - \theta^a, \\ L_\theta &= \text{SmoothL1}(\theta - \Delta\theta). \end{aligned} \tag{9}$$

However, directly regressing the radian offset is normally hard due to the large regression range. Alternatively, the bin-based heading estimation (Qi et al., 2018) is a better solution to learn the heading angle, in which the angle space is first divided into bins, and bin-based classification L_{dir} and residual regression are employed:

$$L_\theta = L_{dir} + \text{SmoothL1}(\theta - \Delta\theta'), \tag{10}$$

where $\Delta\theta'$ is the residual offset within a bin. The sine function can also be utilized to encode the radian offset:

$$\Delta\theta = \sin(\theta^g - \theta^a), \tag{11}$$

and L^θ can be computed following Eqn. 9 or Eqn. 10.

In addition to the loss functions that learn the objects' sizes, locations, and orientations separately, the intersection over union (IoU) loss (Zhou et al., 2019a) that considers all object parameters as a whole can also be applied in 3D object detection:

$$L_{IoU} = 1 - IoU(b^g, b), \tag{12}$$

where b^g and b are the ground truth and predicted 3D bounding boxes, and $IoU(\cdot)$ calculates the 3D IoU in a differential manner. Apart from the IoU loss, the corner loss (Qi et al., 2018) is also introduced to minimize the distances between the eight corners of the ground truth and predicted boxes, that is

$$L_{corner} = \sum_{i=1}^{8} ||c_i^g - c_i||, \tag{13}$$

where c_i^g and c_i are the ith corner of the ground truth and predicted cuboid respectively.

Analysis: potentials and challenges of the anchor-based approaches The anchor-based methods can benefit from the prior knowledge that 3D objects of the same category should have similar shapes, so they can generate accurate object predictions with the help of 3D anchors. However, since 3D objects are relatively small with respect to the detection range, a large number of anchors are required to ensure complete coverage of the whole detection range, e.g. around 70k anchors are utilized in Yan et al. (2018) on the KITTI (Geiger et al., 2012) dataset. Furthermore, for those extremely small objects such as pedestrians and cyclists, applying anchor-based assignments can be quite challenging. Considering the fact that anchors are generally placed at the center of each grid cell, if the grid cell is large and objects in the cell are small, the anchor of this cell may have a low IoU with the small objects, which may hamper the training process.

3.2.2 Anchor-Free 3D Object Detection

Anchor-free approaches eliminate the complicated anchor designs and can be flexibly applied to diverse views, e.g. the bird's-eye view, point view, and range view. An illustration of anchor-free learning objectives is shown in Fig. 9 and a taxonomy is in Table 7. The major difference between the anchor-based and anchor-free methods lies in the selection of positive and negative samples. We will introduce the anchor-free methods from the perspective of positive assignments, including grid-based, point-based, range-based, and set-to-set assignments. We still adopt the notations in Sect. 3.2.1 for simplicity.

Grid-based assignment In contrast to the anchor-based methods that rely on the IoUs with anchors to determine the positive and negative samples, the anchor-free methods leverage various grid-based assignment strategies for BEV grid cells, pillars, and voxels. PIXOR Yang et al. (2018b) is a pioneering work that leverages the grid cells inside the ground truth 3D objects as positives, and the others as negatives. This inside-object assignment strategy is adopted in Wang et al. (2020f), and further improved in Ge et al. (2020), Hu et al. (2021), Chen et al. (2020b) by selecting the grid cells nearest to the object center. CenterPoint Yin et al. (2021a) utilizes a Gaussian kernel at each object center to assign positive labels. These methods can still use Eqn. 5 or Eqn. 6 as the classification loss, and the regression target is

$$\Delta = [dx, dy, z^g, \log(l^g), \log(w^g), \log(h^g), \sin(\theta^g), \cos(\theta^g)], \tag{14}$$

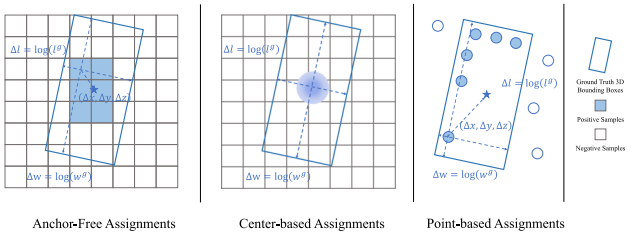

Fig. 9 An illustration of anchor-free learning objectives

where dx and dy are the offsets between positive grid cells and object centers. The SmoothL1 loss is leveraged to regress Δ.

Point-based assignment Most point-based detection approaches resort to the anchor-free and point-based assignment strategy, in which the points are first segmented and those foreground points inside or near 3D objects are selected as positive samples, and 3D bounding boxes are finally learned from those foreground points. This foreground point segmentation strategy has been adopted in most point-based detectors (Shi et al., 2019; Yang et al., 2020c, 2018c; Pan et al., 2021), with improvements such as adding centerness scores (Yang et al., 2020c), etc.

Range-based assignment Anchor-free assignments can also be employed on range images. A common solution is to select the range pixels inside 3D objects as positive samples, which has been adopted in Meyer et al. (2019b), Fan et al. (2021). Different from other methods where the regression targets are based on the global 3D coordinate system, the range-based methods resort to an object-centric coordinate system for regression. Eqn. 14 can still be applied in these methods with an additional coordinate transform.

Set-to-set assignment DETR Carion et al. (2020) is an influential 2D detection method that introduces a set-to-set assignment strategy to automatically assign the predictions to the respective ground truths via the Hungarian algorithm (Kuhn, 1955):

$$\mathcal{M}^* = \underset{\mathcal{M}}{\operatorname{argmin}} \sum_{(i \to j) \in \mathcal{M}} L_{det}(b_i^g, b_j), \tag{15}$$

where $\mathcal{M}$ is a one-to-one mapping from each positive sample to a 3D object. The set-to-set assignments have also been explored in 3D object detection approaches (Misra et al., 2021; Wang & Solomon, 2021; Xue et al., 2022), and (Xue et al., 2022) further introduces a novel cost function for the Hungarian matching.

Analysis: potentials and challenges of the anchor-free approaches The anchor-free detection methods abandon the complicated anchor design and exhibit stronger flexibility in terms of the assignment strategies. With the anchor-free assignments, 3D objects can be predicted directly on various representations, including points, range pixels, voxels,

Table 7 A taxonomy of anchor-free detection methods based on the sample types for prediction and the assignment strategies

Method	Samples for prediction	Positive samples selection
PIXOR Yang et al. (2018b)	BEV grid cells	Inside objects
CenterPoint Yin et al. (2021a)	BEV grid cells	Gaussian radii on centers
ObjectDGCNN Wang and Solomon (2021)	BEV grid cells	Bipartite matching
Pillar-OD Wang et al. (2020f)	Pillars	Inside objects
HotSpotNet Chen et al. (2020b)	Voxels	K-nearest to object centers
PointRCNN Shi et al. (2019)	Points	foreground
3DSSD Yang et al. (2020c)	Points	Foreground & near centers
LaserNet Meyer et al. (2019b)	Range pixels	inside objects
RangeDet Fan et al. (2021)	Range pixels	Inside objects

pillars, and BEV grid cells. The learning process is also greatly simplified without introducing additional shape priors. Among those anchor-free methods, the center-based methods (Yin et al., 2021a) have shown great potential in detecting small objects and have outperformed the anchor-based detection methods on the widely used benchmarks (Caesar et al., 2020; Sun et al., 2020c).

Despite these merits, a general challenge to the anchor-free methods is to properly select positive samples to generate 3D object predictions. In contrast to the anchor-based methods that only select those high IoU samples, the anchor-free methods may possibly select some bad positive samples that yield inaccurate object predictions. Hence, careful design to filter out those bad positives is important in most anchor-free methods.

3.2.3 3D Object Detection with Auxiliary Tasks

Numerous approaches resort to auxiliary tasks to enhance the spatial features and provide implicit guidance for accurate 3D object detection. The commonly used auxiliary tasks include semantic segmentation, intersection over union prediction, object shape completion, and object part estimation.

Semantic segmentation Semantic segmentation can help 3D object detection in 3 aspects: (1) Foreground segmentation could provide implicit information on objects' locations. Point-wise foreground segmentation has been broadly adopted in most point-based 3D object detectors (Shi et al., 2019; Zhou et al., 2020a; Yang et al., 2019; He et al., 2020a) for proposal generation. (2) Spatial features can be enhanced by segmentation. In Yi et al. (2020), a semantic context encoder is leveraged to enhance spatial features with semantic knowledge. (3) Semantic segmentation can be utilized as a pre-processing step to filter out background samples and make 3D object detection more efficient. Yang et al. (2018c) and Sun et al. (2021) leverage semantic segmentation to remove those redundant points to speed up the subsequent detection model.

Table 8 A taxonomy of detection methods based on auxiliary tasks

Auxiliary Task	Methods
Semantic segmentation	(Shi et al., 2019; Yang et al., 2019; He et al., 2020a; Zhou et al., 2020a; Yi et al., 2020; Yang et al., 2018c; Sun et al., 2021)
IoU prediction	(Zheng et al., 2021a, b; Liang et al., 2021c; Ge et al., 2020; Hu et al., 2021)
Object shape completion	(Najibi et al., 2020; Zhu et al., 2020; Xu et al., 2021b, a)
Object part estimation	(Chen et al., 2020b; Shi et al., 2020b)

IoU prediction Intersection over union (IoU) can serve as a useful supervisory signal to rectify the object confidence scores. Zheng et al. (2021a) proposes an auxiliary branch to predict an IoU score S_{IoU} for each detected 3D object. During inference, the original confidence scores $S_{conf} = S_{cls}$ from the conventional classification branch are further rectified by the IoU scores S_{IoU}:

$$S_{conf} = S_{cls} \cdot (S_{IoU})^{\beta}, \tag{16}$$

where the hyper-parameter β controls the degrees of suppressing the low-IoU predictions and enhancing the high-IoU predictions. With the IoU rectification, the high-quality 3D objects are easier to be selected as the final predictions. Similar designs have also been adopted in Zheng et al. (2021b), Liang et al. (2021c), Ge et al. (2020), Hu et al. (2021).

Object shape completion Due to the nature of LiDAR sensors, faraway objects generally receive only a few points on their surfaces, so 3D objects are generally sparse and incomplete. A straightforward way of boosting the detection performance is to complete object shapes from sparse point clouds. Complete shapes could provide more useful information for accurate and robust detection. Many shape completion techniques have been proposed in 3D detection,

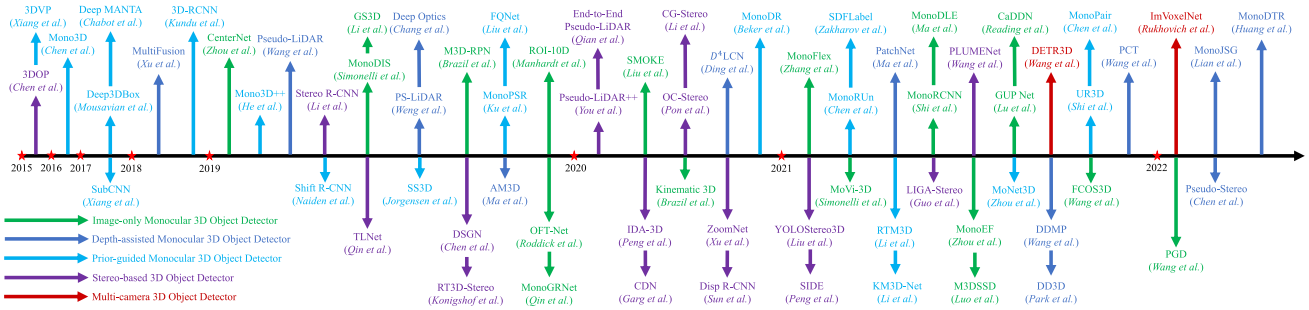

Fig. 10 Chronological overview of the camera-based 3D object detection methods

including a shape decoder (Najibi et al., 2020), shape signatures (Zhu et al., 2020), and a probabilistic occupancy grid (Xu et al., 2021b, a).

Object part estimation Identifying the part information inside objects is helpful in 3D object detection, as it reveals more fine-grained 3D structure information of an object. Object part estimation has been explored in some works (Chen et al., 2020b; Shi et al., 2020b).

Analysis: future prospects of multitask learning for 3D object detection 3D object detection is innately correlated with many other 3D perception and generation tasks. Multitask learning of 3D detection and segmentation is more beneficial compared to training 3D object detectors independently, and shape completion can also help 3D object detection. There are also other tasks that can help boost the performance of 3D object detectors. For instance, scene flow estimation could identify static and moving objects, and tracking the same 3D object in a point cloud sequence yields a more accurate estimation of this object. Hence, it will be promising to integrate more perception tasks into the existing 3D object detection pipeline.

4 Camera-Based 3D Object Detection

In this section, we introduce camera-based 3D object detection methods. In Sect. 4.1, we review and analyze the monocular 3D object detection methods, which can be further divided into the image-only, depth-assisted, and prior-guided approaches. In Sect. 4.2, we investigate the 3D object detection methods based on stereo images. In Sect. 4.3, we introduce the 3D object detection methods with multiple cameras. A chronological overview of the camera-based 3D object detection methods is shown in Fig. 10.

4.1 Monocular 3D Object Detection

Problem and Challenge Detecting objects in the 3D space from monocular images is an ill-posed problem since a single image cannot provide sufficient depth information. Accurately predicting the 3D locations of objects is the major challenge in monocular 3D object detection. Many endeavors have been made to tackle the object localization problem, e.g. inferring depth from images, leveraging geometric constraints and shape priors. Nevertheless, the problem is far from being solved. Monocular 3D detection methods still perform much worse than the LiDAR-based methods due the poor 3D localization ability, which leaves an open challenge to the research community.

4.1.1 Image-Only Monocular 3D Object Detection

Inspired by the 2D detection approaches, a straightforward solution to monocular 3D object detection is to directly regress the 3D box parameters from images via a convolutional neural network. The direct-regression methods naturally borrow designs from the 2D detection network architectures, and can be trained in an end-to-end manner. These approaches can be divided into the single-stage/two-stage, or anchor-based/anchor-free methods. An illustration of image-only 3D object detection is shown in Fig. 11 and a taxonomy is in Table 9.

Single-stage anchor-based methods Anchor-based monocular detection approaches rely on a set of 2D-3D anchor boxes placed at each image pixel, and use a 2D convolutional neural network to regress object parameters from the anchors. Specifically, for each pixel $[u, v]$ on the image plane, a set of 3D anchors $[w^a, h^a, l^a, \theta^a]_{3D}$, 2D anchors $[w^a, h^a]_{2D}$, and depth anchors d^a are pre-defined. An image is passed through a convolutional network to predict the 2D box offsets $\delta_{2D} = [\delta_x, \delta_y, \delta_w, \delta_h]_{2D}$ and the 3D box offsets $\delta_{3D} = [\delta_x, \delta_y, \delta_d, \delta_w, \delta_h, \delta_l, \delta_\theta]_{3D}$ based on each anchor. Then, the 2D bounding boxes $b_{2D} = [x, y, w, h]_{2D}$ can be decoded as

$$\begin{aligned} [x, y]_{2D} &= [u, v] + [\delta_x, \delta_y]_{2D} \cdot [w^a, h^a]_{2D}, \\ [w, h]_{2D} &= e^{[\delta_w, \delta_h]_{2D}} \cdot [w^a, h^a]_{2D}, \end{aligned} \tag{17}$$

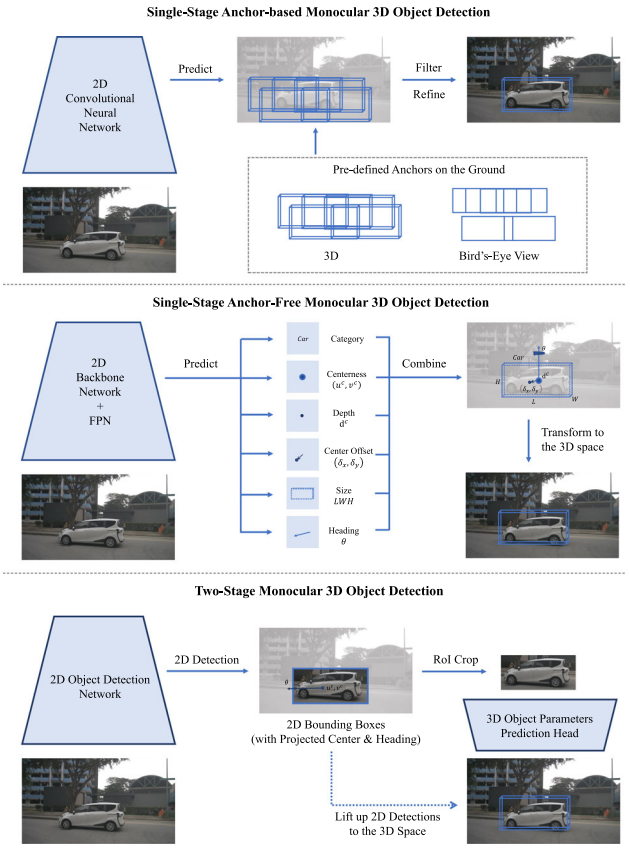

Fig. 11 An illustration of image-only monocular 3D object detection methods. Image samples are from (Wang et al., 2021g)

and the 3D bounding boxes $b_{3D} = [x, y, z, l, w, h, \theta]_{3D}$ can be decoded from the anchors and δ_{3D}:

$$\begin{aligned} [u^c, u^c] &= [u, v] + [\delta_x, \delta_y]_{3D} \cdot [w^a, h^a]_{2D}, \\ [w, h, l]_{3D} &= e^{[\delta_w, \delta_h, \delta_l]_{3D}} \cdot [w^a, h^a, l^a]_{3D}, \\ d^c &= d^a + \delta_{d_{3D}}, \theta_{3D} = \theta^a_{3D} + \delta_{\theta_{3D}}, \end{aligned} \tag{18}$$

where $[u^c, v^c]$ is the projected object center on the image plane. Finally, the projected center $[u^c, v^c]$ and its depth d^c are transformed into the 3D object center $[x, y, z]_{3D}$:

$$d^c \cdot \begin{bmatrix} u^c \\ v^c \\ 1 \end{bmatrix} = K \; T \; \begin{bmatrix} x \\ y \\ z \\ 1 \end{bmatrix}_{3D}, \tag{19}$$

where K and T are the camera intrinsics and extrinsics.

M3D-RPN Brazil and Liu (2019) is a seminal paper that proposes the anchor-based framework, and many papers have tried to improve this framework, e.g. extending it into video-based 3D detection (Brazil et al., 2020), introducing differential non-maximum suppression (Kumar et al., 2021), designing an asymmetric attention module (Luo et al., 2021a).

Single-stage anchor-free methods Anchor-free monocular detection approaches predict the attributes of 3D objects from images without the aid of anchors. Specifically, an image is passed through a 2D convolutional neural network and then multiple heads are applied to predict the object attributes separately. The prediction heads generally include a category head to predict the object's category, a keypoint head to predict the coarse object center $[u, v]$, an offset head to predict the center offset $[\delta_x, \delta_y]$ based on $[u, v]$, a depth head to predict the depth offset δ_d, a size head to predict the object size $[w, h, l]$, and an orientation head to predict the observation angle α. The 3D object center $[x, y, z]$ can be converted from the projected center $[u^c, v^c]$ and depth d^c:

$$\begin{aligned} &d^c = \sigma^{-1}(\frac{1}{\delta_d + 1}), u^c = u + \delta_x, v^c = v + \delta_y, \\ &d^c \cdot \begin{bmatrix} u^c \\ v^c \\ 1 \end{bmatrix} = K \; T \; \begin{bmatrix} x \\ y \\ z \\ 1 \end{bmatrix}_{3D}, \end{aligned} \tag{20}$$

where σ is the sigmoid function. The yaw angle θ of an object can be converted from the observation angle α using

$$\theta = \alpha + \arctan(\frac{x}{z}). \tag{21}$$

Table 9 A taxonomy of image-only monocular detection methods based on frameworks

Framework		Methods
Single-stage	Anchor-based	(Brazil & Liu, 2019; Brazil et al., 2020; Kumar et al., 2021; Luo et al., 2021a; Liu et al., 2019a)
	Anchor-free	(Zhou et al., 2019b; Mousavian et al., 2017; Liu et al., 2020c; Wang et al., 2021g, 2022a; Ma et al., 2021; Li et al., 2020c; Zhang et al., 2021c; Zhou et al., 2021; Roddick et al., 2019; Reading et al., 2021; Simonelli et al., 2020)
Two-stage		(Manhardt et al., 2019; Simonelli et al., 2019; Li et al., 2019a; Qin et al., 2019a; Shi et al., 2021b; Lu et al., 2021)

CenterNet Zhou et al. (2019b) first introduces the single-stage anchor-free framework for monocular 3D object detection. Many following papers work on improving this framework, including novel depth estimation schemes (Liu et al., 2020c; Wang et al., 2022a; Zhang et al., 2021c), an FCOS (Tian et al., 2019)-like architecture (Wang et al., 2021g), a new IoU-based loss function (Ma et al., 2021), key-points (Li et al., 2020c), pair-wise relationships (Chen et al., 2020f), camera extrinsics prediction (Zhou et al., 2021), and view transforms (Roddick et al., 2019; Reading et al., 2021; Simonelli et al., 2020).

Two-stage methods Two-stage monocular detection approaches generally extend the conventional two-stage 2D detection architectures to 3D object detection. Specifically, they utilize a 2D detector in the first stage to generate 2D bounding boxes from an input image. Then in the second stage, the 2D boxes are lifted up to the 3D space by predicting the 3D object parameters from the 2D RoIs. ROI-10D Manhardt et al. (2019) extends the conventional Faster R-CNN (Ren et al., 2015b) architecture with a novel head to predict the parameters of 3D objects in the second stage. A similar design paradigm has been adopted in many works with improvements like disentangling the 2D and 3D detection loss (Simonelli et al., 2019), predicting heading angles in the first stage (Li et al., 2019a), learning more accurate depth information (Qin et al., 2019a; Shi et al., 2021b; Lu et al., 2021).

Analysis: potentials and challenges of the image-only methods The image-only methods aim to directly regress the 3D box parameters from images via a modified 2D object detection framework. Since these methods take inspiration from the 2D detection methods, they can naturally benefit from the advances in 2D object detection and image-based network architectures. Most methods can be trained end-to-end without pre-training or post-processing, which is quite simple and efficient.

A critical challenge of the image-only methods is to accurately predict depth d^c for each 3D object. As shown in Wang et al. (2022a), simply replacing the predicted depth with ground truth yields more than 20% car AP gain on the KITTI (Geiger et al., 2012) dataset, while replacing other parameters only results in an incremental gain. This observation indicates that the depth error dominates the total errors and becomes the most critical factor hampering accurate monocular detection. Nevertheless, depth estimation from monocular images is an ill-posed problem, and the problem becomes severer with only box-level supervisory signals.

4.1.2 Depth-Assisted Monocular 3D Object Detection

Depth estimation is critical in monocular 3D object detection. To achieve more accurate monocular detection results, many papers resort to pre-training an auxiliary depth estimation network. Specifically, a monocular image is first passed through a pre-trained depth estimator, e.g. MonoDepth (Godard et al., 2017) or DORN (Fu et al., 2018), to generate a depth image. Then, there are mainly two categories of methods to deal with depth images and monocular images. The depth-image based methods fuse images and depth maps with a specialized neural network to generate depth-aware features that could enhance the detection performance. The pseudo-LiDAR based methods convert a depth image into a pseudo-LiDAR point cloud, and LiDAR-based detectors can then be applied to the point cloud to predict 3D objects. An illustration of depth-assisted monocular 3D object detection is shown in Fig. 12 and a taxonomy of these methods is in Table 10.

Depth-image based methods Most depth-image based methods leverage two backbone networks for RGB and depth images respectively. They obtain depth-aware image features by fusing the information from the two backbones with specialized operators. More accurate 3D bounding boxes can be learned from the depth-ware features and can be further refined with depth images. MultiFusion Xu and Chen (2018) is a pioneering work that introduces the depth-image based detection framework. Following papers adopt similar design paradigms with improvements in network architectures, operators, and training strategies, e.g. a point-based attentional network (Bao et al., 2019), depth-guided convolutions (Ding et al., 2020), depth-conditioned message passing (Wang et al., 2021d), disentangling appearance and localization features (Zou et al., 2021), and a novel depth pre-training framework (Park et al., 2021).

Pseudo-LiDAR based methods Pseudo-LiDAR based methods transform a depth image into a pseudo-LiDAR point cloud, and LiDAR-based detectors can then be employed to detect 3D objects from the point cloud. Pseudo-LiDAR point cloud is a data representation first introduced in Wang et al. (2019a), where they convert a depth map $\mathcal{D} \in R^{H \times W}$ into a pseudo point cloud $\mathcal{P} \in R^{HW \times 3}$. Specifically, for each pixel $[u, v]$ and its depth value d in a depth image, the corresponding 3D point coordinate $[x, y, z]$ in the camera coordinate system is computed as

$$x = \frac{(u - c_u) \times z}{f_u}, y = \frac{(v - c_v) \times z}{f_v}, z = d, \tag{22}$$

where $[c_u, c_v]$ is the camera principal point, and f_u and f_v are the focal lengths along the horizontal and vertical axis respectively. Thus $\mathcal{P}$ can be obtained by back-projecting each pixel in $\mathcal{D}$ into the 3D space. $\mathcal{P}$ is referred as the pseudo-LiDAR representation: it is essentially a 3D point cloud but is extracted from a depth image instead of a real LiDAR sensor. Finally, LiDAR-based 3D object detectors can be directly applied on the pseudo-LiDAR point cloud $\mathcal{P}$ to predict 3D objects. Many papers have worked on improving the

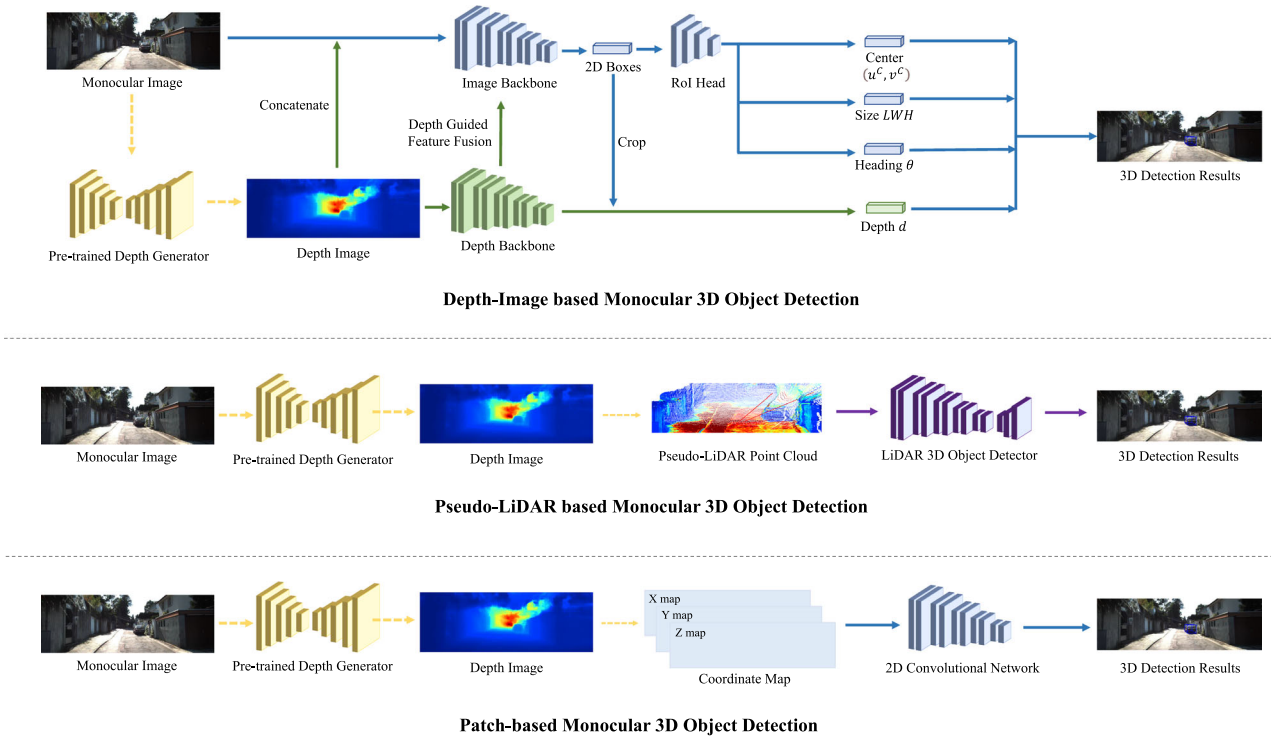

Fig. 12 An illustration of depth-assisted monocular 3D object detection methods. Image and depth samples are from (Ma et al., 2019a)

pseudo-LiDAR detection framework, including augmenting pseudo point cloud with color information (Ma et al., 2019a), introducing instance segmentation (Weng & Kitani, 2019), designing a progressive coordinate transform scheme (Wang et al., 2021e), improving pixel-wise depth estimation with separate foreground and background prediction (Wang et al., 2020d), domain adaptation from real LiDAR point cloud (Ye et al., 2020b), and a new physical sensor design (Chang & Wetzstein, 2019).

PatchNet Ma et al. (2020) challenges the conventional idea of leveraging the pseudo-LiDAR representation $\mathcal{P} \in R^{HW\times 3}$ for monocular 3D object detection. They conduct an in-depth investigation and provide an insightful observation that the power of pseudo-LiDAR representation comes from the coordinate transformation instead of the point cloud representation. Hence, a coordinate map $\mathcal{M} \in R^{H\times W\times 3}$ where each pixel encodes a 3D coordinate can attain a comparable monocular detection result with the pseudo-LiDAR point cloud representation. This observation enables us to directly apply a 2D neural network on the coordinate map to predict 3D objects, eliminating the need of leveraging the time-consuming LiDAR-based detectors on point clouds.

Analysis: potentials and challenges of the depth-assisted approaches The depth-assisted approaches pursue more

Table 10 A taxonomy of depth-assisted monocular detection methods based on data representations and detection networks (2D: convolutional networks; 3D: point cloud networks)

Method	Representation				Network	
	RGB	Depth	Pseudo LiDAR	Coord. Map	2D	3D
MultiFusion Xu and Chen (2018)	✓	✓			✓	
D^{4}LCN Ding et al. (2020)	✓	✓			✓	
DDMP Wang et al. (2021d)	✓	✓			✓	
Pseudo-LiDAR Wang et al. (2019a)			✓			✓
Deep Optics Chang and Wetzstein (2019)			✓			✓
AM3D Ma et al. (2019a)	✓		✓		✓	✓
Weng et al. Weng and Kitani (2019)	✓		✓		✓	✓
PatchNet Ma et al. (2020)				✓	✓	

accurate depth estimation by leveraging a pre-trained depth prediction network. Both the depth image representation and the pseudo-LiDAR presentation could significantly boost the monocular detection performance. Nevertheless, compared to the image-only methods that only require 3D box annotations, pre-training a depth prediction network requires expensive ground truth depth maps, and it also hampers the end-to-end training of the whole framework. Furthermore, pre-trained depth estimation networks suffer from poor generalization ability. Pretrained depth maps are usually not well calibrated on the target dataset and typically the scale needs to be adapted to the target dataset. Thus there remains a non-negligible domain gap between the source domain leveraged for depth pre-training and the target domain for monocular detection. Given the fact that driving scenarios are normally diverse and complex, pre-training depth networks on a restricted domain may not work well in real-world applications.

4.1.3 Prior-Guided Monocular 3D Object Detection

Numerous approaches try to tackle the ill-posed monocular 3D object detection problem by leveraging the hidden prior knowledge of object shapes and scene geometry from images. The prior knowledge can be learned by introducing pre-trained sub-networks or auxiliary tasks, and they can provide extra information or constraints to help accurately localize 3D objects. The broadly adopted prior knowledge includes object shapes, geometry consistency, temporal constraints, and segmentation information. An illustration of the prior types is shown in Fig. 13.

Object shapes Many methods resort to shape reconstruction of 3D objects directly from images. The reconstructed shapes can be further leveraged to determine the locations and poses of the 3D objects. There are 5 types of reconstructed representations: computer-aided design (CAD) models, wireframe models, signed distance function (SDF), points, and voxels.

Some papers (Zeeshan Zia et al., 2014; Chabot et al., 2017; He & Soatto, 2019) learn morphable wireframe models to represent 3D objects. Other works (Kundu et al., 2018; Manhardt et al., 2019; Zakharov et al., 2020; Beker et al., 2020) leverage DeepSDF (Park et al., 2019) to learn implicit signed distance functions or low-dimensional shape parameters from CAD models, and they further propose a render-and-compare approach to learn the parameters of 3D objects. Some works (Xiang et al., 2015, 2017) utilize voxel patterns to represent 3D objects. Other papers (Ku et al., 2019; Chen et al., 2021a) resort to point cloud reconstruction from images and estimate the locations of 3D objects with 2D-3D correspondences.

Geometric consistency Given the extrinsics matrix $T \in SE(3)$ that transforms a 3D coordinate in the object frame to the camera frame, and the camera intrinsics matrix K that

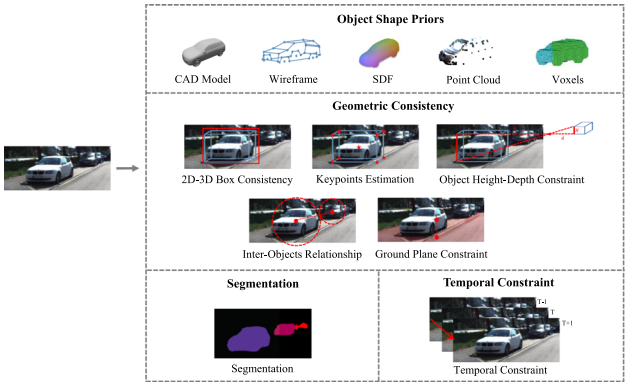

Fig. 13 An illustration of the prior types in monocular 3D object detection methods. Samples are from (Chen et al., 2016; He & Soatto, 2019; Xiang et al., 2015; Beker et al., 2020; Park et al., 2019)

project the 3D coordinate onto the image plane, the projection of a 3D point $[x, y, z]$ in the object frame into the image pixel coordinate $[u, v]$ can be represented as

$$d \cdot \begin{bmatrix} u \\ v \\ 1 \end{bmatrix} = K \; T \begin{bmatrix} x \\ y \\ z \\ 1 \end{bmatrix}, \tag{23}$$

where d is the depth of transformed 3D coordinate in the camera frame. Equation 23 provides a geometric relationship between 3D points and 2D image pixel coordinates, which can be leveraged in various ways to encourage consistency between the predicted 3D objects and the 2D objects on images. There are mainly 5 types of geometric constraints in monocular detection: 2D-3D boxes consistency, keypoints, object's height-depth relationship, inter-objects relationship, and ground plane constraints.

Some works (Mousavian et al., 2017; Brazil & Liu, 2019; Jörgensen et al., 2019; Naiden et al., 2019) propose to encourage the consistency between 2D and 3D boxes by minimizing reprojection errors. These methods introduce a post-processing step to optimize the 3D object parameters by gradually fitting the projected 3D boxes to 2D bounding boxes on images. There is also a branch of papers (Li et al., 2020c; Shi et al., 2020c; Li & Zhao, 2021) that predict the object keypoints from images, and the keypoints can be leveraged to calibrate the sizes of locations of 3D objects. Object's height-depth relationship can also serve as a strong geometric prior. Specifically, given the physical height of an object H in the 3D space, the visual height h on images, and the corresponding depth of the object d, there exists a geometric constraint: $d = f \cdot H/h$, where f is the camera focal length. This constraint can be leveraged to obtain more accurate depth estimation and has been broadly applied in a lot of works (Cai et al., 2020; Zhang et al., 2021c; Lu et al., 2021; Shi et al., 2021b). There are also some papers (Zhou et al., 2020b; Chen et al., 2020f) trying to model the inter-objects

Table 11 A taxonomy of prior-guided monocular detection methods based on prior types

Prior types		Methods
Object shape	Wireframe	(Zeeshan Zia et al., 2014; Chabot et al., 2017; He & Soatto, 2019)
	SDF	(Kundu et al., 2018; Manhardt et al., 2019; Beker et al., 2020; Zakharov et al., 2020)
	Points	(Chen et al., 2021a; Ku et al., 2019)
	Voxels	(Xiang et al., 2015, 2017)
Geometric consistency	2D-3D boxes	(Mousavian et al., 2017; Brazil & Liu, 2019; Jörgensen et al., 2019; Naiden et al., 2019)
	Keypoints	(Li et al., 2020c; Li & Zhao, 2021; Shi et al., 2020c)
	Height-depth	(Cai et al., 2020; Lu et al., 2021; Zhang et al., 2021c; Shi et al., 2021b)
	Inter-objects	(Zhou et al., 2020b; Chen et al., 2020f)
	Ground plane	(Chen et al., 2016; Liu et al., 2021b; Brazil & Liu, 2019; Liu et al., 2019a)
Temporal constraints		(Hu et al., 2019; Brazil et al., 2020)
Segmentation		(Xiang et al., 2015; Beker et al., 2020; Chen et al., 2016; Heylen et al., 2021)

relationships by exploiting new geometric relations among objects. Other papers (Chen et al., 2016; Brazil & Liu, 2019; Liu et al., 2019a, 2021b) leverage the assumption that 3D objects are generally on the ground plane to better localize those objects.

Temporal constraints Temporal association of 3D objects can be leveraged as strong prior knowledge. The temporal object relationships have been exploited as depth-ordering (Hu et al., 2019) and multi-frame object fusion with a 3D Kalman filter (Brazil et al., 2020).

Segmentation Image segmentation helps monocular 3D object detection mainly in two aspects. First, object segmentation masks are crucial for instance shape reconstruction in some works (Xiang et al., 2015; Beker et al., 2020). Second, segmentation indicates whether an image pixel is inside a 3D object from the perspective view, and this information has been utilized in Chen et al. (2016), Heylen et al. (2021) to help localize 3D objects.

Analysis: potentials and challenges of leveraging prior knowledge in monocular 3D detection With shape reconstruction, we could obtain more detailed object shape information from images, which is beneficial to 3D object detection. We can also attain more accurate detection results through the projection or render-and-compare loss. However, there exist two challenges for shape reconstruction applied in monocular 3D object detection. First, shape reconstruction normally requires an additional step of pre-training a reconstruction network, which hampers end-to-end training of the monocular detection pipeline. Second, object shapes are generally learned from CAD models instead of real-world instances, which imposes the challenge of generalizing the reconstructed objects to real-world scenarios.

Geometric consistencies are broadly adopted and can help improve detection accuracy. Nevertheless, some methods formulate the geometric consistency as an optimization problem and optimize object parameters in post-processing, which is quite time-consuming and hampers end-to-end training.

Image segmentation is useful information in monocular 3D detection. However, training segmentation networks requires expensive pixel annotations. Pre-training segmentation models on external datasets will suffer from the generalization problem.

4.2 Stereo-Based 3D Object Detection

Problem and Challenge Stereo-based 3D object detection aims to detect 3D objects from a pair of images. Compared to monocular images, paired stereo images provide additional geometric constraints that can be utilized to infer more accurate depth information. Hence, the stereo-based methods generally obtain a better detection performance than the monocular-based methods. Nevertheless, stereo cameras typically require very accurate calibration and synchronization, which are normally difficult to achieve in real applications. An illustration of stereo-based 3D object detection approaches is shown in Fig. 14 and a taxonomy is in Table 12.

Stereo matching and depth estimation A stereo camera can produce a pair of images, i.e. the left image $\mathcal{I}_L$ and the right image $\mathcal{I}_R$, in one shot. With the stereo matching techniques (Mayer et al., 2016; Chang & Chen, 2018), a disparity map can be estimated from the paired stereo images leveraging multi-view geometry (Hartley & Zisserman, 2003). Ideally, for each pixel on the left image $\mathcal{I}_L(u, v)$, there exists a pixel on the right image $\mathcal{I}_R(u, v + p)$ with the disparity value p so that the two pixels picture the same 3D location. Finally, the disparity map can be transformed into a depth image with the following formula:

$$d = \frac{f \times b}{p}, \tag{24}$$

where d is the depth value, f is the focal length, and b is the baseline length of the stereo camera. The pixel-wise disparity

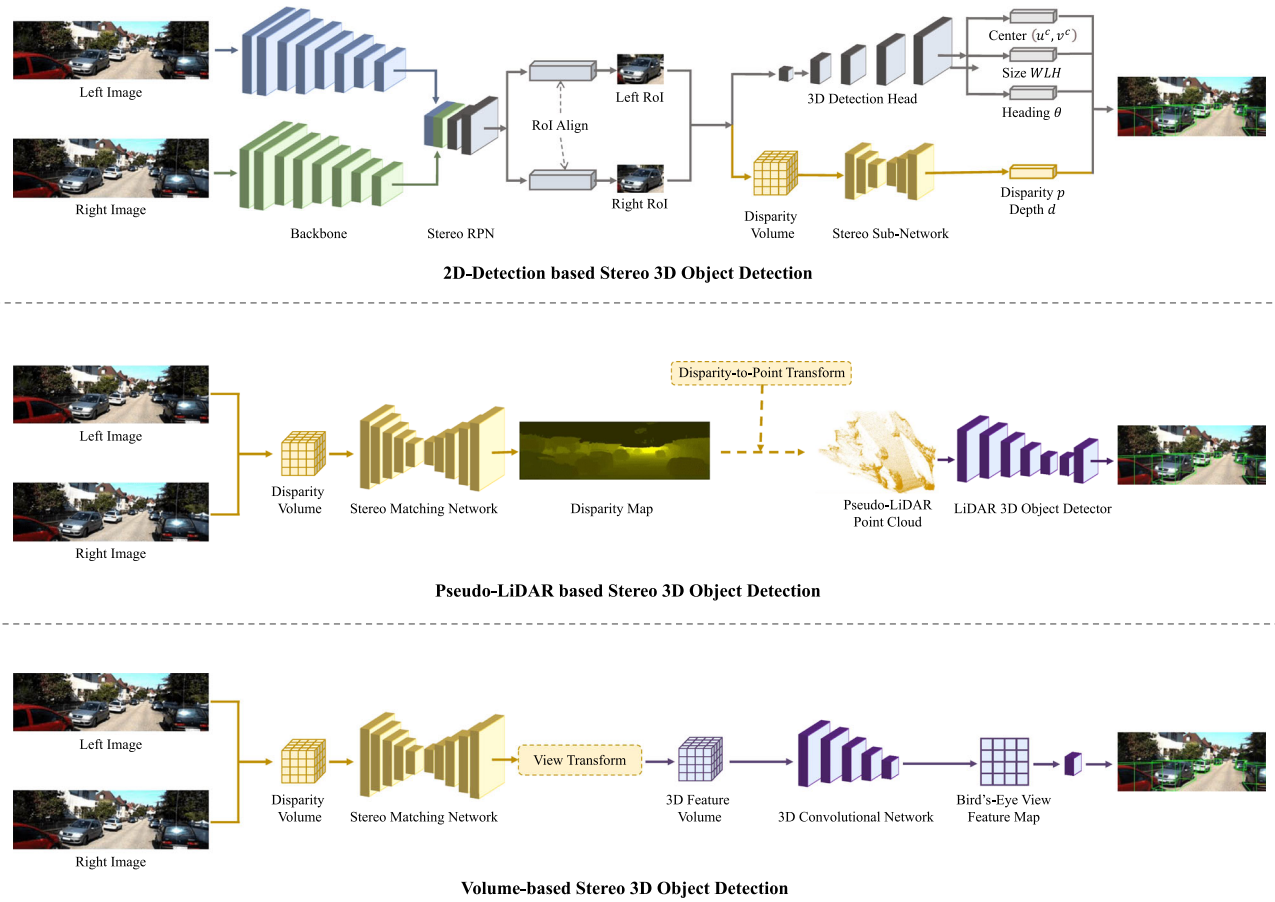

Fig. 14 An illustration of stereo-based 3D object detection methods. Image and disparity samples are from (Qian et al., 2020)

Table 12 A taxonomy of stereo-based detection methods based on auxiliary tasks and data representations

Method	2D det.	2D seg.	Disp./depth	Pseudo LiDAR	3D volume
3DOP Chen et al. (2015)			✓		
TLNet Qin et al. (2019b)	✓				
Stereo R-CNN Li et al. (2019b)	✓				
Disp R-CNN Sun et al. (2020b)	✓	✓	✓		
ZoomNet Xu et al. (2020)	✓	✓	✓		
OC-Stereo Pon et al. (2020)	✓	✓	✓		
IDA-3D Peng et al. (2020)	✓		✓		
YOLOStereo3D Liu et al. (2021a)			✓		
SIDE Peng et al. (2022)	✓		✓		
P-LiDAR++ You et al. (2020)			✓	✓	
Qian et al. (2020)			✓	✓	
CDN Garg et al. (2020)			✓	✓	
RT3D-Stereo Königshof et al. (2019)		✓	✓	✓	
CG-Stereo Li et al. (2020a)		✓	✓	✓	
LIGA-Stereo Guo et al. (2021)			✓		✓
DSGN Chen et al. (2020e)			✓		✓
PLUMENet Wang et al. (2021i)			✓		✓

constraints from stereo images enable more accurate depth estimation compared to monocular depth prediction.

2D-detection based methods Conventional 2D object detection frameworks can be modified to resolve the stereo detection problem. Specifically, paired stereo images are passed through an image-based detector with Siamese backbone networks to generate left and right regions of interest (RoIs) for the left and right images respectively. Then in the second stage, the left and right RoIs are fused to estimate the parameters of 3D objects. Stereo R-CNN (Li et al., 2019b) first proposes to extend 2D detection frameworks to stereo 3D detection. This design paradigm has been adopted in numerous papers. Qin et al. (2019b) proposes a novel stereo triangulation learning sub-network at the second stage; Xu et al. (2020), Pon et al. (2020), Sun et al. (2020b), Chen et al. (2021b) learn instance-level disparity by object-centric stereo matching and instance segmentation; Peng et al. (2020) proposes adaptive instance disparity estimation; Liu et al. (2021a), Peng et al. (2022) introduce single-stage stereo detection frameworks; Chen et al. (2015), Chen et al. (2017a) propose an energy-based framework for stereo-based 3D object detection.

Pseudo-LiDAR based methods The disparity map predicted from stereo images can be transformed into the depth image and then converted into the pseudo-LiDAR point cloud. Hence, similar to the monocular detection methods, the pseudo-LiDAR representation can also be employed in stereo-based 3D object detection methods. Those methods try to improve the disparity estimation in stereo matching for more accurate depth prediction. You et al. (2020) introduces a depth cost volume in stereo matching networks; Qian et al. (2020) proposes an end-to-end stereo matching and detection framework; Königshof et al. (2019), Li et al. (2020a) leverage semantic segmentation and predict disparity for foreground and background regions separately; Garg et al. (2020) proposes a Wasserstein loss for disparity estimation.

Volume-based methods There exists a category of methods that skip the pseudo-LiDAR representation and perform 3D object detection directly on 3D stereo volumes. DSGN Chen et al. (2020e) proposes a 3D geometric volume derived from stereo matching networks and applies a grid-based 3D detector on the volume to detect 3D objects. Guo et al. (2021) and Wang et al. (2021i) improve (Chen et al., 2020e) by leveraging knowledge distillation and 3D feature volumes respectively.

Potentials and challenges of the stereo-based methods Compared to the monocular detection methods, the stereo-based methods can obtain more accurate depth and disparity estimation with stereo matching techniques, which brings a stronger object localization ability and significantly boosts the 3D object detection performance. Nevertheless, an auxiliary stereo matching network brings additional time and memory consumption. Compared to LiDAR-based 3D object detection, detection from stereo images can serve as a much cheaper solution for 3D perception in autonomous driving scenarios. However, there still exists a non-negligible performance gap between the stereo-based and the LiDAR-based 3D object detection approaches.

4.3 Multi-view 3D Object Detection

Problem and Challenge Autonomous vehicles are generally equipped with multiple cameras to obtain complete environmental information from multiple viewpoints. Recently, multi-view 3D object detection has evolved rapidly. Some multi-view 3D detection approaches try to construct a unified BEV space by projecting multi-view images into the bird's-eye view, and then employ a BEV-based detector on top of the unified BEV feature map to detect 3D objects. The transformation from camera views to the bird's-eye view is ambiguous without accurate depth information, so image pixels and their BEV locations are not perfectly aligned. How to build reliable transformations from camera views to the bird's-eye view is a major challenge in these methods. Other methods resort to 3D object queries that are generated from the bird's-eye view and Transformers where cross-view attention is applied to object queries and multi-view image features. The major challenge is how to properly generate 3D object queries and design more effective attention mechanisms in Transformers.

BEV-based multi-view 3D object detection LSS Philion and Fidler (2020) is a pioneering work that proposes a lift-splat-shoot paradigm to solve the problem of BEV perception from multi-view cameras. There are three steps in LSS. *Lift:* bin-based depth prediction is conducted on image pixels and multi-view image features are lifted to 3D frustums with depth bins. *Splat:* 3D frustums are splatted into a unified bird's-eye view plane and image features are transformed into BEV features in an end-to-end manner. *Shoot:* downstream perception tasks are performed on top of the BEV feature map. This paradigm has been successfully adopted by many following works. BEVDet (Huang et al., 2021; Huang & Huang, 2022) improves LSS (Philion & Fidler, 2020) with a four-step multi-view detection pipeline, where the image view encoder encodes features from multi-view images, the view transformer transforms image features from camera views to the bird's-eye view, the BEV encoder further encodes the BEV features, and the detection head is employed on top of the BEV features for 3D detection. The major bottleneck in Huang et al. (2021), Philion and Fidler (2020) is depth prediction, as it is normally inaccurate and will result in inaccurate feature transforms from camera views to the bird's-eye view. To obtain more accurate depth information, many papers resort to mining additional information from multi-view images and past frames, e.g. (Li et al., 2022c) leverages explicit depth supervision, Wang et

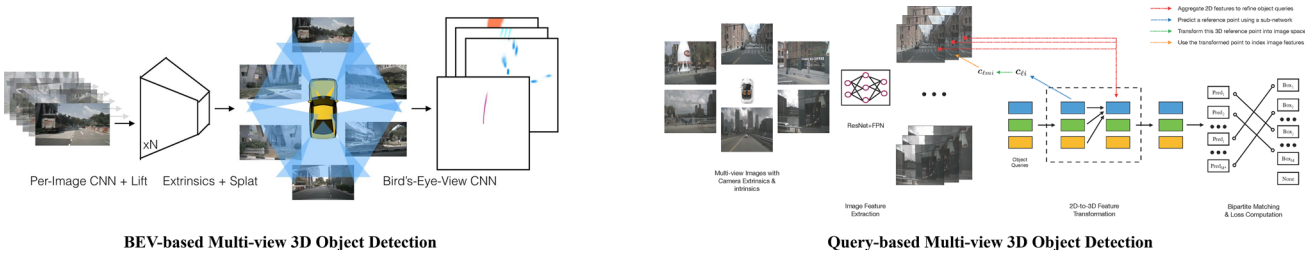

Fig. 15 An illustration of multi-view 3D object detection methods. Figures are from Philion and Fidler (2020) and Wang et al. (2022b)

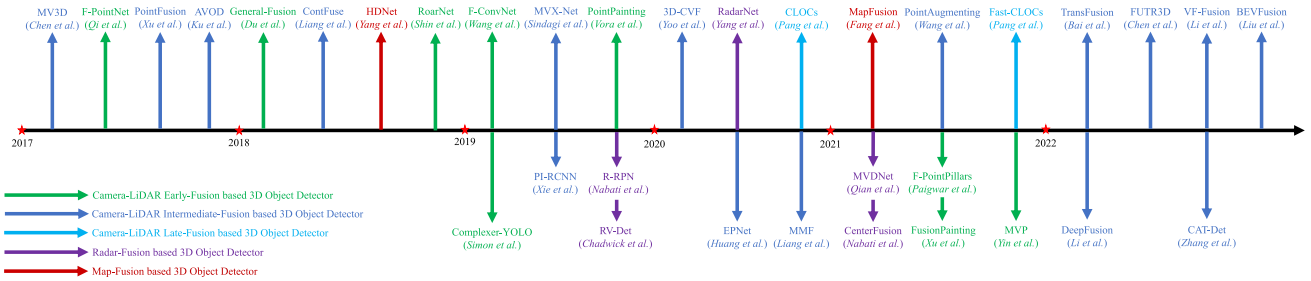

Fig. 16 Chronological overview of the multi-modal 3D object detection methods

al. (2022c) introduces surround-view temporal stereo, Li et al. (2022a) uses dynamic temporal stereo, Park et al. (2022) combines both short-term and long-term temporal stereo for depth prediction. In addition, there are also some papers (Xie et al., 2022; Huang et al., 2022a) that completely abandon the design of depth bins and categorical depth prediction. They simply assume that the depth distribution along the ray is uniform, so the camera-to-BEV transformation can be conducted with higher efficiency.

Query-based multi-view 3D object detection In addition to the BEV-based approaches, there is also a category of methods where object queries are generated from the bird's-eye view and interact with camera view features. Inspired by the advances in Transformers for object detection (Carion et al., 2020), DETR3D (Wang et al., 2022b) introduces a sparse set of 3D object queries, and each query corresponds to a 3D reference point. The 3D reference points can collect image features by projecting their 3D locations onto the multi-view image planes and then object queries interact with image features through Transformer layers. Finally, each object query will decode a 3D bounding box. Many following papers try to improve this design paradigm, such as introducing spatially-aware cross-view attention (Doll et al., 2022) and adding 3D positional embeddings on top of image features (Liu et al., 2022a). BEVFormer Li et al. (2022f) introduces dense grid-based BEV queries and each query corresponds to a pillar that contains a set of 3D reference points. Spatial cross-attention is applied to object queries and sparse image features to obtain spatial information, and temporal self-attention is applied to object queries and past BEV queries to fuse temporal information.

5 Multi-modal 3D Object Detection

In this section, we introduce the multi-modal 3D object detection approaches that fuse multiple sensory inputs. According to the sensor types, the approaches can be divided into three categories: LiDAR-camera, radar, and map fusion-based methods. In Sect. 5.1, we review and analyze the multi-modal detection approaches with LiDAR-camera fusion, including the early-fusion based, the intermediate-fusion based, and the late-fusion based methods. In Sect. 5.2, we investigate the multi-modal detection approaches with radar signals. In Sect. 5.3, we introduce the multi-modal 3D detection approaches with high-definition maps. A chronological overview of the multi-modal 3D object detection approaches is shown in Fig. 16.

5.1 Multi-modal Detection with LiDAR-Camera Fusion

Problem and Challenge Camera and LiDAR are two complementary sensor types for 3D object detection. Cameras provide color information from which rich semantic features can be extracted, while LiDAR sensors specialize in 3D localization and provide rich information about 3D structures. Many endeavors have been made to fuse the information from cameras and LiDARs for accurate 3D object detection. Since LiDAR-based detection methods perform much better than camera-based methods, the state-of-the-art approaches are mainly based on LiDAR-based 3D object detectors and try to incorporate image information into different stages of a LiDAR detection pipeline. In view of the complexity of

Table 13 A taxonomy of multi-sensor fusion-based detection methods based on fused stages, representations, and operators

Method	Fusion stage					Fusion representation		Fusion operator
	Input	Backbone	Proposal	RoI	Output	Camera rep	LiDAR rep	
F-PointNet Qi et al. (2018)	✓					Frustum	Point cloud	Region selection
F-ConvNet Wang and Jia (2019)	✓					Frustum	Point cloud	Region selection
RoarNet Shin et al. (2019)	✓					2D boxes & poses	Point cloud	Region selection
PointPainting Vora et al. (2020)	✓					2D segmentation	Point cloud	Point-wise append
MVX-Net Sindagi et al. (2019)		✓				image features	voxels	concatenation & MLP
ContFuse Liang et al. (2018)		✓				image features	BEV features	continuous convolution
PointFusion Xu et al. (2018)		✓				image features	point features	concatenation & MLP
EPNet Huang et al. (2020b)		✓				image features	point features	point-wise attention
MMF Liang et al. (2019)		✓	✓	✓		image features	BEV features	continuous convolution
3D-CVF Yoo et al. (2020)		✓	✓	✓		image features	BEV features	gated attention
MV3D Chen et al. (2017b)			✓	✓		image features	multi-view features	concatenation & MLP
AVOD Ku et al. (2018)			✓	✓		image features	BEV features	concatenation & MLP
CLOCs Pang et al. (2020)					✓	2D boxes	3D boxes	box consistency

LiDAR-based and camera-based detection systems, combining the two modalities together inevitably brings additional computational overhead and inference time latency. Therefore, how to efficiently fuse the multi-modal information remains an open challenge. A taxonomy of multi-modal 3D object detection methods is in Table 13.

5.1.1 Early-Fusion Based 3D Object Detection

Early-fusion based methods aim to incorporate the knowledge from images into point cloud before they are fed into a LiDAR-based detection pipeline. Hence the early-fusion frameworks are generally built in a sequential manner: 2D detection or segmentation networks are firstly employed to extract knowledge from images, and then the image knowledge is passed to point cloud, and finally the enhanced point cloud is fed to a LiDAR-based 3D object detector. Based on the fusion types, the early-fusion methods can be divided into two categories: region-level knowledge fusion and point-level knowledge fusion. An illustration of the early-fusion based approaches is shown in Fig. 17.

Region-level knowledge fusion Region-level fusion methods aim to leverage knowledge from images to narrow down the object candidate regions in 3D point cloud. Specifically, an image is first passed through a 2D object detector to generate 2D bounding boxes, and then the 2D boxes are extruded into 3D viewing frustums. The 3D viewing frustums are applied on LiDAR point cloud to reduce the searching space. Finally, only the selected point cloud regions are fed into a LiDAR detector for 3D object detection. F-PointNet (Qi et al., 2018) first proposes this fusion mechanism, and many endeavors have been made to improve the fusion framework. Wang and Jia (2019) divides a viewing frustum into grid cells and applies a convolutional network on the grid cells for 3D detection; Shin et al. (2019) proposes a novel geometric agreement search; Paigwar et al. (2021) exploits the pillar representation; Du et al. (2018) introduces a model fitting algorithm to find the object point cloud inside each frustum.

Point-level knowledge fusion Point-level fusion methods aim to augment input point cloud with image features. The augmented point cloud is then fed into a LiDAR detector to attain a better detection result. PointPainting Vora et al. (2020) is a seminal work that leverages image-based semantic segmentation to augment point clouds. Specifically, an image is passed through a segmentation network to obtain pixel-wise semantic labels, and then the semantic labels are attached to the 3D points by point-to-pixel projection. Finally, the points with semantic labels are fed into a LiDAR-based 3D object detector. This design paradigm has been followed by a lot of papers (Xu et al., 2021c; Simon et al., 2019; Meyer et al., 2019a). Apart from semantic segmentation, there also exist some works trying to exploit other

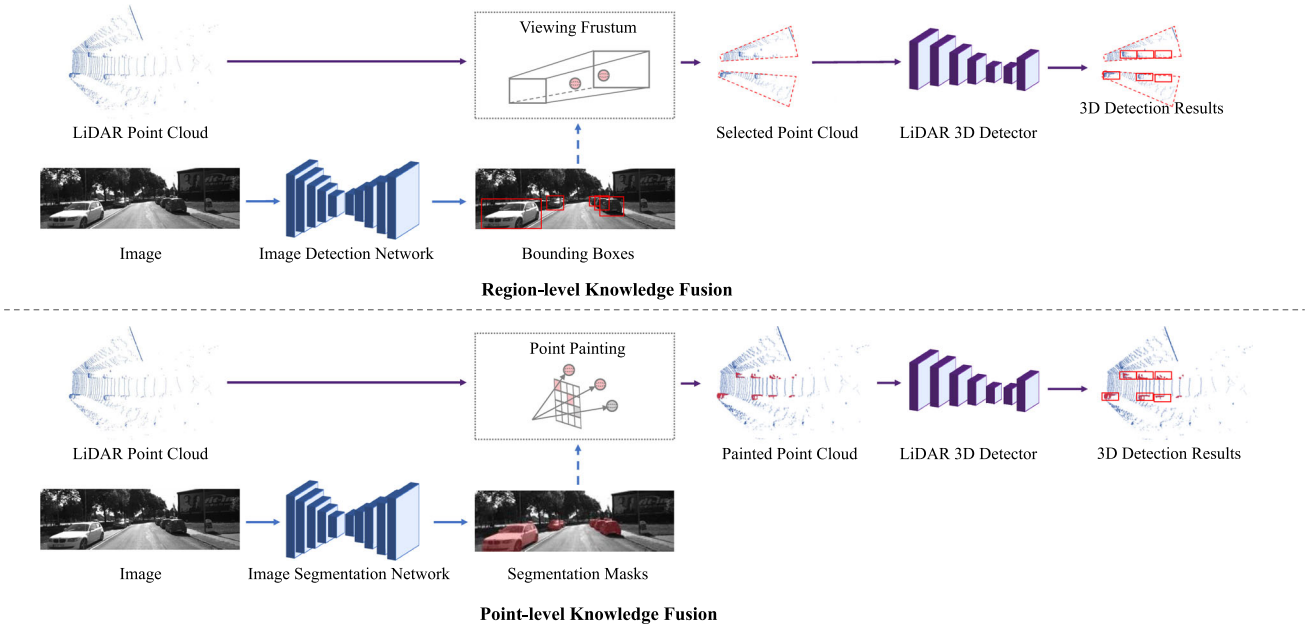

Fig. 17 An illustration of early-fusion based 3D object detection methods

information from images, e.g. depth image completion (Yin et al., 2021b).

Analysis: potentials and challenges of the early-fusion methods The early-fusion based methods focus on augmenting point clouds with image information before they are passed through a LiDAR 3D object detection pipeline. Most methods are compatible with a wide range of LiDAR-based 3D object detectors and can serve as a quite effective pre-processing step to boost detection performance. Nevertheless, the early-fusion methods generally perform multi-modal fusion and 3D object detection in a sequential manner, which brings additional inference latency. Given the fact that the fusion step generally requires a complicated 2D object detection or semantic segmentation network, the time cost brought by multi-modal fusion is normally non-negligible. Hence, how to perform multi-modal fusion efficiently at the early stage has become a critical challenge.

5.1.2 Intermediate-Fusion Based 3D Object Detection

Intermediate-fusion based methods try to fuse image and LiDAR features at the intermediate stages of a LiDAR-based 3D object detector, e.g. in backbone networks, at the proposal generation stage, or at the RoI refinement stage. These methods can also be classified according to the fusion stages. An illustration of intermediate-fusion based approaches is shown in Fig. 18.

Fusion in backbone networks Many endeavors have been made to progressively fuse image and LiDAR features in the backbone networks. In those methods, point-to-pixel correspondences are firstly established by LiDAR-to-camera transform, and then with the point-to-pixel correspondences, features from a LiDAR backbone can be fused with features from an image backbone through different fusion operators. The multi-modal fusion can be conducted in the intermediate layers of a grid-based detection backbone, with novel fusion operators such as continuous convolutions (Wang et al., 2018; Liang et al., 2018, 2019), hybrid voxel feature encoding (Sindagi et al., 2019), and Transformer (Li et al., 2022e; Zhang et al., 2022a). The multi-modal fusion can also be conducted only at the output feature maps of backbone networks, with fusion modules and operators including gated attention (Yoo et al., 2020), unified object queries (Chen et al., 2022a), BEV pooling (Liu et al., 2022b), learnable alignments (Chen et al., 2022c), point-to-ray fusion (Li et al., 2022d), Transformer (Bai et al., 2022), and other techniques (Dou et al., 2019; Chen et al., 2020d; Wang et al., 2021a). In addition to the fusion in grid-based backbones, there also exist some papers incorporating image information into the point-based detection backbones (Xu et al., 2018; Huang et al., 2020b; Xie et al., 2020a; Wang et al., 2021k; Zhu et al., 2021a).

Fusion in proposal generation and RoI head There exists a category of works that conduct multi-modal feature fusion at the proposal generation and RoI refinement stage. In those methods, 3D object proposals are first generated from a LiDAR detector, and then the 3D proposals are projected into multiple views, i.e. the image view and bird's-eye view, to crop features from the image and LiDAR backbone respectively. Finally, the cropped image and LiDAR features are fused in an RoI head to predict parameters for each 3D object. MV3D Chen et al. (2017b) and AVOD Ku et al.

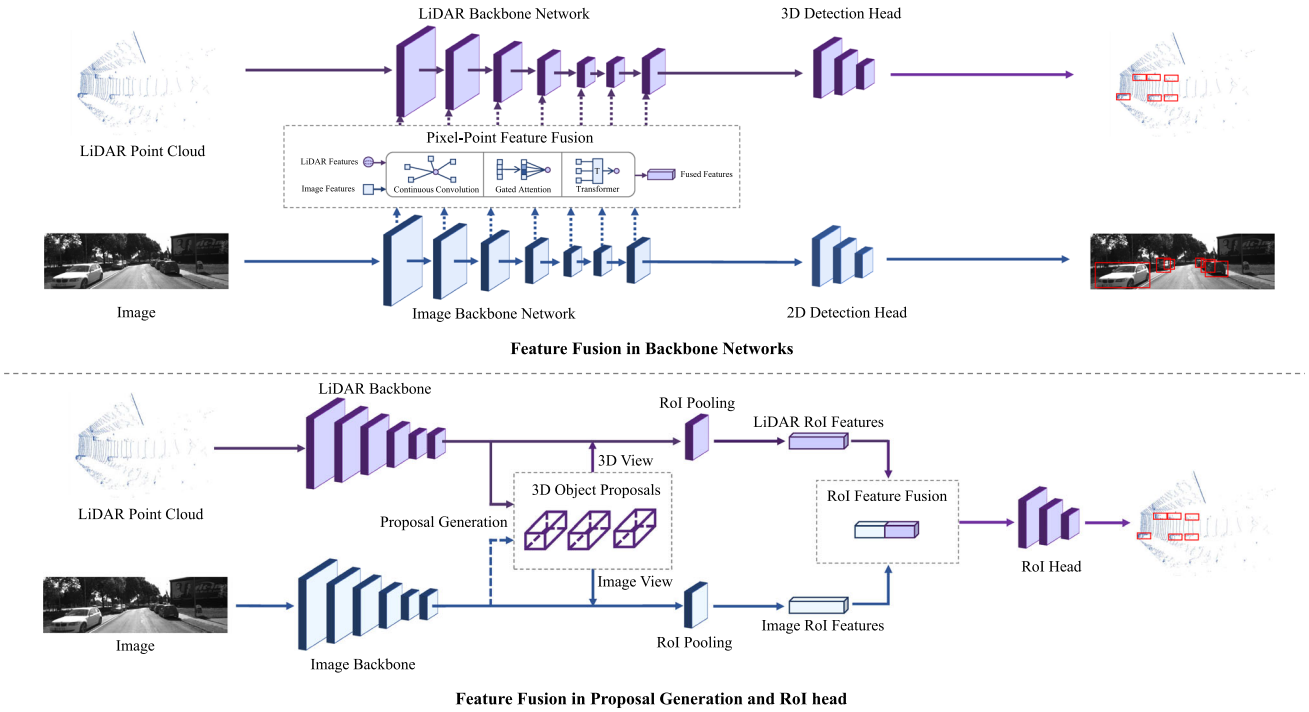

Fig. 18 An illustration of intermediate-fusion based 3D object detection methods

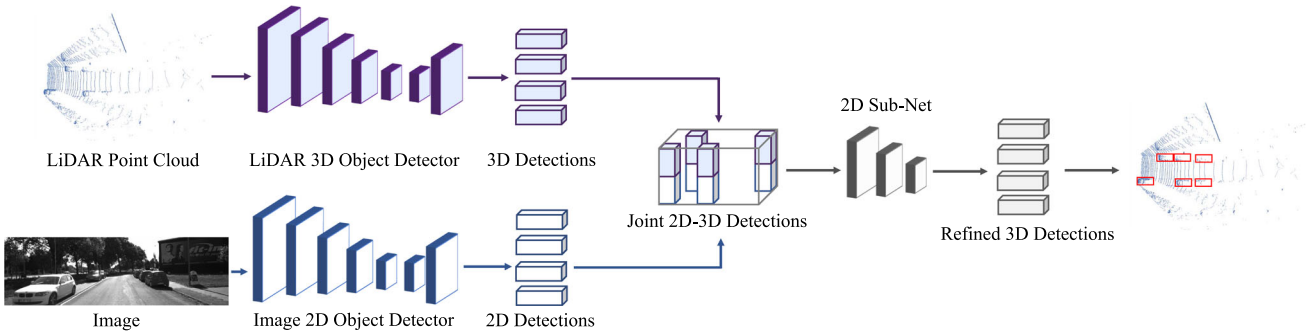

Fig. 19 An illustration of late-fusion based 3D object detection methods

(2018) are pioneering works leveraging multi-view aggregation for multi-modal detection. Other papers (Chen et al., 2022a; Bai et al., 2022) use the Transformer (Vaswani et al., 2017) decoder as the RoI head for multi-modal feature fusion.

Analysis: potentials and challenges of the intermediate-fusion methods The intermediate methods encourage deeper integration of multi-modal representations and yield 3D boxes of higher quality. Nevertheless, camera and LiDAR features are intrinsically heterogeneous and come from different viewpoints, so there still exist some problems on the fusion mechanisms and view alignments. Hence, how to fuse the heterogeneous data effectively and how to deal with the feature aggregation from multiple views remain a challenge to the research community.

5.1.3 Late-Fusion Based 3D Object Detection

Fusion at the box level Late-fusion based approaches operate on the outputs, i.e. 3D and 2D bounding boxes, from a LiDAR-based 3D object detector and an image-based 2D object detector respectively. An illustration of late-fusion based approaches is shown in Fig. 19. In those methods, object detection with camera and LiDAR sensor can be conducted in parallel, and the output 2D and 3D boxes are fused to yield more accurate 3D detection results. CLOCs Pang et al. (2020) introduces a sparse tensor that contains paired 2D-3D boxes and learns the final object confidence scores from this sparse tensor. Pang et al. (2022) improves (Pang et al., 2020) by introducing a light-weight 3D detector-cued image detector.

Analysis: potentials and challenges of the late-fusion methods The late-fusion based approaches focus on the instance-level aggregation and perform multi-modal fusion only on the outputs of different modalities, which avoids complicated interactions on the intermediate features or on the input point cloud. Hence these methods are much more efficient compared to other approaches. However, without resorting to deep features from camera and LiDAR sensors, these methods fail to integrate rich semantic information of different modalities, which limits the potential of this category of methods.

5.2 Multi-modal Detection with Radar Signals

Problem and Challenge Radar is an important sensory type in driving systems. In contrast to LiDAR sensors, radar has four irreplaceable advantages in real-world applications: Radar is much cheaper than LiDAR sensors; Radar is less vulnerable to extreme weather conditions; Radar has a larger detection range; Radar provides additional velocity measurements. Nevertheless, compared to LiDAR sensors that generate dense point clouds, radar only provides sparse and noisy measurements. Hence, how to effectively handle the radar signals remains a critical challenge.

Radar-LiDAR fusion Many papers try to fuse the two modalities by introducing new fusion mechanisms to enable message passing between the radar and LiDAR signals, including voxel-based fusion (Yang et al., 2020a), attention-based fusion (Qian et al., 2021a), introducing a range-azimuth-doppler tensor (Major et al., 2019), leveraging graph neural networks (Meyer et al., 2021), exploiting dynamic occupancy maps (Wang & Goldluecke, 2021), and introducing 4D radar data (Palffy et al., 2022).

Radar-camera fusion Radar-camera fusion is quite similar to LiDAR-camera fusion, as both radar and LiDAR data are 3D point representations. Most radar-camera approaches (Chadwick et al., 2019; Nabati & Qi, 2019, 2021) adapt the existing LiDAR-based detection architectures to handle sparse radar points and adopt similar fusion strategies as LiDAR-camera based methods.

5.3 Multi-modal Detection with High-Definition Maps

Problem and Challenge High-definition maps (HD maps) contain detailed road information such as road shape, road marking, traffic signs, barriers, etc. HD maps provide rich semantic information on surrounding environments and can be leveraged as a strong prior to assist 3D object detection. How to effectively incorporate map information into a 3D object detection framework has become an open challenge to the research community (Fig. 20).

Multi-modal detection with map information High-definition maps can be readily transformed into a bird's-eye view representation and fused with rasterized BEV point clouds or feature maps. The fusion can be conducted by simply concatenating the channels of a rasterized point cloud and an HD map from the bird's-eye view (Yang et al., 2018a), feeding LiDAR point cloud and HD map into separate backbones and fusing the output feature maps of the two modalities (Fang et al., 2021a), or simply filtering out those predictions that do not fall into the relevant map regions (Caesar et al., 2020). Other map types have also been explored, e.g. visibility map (Hu et al., 2020), vectorized map (Jiang et al., 2022).

6 Transformer-Based 3D Object Detection

In this section, we introduce the Transformer-based 3D object detection methods. Transformers (Vaswani et al., 2017) have shown prominent performance in many computer vision tasks, and many endeavors have been made to adapt Transformers to 3D object detection. In Sect. 6.1, we review the Transformers tailored for 3D object detection from an architectural perspective. In Sect. 6.2, we introduce the applications of Transformers in different 3D object detectors.

6.1 Transformer Architectures for 3D Object Detection

Problem and Challenge While most 3D object detectors are based on convolutional architectures, recently Transformer-based 3D detectors have shown great potential and dominated 3D object detection leaderboards. Compared to convolutional networks, the query-key-value design in Transformers enables more flexible interactions between different representations and the self-attention mechanism results in a larger receptive field than convolutions. However, fully-connected self-attention has quadratic time and space complexity *w.r.t.* the number of inputs, training Transformers can easily fall into sub-optimal results when the data size is small. Hence, it's critical to define proper query-key-value triplets and design specialized attention mechanisms for Transformer-based 3D object detectors.

Transformer architectures The development of Transformer architectures in 3D object detection has experienced three stages: (1) Inspired by vanilla Transformer (Vaswani et al., 2017), new Transformer modules with special attention mechanisms are proposed to obtain more powerful features in 3D object detection. (2) Inspired by DETR (Carion et al., 2020), query-based Transformer encoder-decoder designs are introduced to 3D object detectors. (3) Inspired by ViT (Dosovitskiy et al., 2021), patch-based inputs and architectures similar to Vision Transformers are introduced in 3D object detection.

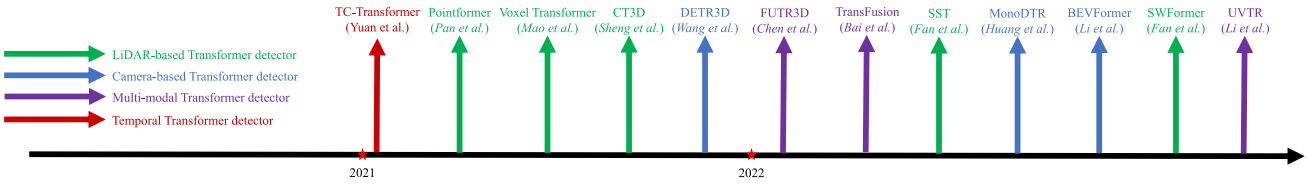

Fig. 20 Chronological overview of Transformer-based 3D object detectors

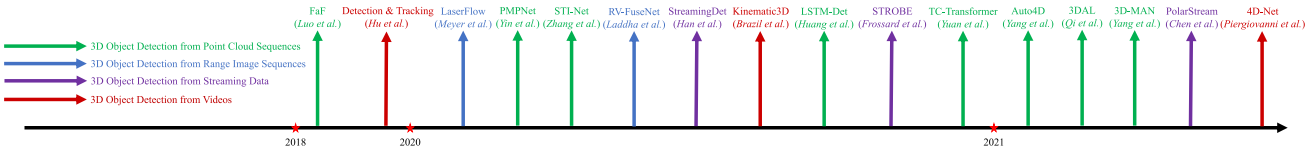

Fig. 21 Chronological overview of the temporal 3D object detection methods

In the first stage, many papers try to introduce novel Transformer modules into conventional 3D detection pipelines. In these papers, the choices of query, key, and value are quite flexible and new attention mechanisms are proposed. Pointformer (Pan et al., 2021) introduces Transformer modules to point backbones. It takes point features and coordinates as queries and applies self-attention to a group of point clouds. Voxel Transformer (Mao et al., 2021c) replaces convolutional voxel backbones with Transformer modules, where sparse and submanifold voxel attention are proposed and applied to voxels. CT3D Sheng et al. (2021) proposes a novel Transformer-based detection head, where proposal-to-point attention and channel-wise attention are introduced.

In the second stage, many papers propose DETR-like architectures for 3D object detection. They leverage a set of object queries and use those queries to interact with different features to predict 3D boxes. DETR3D Wang et al. (2022b) introduces object queries and generates a 3D reference point for each query. They use reference points to aggregate multi-view image features as keys and values, and apply cross-attention between object queries and image features. Finally, each query can decode a 3D bounding box for detection. Many following works have adopted the design of object queries and reference points. BEVFormer Li et al. (2022f) generates dense queries from BEV grids. TransFusion Bai et al. (2022) produces object queries from initial detections and applies cross-attention to LiDAR and image features in a Transformer decoder. UVTR Li et al. (2022b) fuses object queries with image and LiDAR voxels in a Transformer decoder. FUTR3D Chen et al. (2022a) fuses object queries with features from different sensors in a unified way.

In the third stage, many papers try to apply the designs of Vision Transformers to 3D object detectors. Following (Dosovitskiy et al., 2021; Liu et al., 2021c), they split inputs into patches and apply self-attention within each patch and across different patches. SST Fan et al. (2022) proposes a sparse Transformer, in which voxels in a local region are grouped into a patch and sparse regional attention is applied to the voxels in a patch, and then region shift is applied to change the grouping so new patches can be generated. SWFormer Sun et al. (2022) improves (Fan et al., 2022) with multi-scale feature fusion and voxel diffusion.

6.2 Transformer Applications in 3D Object Detection

Applications of Transformer-based 3D detectors Transformer architectures have been broadly adopted in various types of 3D object detectors. For point-based 3D object detectors, a point-based Transformer (Pan et al., 2021) has been developed to replace the conventional PointNet backbone. For voxel-based 3D detectors, a lot of papers (Mao et al., 2021c; Fan et al., 2022; Sun et al., 2022) propose novel voxel-based Transformers to replace the conventional convolutional backbone. For point-voxel based 3D object detectors, a new Transformer-based detection head (Sheng et al., 2021) has been proposed for better proposal refinement. For monocular 3D object detectors, Transformers can be used to fuse image and depth features (Huang et al., 2022b). For multi-view 3D object detectors, Transformers are utilized to fuse multi-view image features for each query (Li et al., 2022f; Wang et al., 2022b). For multi-modal 3D object detectors, many papers (Bai et al., 2022; Li et al., 2022b; Chen et al., 2022a) leverage Transformer architectures and special cross-attention mechanisms to fuse features of different modalities. For temporal 3D object detectors, Temporal-Channel Transformer (Yuan et al., 2021) is proposed to model temporal relationships across LiDAR frames.

7 Temporal 3D Object Detection

In this section, we introduce the temporal 3D object detection methods. Based on the data types, these methods can be divided into three categories: detection from LiDAR

Table 14 A taxonomy of temporal 3D object detection methods based on input representations

Input		Methods
LiDAR	multi-frame point clouds	(Huang et al., 2020a; Yang et al., 2021c; Zhang et al., 2020a; Yin et al., 2020; Yuan et al., 2021; Qi et al., 2021; Yang et al., 2021a)
	multi-frame range images	(Meyer et al., 2020; Laddha et al., 2020)
	streaming inputs	(Han et al., 2020; Frossard et al., 2021; Chen et al., 2021c)
Camera	videos	(Hu et al., 2019; Brazil et al., 2020; Piergiovanni et al., 2021)

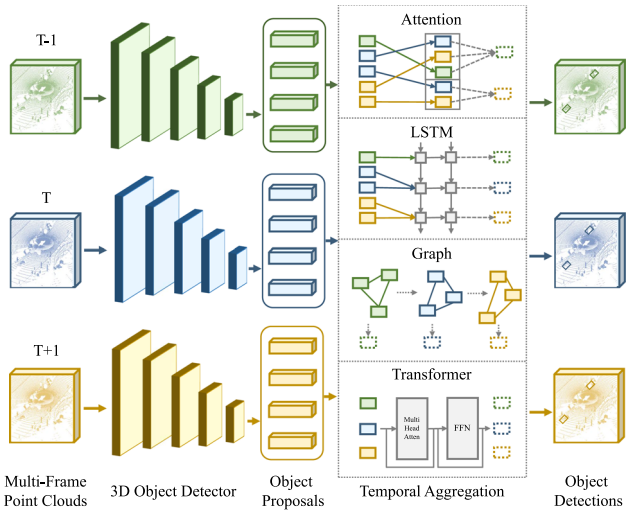

Fig. 22 An illustration of detection from LiDAR sequences

sequences, detection from streaming inputs, and detection from videos. In Sect. 7.1, we review the 3D object detection methods leveraging sequential LiDAR sweeps. In Sect. 7.2, we introduce the detection approaches with streaming data as input. In Sect. 7.3, we investigate 3D detection from videos and multi-modal temporal data. A chronological overview of the temporal detection approaches is shown in Fig. 21 and a taxonomy is in Table 14.

7.1 3D Object Detection from LiDAR Sequences

Problem and Challenge While most methods focus on detection from a single-frame point cloud, there also exist many approaches leveraging multi-frame point clouds for more accurate 3D object detection. These methods are trying to tackle the temporal detection problem by fusing multi-frame features via various temporal modeling tools, and they can also obtain more complete 3D shapes by merging multi-frame object points into a single frame. Temporal 3D object detection has exhibited great success in offline 3D auto-labeling pipelines. However, in onboard applications, these methods still suffer from memory and latency issues, as processing multiple frames inevitably brings additional time and memory costs, which can become severe when models are running on embedded devices. An illustration of temporal 3D object detection from LiDAR sequences is shown in Fig. 22.

3D object detection from sequential sweeps Most detection approaches using multi-frame point clouds resort to proposal-level temporal information aggregation. Namely, 3D object proposals are first generated independently from each frame of point cloud through a shared detector, and then various temporal modules are applied on the object proposals and the respective RoI features to aggregate the information of objects across different frames. The adopted temporal aggregation modules include temporal attention (Yang et al., 2021c), ConvGRU (Yin et al., 2020), graph network (Zhang et al., 2020a), LSTM (Huang et al., 2020a), and Transformer (Yuan et al., 2021). Temporal 3D object detection is also applied in the 3D object auto-labeling pipelines (Qi et al., 2021; Yang et al., 2021a). In addition to temporal detection from multi-frame point clouds, there are also some works (Meyer et al., 2020; Laddha et al., 2020) leveraging sequential range images for 3D object detection.

7.2 3D Object Detection from Streaming Data

Problem and Challenge Point clouds collected by rotating LiDARs are intrinsically a streaming data source in which LiDAR packets are sequentially recorded in a sweep. It typically takes 50-100 ms for a rotating LiDAR sensor to generate a 360° complete LiDAR sweep, which means that by the time a point cloud is produced, it no longer accurately reflects the scene at the exact time. This poses a challenge to autonomous driving applications which generally require minimal reaction times to guarantee driving safety. Many endeavors have been made to directly detect 3D objects from the streaming data. These methods generally detect 3D objects on the active LiDAR packets immediately without waiting for the full sweep to be built. Streaming 3D object detection is a more accurate and low-latency solution to vehicle perception compared to detection from full LiDAR sweeps. An illustration of 3D object detection from streaming data is shown in Fig. 23.

Streaming 3D object detection Similar to temporal detection from multi-frame point clouds, streaming detection methods (Han et al., 2020) can treat each LiDAR packet as an independent sample to detect 3D objects and apply temporal modules on the sequential packets to learn the inter-packets relationships. However, a LiDAR packet normally contains an incomplete point cloud and the information from a single packet is generally not sufficient for accurately detecting 3D

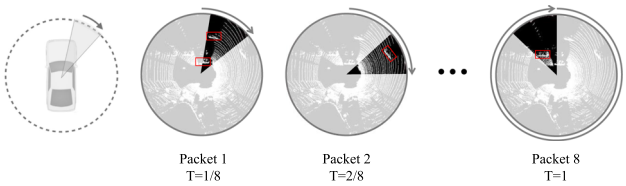

Fig. 23 An illustration of streaming 3D object detection. References Frossard et al. (2021) and Chen et al. (2021c)

objects. To this end, some papers try to provide more context information for detection in a single packet. The proposed techniques include a spatial memory bank (Frossard et al., 2021) and a multi-scale context padding scheme (Chen et al., 2021c).

7.3 3D Object Detection from Videos

Problem and Challenge Video is an important data type and can be easily obtained in autonomous driving applications. Compared to single-image based 3D object detection, video-based 3D detection naturally benefits from the temporal relationships of sequential images. While numerous works focus on single-image based 3D object detection, only a few papers investigate the problem of 3D object detection from videos, which leaves an open challenge to the research community.

Video-based 3D object detection Video-based detection approaches generally extend the image-based 3D object detectors by tracking and fusing the same objects across different frames. The proposed trackers include LSTM (Hu et al., 2019) and the 3D Kalman filter (Brazil et al., 2020). In addition, there are some works (Piergiovanni et al., 2021; Zeng et al., 2022) leveraging both videos and multi-frame point clouds for more accurate 3D object detection. Those methods propose 4D sensor-time fusion to learn features from both temporal and multi-modal data.

8 Label-Efficient 3D Object Detection

In this section, we introduce the methods of label-efficient 3D object detection. In previous sections, we generally assume the 3D detectors are trained under full supervision on a specific data domain and with a sufficient amount of annotations. However, in real-world applications, the 3D object detection methods inevitably face the problems of poor generalizability and lacking annotations. To address these issues, label-efficient techniques can be employed in 3D object detection, including domain adaptation (Sect. 8.1), weakly-supervised learning (Sect. 8.2), semi-supervised learning (Sect. 8.3), and self-supervised learning (Sect. 8.4) for 3D object detection. We will introduce those techniques in the following sections.

8.1 Domain Adaptation for 3D Object Detection

Problem and Challenge Domain gaps are ubiquitous in the data collection process. Different sensor settings and placements, different geographical locations, and different weathers will result in completely different data domains. In most conditions, 3D object detectors trained on a certain domain cannot perform well on other domains. Many techniques have been proposed to address the domain adaptation problem for 3D object detection, e.g. leveraging consistency between source and target domains, and self-training on target domains. Nevertheless, most methods only focus on solving one specific domain transfer problem. Designing a domain adaptation approach that can be generally applied in any domain transfer tasks in 3d object detection will be a promising research direction. An illustration of the domain gaps in 3D object detection is shown in Fig. 24 and a taxonomy of the domain adaptive methods is in Table 15

Cross-sensor domain adaptation Different datasets have different sensory settings, e.g. a 32-beam LiDAR sensor used in nuScenes (Caesar et al., 2020) versus a 64-beam LiDAR sensor in KITTI (Geiger et al., 2012), and the data is also collected at different geographic locations, e.g. KITTI (Geiger et al., 2012) is collected in Germany while Waymo (Sun et al., 2020c) is collected in United States. These factors will lead to severe domain gaps between different datasets, and the detectors trained on a dataset generally exhibit quite poor performance when they are tested on other datasets. Wang et al. (2020e) is a notable work that observes the domain gaps between datasets, and they introduce a statistic normalization approach to handle the gaps. Many following works leverage self-training to resolve the domain adaptation problem. In those methods, a detector pre-trained on the source dataset will produce pseudo labels for the target dataset, and then the detector is re-trained on the target dataset with pseudo labels. These methods make improvements mainly on obtaining pseudo labels of higher quality, e.g. (Saltori et al., 2020) proposes a scale-and-detect strategy, Yang et al. (2021b) introduces a memory bank, Fruhwirth-Reisinger et al. (2021) leverages the scene flow information, and You et al. (2021) exploits playbacks to enhance the quality of pseudo labels. In addition to the self-training approaches, there also exist some papers building alignments between source and target domains. The domain alignments can be established through a scale-aware and range-aware alignment strategy (Zhang et al., 2021a), multi-level consistency (Luo et al., 2021b), and a contrastive co-training scheme (Yihan et al., 2021).

In addition to the domain gaps among datasets, different sensors also produce data of distinct characteristics. A 32-beam LiDAR produces much sparser point clouds compared to a 64-beam LiDAR, and images obtained from different cameras also have diverse sizes and intrinsics. Rist et al. (2019) introduces a multi-task learning scheme to tackle the

Table 15 A taxonomy of domain adaptation methods for 3D object detection based on transferred domains and techniques

Method	Transferred Domain	Technique
Wang et al. (2020e)	cross-sensor	statistics normalization
SF-UDA^{3D} Saltori et al. (2020)	cross-sensor	self-training
ST3D Yang et al. (2021b)	cross-sensor	self-training
FAST3D Fruhwirth-Reisinger et al. (2021)	cross-sensor	self-training
SRDAN Zhang et al. (2021a)	cross-sensor	domain alignments
MLC-Net Luo et al. (2021b)	cross-sensor	domain alignments
3D-CoCo Yihan et al. (2021)	cross-sensor	domain alignments
Rist et al. (2019)	cross-sensor	multi-task learning
PIT Gu et al. (2021)	cross-sensor	image transform
SPG Xu et al. (2021b)	cross-weather	semantic point generation
Saleh et al. (2019)	sim-to-real	Cycle GAN
DeBortoli et al. (2021)	sim-to-real	adversarial training

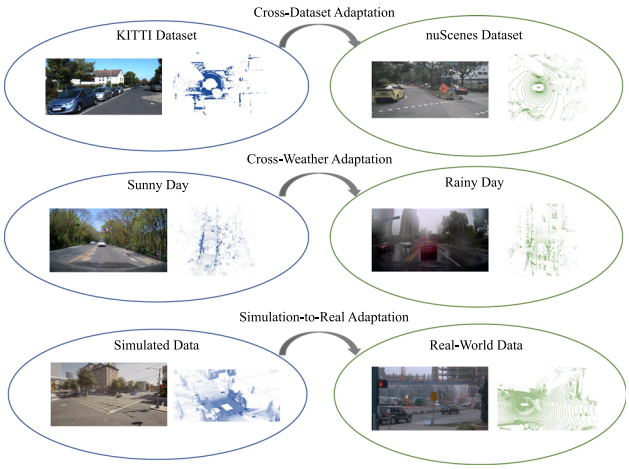

Fig. 24 An illustration of domain gaps in 3D detection

Fig. 25 An illustration of weakly-supervised 3D detection

domain gaps between different LiDAR sensors, and Gu et al. (2021) proposes the position-invariant transform to address the domain gaps between different cameras.

Cross-weather domain adaptation Weather conditions have a huge impact on the quality of collected data. On rainy days, raindrops will change the surface property of objects so that fewer LiDAR beams can be reflected and detected, so point clouds collected on rainy days are much sparser than those obtained under dry weather. Besides fewer reflections, rain also causes false positive reflections from raindrops in mid-air. Xu et al. (2021b) addresses the cross-weather domain adaptation problem with a novel semantic point generation scheme.

Sim-to-real domain adaptation Simulated data has been broadly adopted in 3D object detection, as the collected real-world data cannot cover all driving scenarios. However, the synthetic data has quite different characteristics from the real-world data, which gives rise to a sim-to-real adaptation problem. Many approaches are proposed to resolve this problem, including GAN (Zhu et al., 2017) based training (Saleh et al., 2019) and introducing an adversarial discriminator (DeBortoli et al., 2021) to distinguish real and synthetic data.

8.2 Weakly-Supervised 3D Object Detection

Problem and Challenge Existing 3D object detection methods highly rely on training with vast amounts of manually labeled 3D bounding boxes, but annotating those 3D boxes is quite laborious and expensive. Weakly-supervised learning can be a promising solution to this problem, in which weak supervisory signals, e.g. less expensive 2D annotations, are exploited to train the 3D object detection models. Weakly-supervised 3D object detection requires fewer human efforts for data annotation, but there still exists a non-negligible performance gap between the weakly-supervised and the fully-supervised methods. An illustration of weakly-supervised 3D object detection is shown in Fig. 25.

Weakly-supervised 3D object detection Weakly-supervised approaches leverage weak supervision instead of fully annotated 3D bounding boxes to train 3D object detectors. The weak supervisions include 2D image bounding boxes (Wei et al., 2021b; Peng et al., 2021), a pre-trained image detector (Qin et al., 2020), BEV object centers and vehicle instances (Meng et al., 2020, 2021). Those methods generally design novel learning mechanisms to skip the 3D box

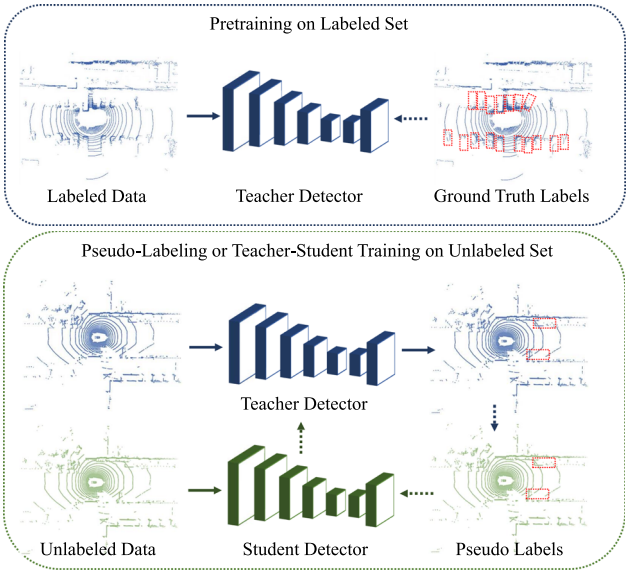

Fig. 26 An illustration of semi-supervised 3D detection

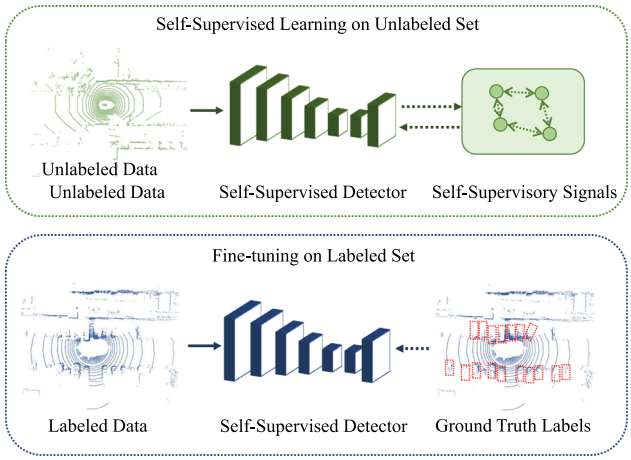

Fig. 27 An illustration of self-supervised 3D detection

supervision and learn to detect 3D objects by mining useful information from weak signals.

8.3 Semi-supervised 3D Object Detection

Problem and Challenge In real-world applications, data annotation requires much more human effort than data collection. Typically a data acquisition vehicle can collect more than 100k frames of point clouds in a day, while a skilled human annotator can only annotate 100-1k frames per day. This will inevitably lead to a rapid accumulation of a large amount of unlabeled data. Hence how to mine useful information from large-scale unlabeled data has become a critical challenge to both the research community and the industry. Semi-supervised learning, which exploits a small amount of labeled data and a huge amount of unlabeled data to jointly train a stronger model, is a promising direction. Combining 3D object detection with semi-supervised learning can boost detection performance. An illustration of semi-supervised 3D object detection is shown in Fig. 26.

Semi-supervised 3D object detection There are mainly two categories of approaches in semi-supervised 3D object detection: pseudo-labeling and teacher-student learning. The pseudo labeling approaches (Caine et al., 2021; Wang et al., 2021b) first train a 3D object detector with the labeled data, and then use the 3D detector to produce pseudo labels for the unlabeled data. Finally, the 3D object detector is re-trained with the pseudo labels on the unlabeled domain. The teacher-student methods (Zheng et al., 2021c) adapt the Mean Teacher (Tarvainen & Valpola, 2017) training paradigm to 3D object detection. Specifically, a teacher detector is first trained on the labeled domain, and then the teacher detector guides the training of a student detector on the unlabeled domain by encouraging the output consistencies between the two detection models.

8.4 Self-supervised 3D Object Detection

Problem and Challenge Self-supervised pre-training has become a powerful tool when there exists a large amount of unlabeled data and limited labeled data. In self-supervised learning, models are first pre-trained on large-scale unlabeled data and then fine-tuned on the labeled set to obtain a better performance. In autonomous driving scenarios, self-supervised pre-training for 3D object detection has not been widely explored. Existing methods are trying to adapt the self-supervised methods, e.g. contrastive learning, to the 3D object detection problem, but the rich semantic information in multi-modal data has not been well exploited. How to effectively handle the raw point clouds and images to pre-train an effective 3D object detector remains an open challenge. An illustration of self-supervised 3D object detection is in Fig. 27.

Self-supervised 3D object detection Self-supervised methods generally apply the contrastive learning techniques (He et al., 2020b; Chen et al., 2020c) to 3D object detection. Specifically, an input point cloud is first transformed into two views with augmentations, and then contrastive learning is employed to encourage the feature consistencies of the same 3D locations across the two views. Finally, the 3D detector pre-trained with contrastive learning is further fine-tuned on the labeled set to attain better performance. PointContrast Xie et al. (2020b) first introduces the contrastive learning paradigm in 3D object detection, and the following papers improve this paradigm by leveraging the depth information (Zhang et al., 2021d) and clustering (Liang et al., 2021a). In addition to self-supervised learning for point cloud detectors, there are also some works trying to

exploit both point clouds and images for self-supervised 3D detection, e.g. (Li et al., 2021f) proposes an intra-modal and inter-modal contrastive learning scheme on the multi-modal inputs.

9 3D Object Detection in Driving Systems

In this section, we introduce some critical problems of 3D object detection in driving systems. In Sect. 9.1, we review and analyze the approaches in which 3D object detection is trained together with other tasks, e.g. tracking, trajectory prediction, motion planning, localization, in an end-to-end manner. In Sect. 9.2, we introduce the simulation systems designed for 3D object detection and autonomous driving. In Sect. 9.3, we investigate the research topics on the robustness of 3D object detectors and safety-aware 3D object detection. In Sect. 9.4, we review the approaches of collaborative 3D object detection.

9.1 End-to-End Learning for Autonomous Driving

Problem and Challenge 3D object detection is a critical component of perception systems, and the performance of 3D object detectors will have a profound influence on downstream tasks like tracking, prediction, and planning. Hence from the systematic perspective, jointly training 3D object detection models with other perception tasks as well as the downstream tasks will be a better solution to autonomous driving. An open challenge is how to involve all driving tasks in a unified framework and jointly train these tasks in an end-to-end manner. An illustration of end-to-end autonomous driving is shown in Fig. 28.

Joint perception and prediction There are many works learning to perceive and track 3D objects and then predict their future trajectories in an end-to-end manner. FaF Luo et al. (2018) is a seminal work that proposes to jointly reason about 3D object detection, tracking, and trajectory prediction with a single 3D convolutional network. This design paradigm is followed by a lot of papers with improvements, e.g. (Casas et al., 2018) leverages the map information, Li et al. (2020b) introduces an interactive Transformer, Zhang et al. (2020b) designs a spatial-temporal-interactive network, Wu et al. (2020b) proposes a spatio-temporal pyramid network, Liang et al. (2020a) conducts all the tasks in a loop, Phillips et al. (2021) involves the localization task into the system.

Joint perception, prediction, and planning Many endeavors have been made to involve perception, prediction, and planning in a unified framework. Compared to the joint perception and prediction approaches, the whole system can benefit from the planner's feedback by adding motion planning to the end-to-end pipeline. Many techniques have been proposed to improve this framework, e.g. (Sadat et al., 2020) introduces a semantic occupancy map to produce interpretable intermediate representations, Wei et al. (2021a) incorporates spatial attention into the framework, Zeng et al. (2020) proposes a deep structured network, Casas et al. (2021) proposes a map-free approach, Cui et al. (2021) produces a diverse set of future trajectories.

End-to-end learning for autonomous driving Some methods try to build a completely end-to-end autonomous driving system, in which an autonomous vehicle takes sensory inputs and sequentially performs perception, prediction, planning, and motion control in a loop, and finally produces steering and speed signals for driving. Bojarski et al. (2016) first introduces the idea and implements the image-based end-to-end driving system with a convolutional neural network. Xiao et al. (2020) proposes an end-to-end architecture with multi-modal inputs. Codevilla et al. (2018) and Kendall et al. (2019) propose to learn end-to-end driving systems with conditional imitation learning and deep reinforcement learning respectively.

9.2 Simulation for 3D Object Detection

Problem and Challenge 3D object detection models generally require a large amount of data for training. While the data can be collected in real-world scenarios, the real-world data generally suffers from a long-tail distribution. For example, the scenarios of traffic accidents or extreme weather are seldom recorded but are quite important for training a robust 3D object detector. Simulation is a promising solution to address the long tail data distribution problem, as we can create synthetic data for those rare but critical scenarios. An open challenge for simulation is how to create more realistic synthetic data.

Visual simulation Many endeavors have been made to generate photo-realistic synthetic images in driving scenarios. The ideas of those methods include leveraging a graphics engine (Abu Alhaija et al., 2018; Ros et al., 2016), exploiting texture-mapped surfels (Yang et al., 2020b), leveraging real-world data (Chen et al., 2021d), and learning a controllable neural simulator (Kim et al., 2021).

LiDAR simulation In addition to generating synthetic images, many approaches try to generate LiDAR point clouds by simulation. Some methods (Fang et al., 2020; Nakashima & Kurazume, 2021; Fang et al., 2021b) propose novel point cloud rendering mechanisms by simulating the real-world effects. Some approaches (Manivasagam et al., 2020) leverage real-world instances to reconstruct 3D scenes. Other papers focus on simulation for safety-critical scenarios (Wang et al., 2021c) or under adverse weather conditions (Hahner et al., 2021).

Driving simulation Many papers try to build an interactive driving simulation platform where a virtual vehicle can per-

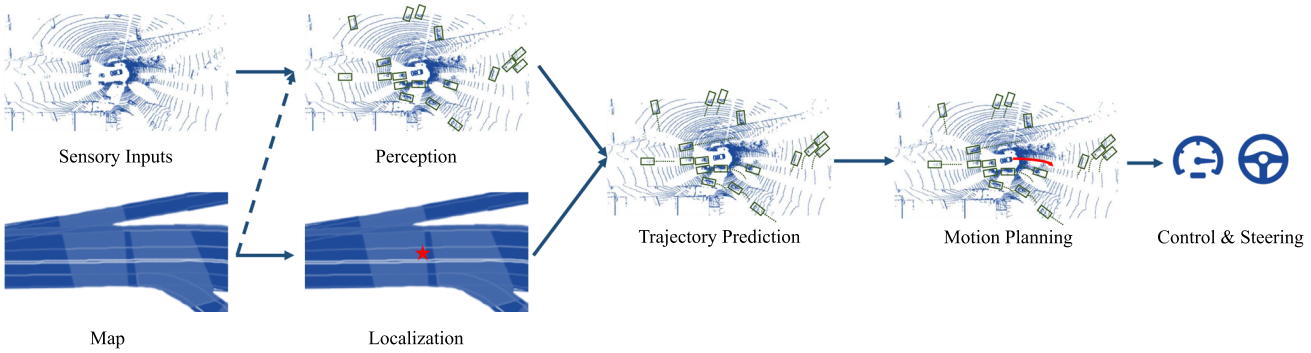

Fig. 28 An illustration of the autonomous driving pipeline. Point cloud and map samples are from (Casas et al., 2018)

ceive and interact with the virtual environments and finally plan the maneuvers. CARLA Dosovitskiy et al. (2017) is a pioneering open-source simulator for autonomous driving. Other papers utilize a graphics engine (Shah et al., 2018), leverage real-world data (Caesar et al., 2021), or develop a data-driven method (Amini et al., 2020) for driving simulation. There are also some works simulating the traffic flows (Tan et al., 2021; Suo et al., 2021) or testing the safety of vehicles by simulation (Wong et al., 2020).

9.3 Robustness for 3D Object Detection

Problem and Challenge Learning-based 3D object detectors are generally vulnerable to adversarial attacks. Adding perturbations or objects to the sensory inputs in an adversarial manner can fool the perception models and lead to misdetections. An open challenge of robust 3D object detection is to develop practical adversarial attack and defense algorithms that can be easy to implement and can be applied to most detection models.

Adversarial attacks on the LiDAR sensors Many endeavors have been made to attack the LiDAR sensors and fool the LiDAR-based perception models with adversarial machine learning. Cao et al. (2019) attack the LiDAR sensor and spoof obstacles close to the front of a victim autonomous vehicle. To achieve this goal, they introduce a novel algorithm to strategically control the spoofed attack to fool the LiDAR-based 3D object detection model. Wicker et al. Wicker and Kwiatkowska (2019) study the problem of adversarial attacks on the point-based detection models. They propose an iterative saliency occlusion approach to generate adversarial point cloud examples by dropping critical points. Tu et al. (2020) propose a method to generate physically realizable adversarial examples that can be placed on a vehicle and make this vehicle invisible to the LiDAR-based 3D object detectors. Sun et al. (2020a) study the general vulnerability of current LiDAR-based 3D object detection models and identify the ignored occlusion patterns in LiDAR point clouds that make vehicles vulnerable to spoofing attacks. They further propose a black-box spoofing attack method that can fool all target detection models. Zhu et al. (2021b) propose to use arbitrary objects to attack LiDAR-based 3D object detection models. Towards this goal, they introduce a method to identify the adversarial locations in a 3D scene, so that arbitrary objects placed at these locations can fool the LiDAR perception systems. Li et al. (2021e) exploit the fact that LiDAR point clouds collected from a moving vehicle need calibration based on the moving trajectories, so they propose to spoof the vehicle's trajectory with adversarial perturbations, which can distort the LiDAR sweeps and fool the 3D object detectors. Tu et al. (2021) perform adversarial attacks on the LiDAR perception models under the setting of multi-agent collaborative perception. Specifically, they fool the perception model of an agent by sending an adversarial message from the attacker in the multi-agent communication system.

Adversarial attacks on the multi-modal sensory inputs In addition to attacking the LiDAR-based perception models, there exist some works trying to perform adversarial attacks on both cameras and LiDAR sensors simultaneously. Cao et al. (2021) propose to generate a physically-realizable and adversarial 3D object that is invisible to both the camera and LiDAR sensor. The adversarial object is generated through optimization and can be leveraged to attack the multi-sensor fusion-based 3D object detection models. Tu et al. (2022) perform adversarial attacks on multi-modal perception models by introducing an adversarial textured mesh that can be placed on a vehicle and make this vehicle invisible to the multi-modal perception models. Specifically, the adversarial mesh is first rendered into both LiDAR points and image pixels in a differentiable manner, and then the multi-modal inputs are passed through a fusion-based detector. Finally, an adversarial loss is employed to adjust the mesh parameters.

9.4 Collaborative 3D Object Detection

Problem and Challenge Existing 3D detection approaches are mainly based on a single ego-vehicle. However, detecting 3D objects with a single vehicle inevitably meets two chal-

lenges: occlusion and sparsity of the far-away objects. To this end, some papers resort to detection under the multi-agent collaborative setting, where an ego-vehicle can communicate with other agents, e.g. vehicles or infrastructures, and exploit the information from other agents to improve the perception accuracy. A challenge of collaborative perception is how to properly balance the accuracy improvements and the communication bandwidth requirements.

Collaborative 3D object detection Collaborative detection approaches fuse the information from multiple agents to boost the performance of a 3D object detector. The fused information can be raw sensory inputs from other agents (Chen et al., 2019b; Zhang et al., 2021b), which cost little communication bandwidth and is quite efficient for detection, and it can also be compressed feature maps (Chen et al., 2019a; Wang et al., 2020c; Vadivelu et al., 2021; Li et al., 2021d), which cost non-negligible communication bandwidth but generally lead to better detection performance. There are also some papers studying when to communicate with other agents (Liu et al., 2020a) and which agent to communicate (Liu et al., 2020b).

10 Analysis and Outlooks

In this section, we conduct a systematic comparison and analysis of the 3D object detection approaches and prospect the future research directions of 3D object detection for autonomous driving. In Sect. 10.1, we conduct a comprehensive analysis of the detection performances and the inference speeds of various 3D object detection methods, i.e. LiDAR-based, camera-based, multi-modal approaches, on multiple datasets, from which we further summarize the research trends over the years. In Sect. 10.2, we propose future research directions in this area.

10.1 Research Trends

We comprehensively collect the statistics of various types of 3D object detection methods in recent years. The statistics include performances and inference time of the 3D object detectors on the most broadly-adopted KITTI (Geiger et al., 2012), nuScenes (Caesar et al., 2020), and Waymo (Sun et al., 2020c) dataset. Table 16, Table 17 and Table 18 show the statistical data. By analyzing these data, we obtain some intriguing findings on the research trends of 3D object detection.

10.1.1 Trends of Dataset Selection

Before 2018, most methods were evaluated on the KITTI dataset, and the evaluation metric they adopted is 2D average precision (AP_{2D}), where they project the 3D bounding boxes into the image plane and compare them with the ground truth 2D boxes. From 2018 until now, more and more papers have adopted the 3D or BEV average precision (AP_{3D} or AP_{BEV}), which is a more direct metric to measure 3D detection quality. For the LiDAR-based methods, the detection performances on KITTI quickly get converged over the years, e.g. AP_{3D} of easy cases increases from 71.40% (Yang et al., 2018c) to 90.90% (Shi et al., 2021a), and even AP_{3D} of hard cases reaches 79.14% (Mao et al., 2021c). Therefore, since 2019, more and more LiDAR-based approaches have turned to larger and more diverse datasets, such as the nuScenes dataset and the Waymo Open dataset. Large-scale datasets also provide more useful data types, e.g. raw range images provided by Waymo facilitate the development of range-based methods. For the camera-based detection methods, AP_{3D} of monocular detection on KITTI increases from 1.32% (Roddick et al., 2019) to 23.22% (Park et al., 2021), leaving huge room for improvement. Until now, only a few monocular methods have been evaluated on the Waymo dataset. For the multi-modal detection approaches, the methods before 2019 are mostly tested on the KITTI dataset, and after that most papers resort to the nuScenes dataset, as it provides more multi-modal data.

10.1.2 Trends of Inference Time

PointPillars Lang et al. (2019) has achieved remarkable inference speed with only 16ms latency, and its architecture has been adopted by many following works (Yin et al., 2021a; Wang et al., 2020f; Shi et al., 2022). However, even with the emergence of more powerful hardware, the inference speed didn't exhibit a significant improvement over the years. This is mainly because most methods focus on performance improvement and pay less attention to efficient inference. Many papers have introduced new modules into the existing detection pipelines, which also brings additional time costs. For the pseudo-LiDAR based detection methods, the stereo-based methods, and most multi-modal methods, the inference time is generally more than 100 ms, which cannot satisfy the real-time requirement and hampers the deployment in real-world applications.

10.1.3 Trends of the LiDAR-Based Methods

LiDAR-based 3D object detection has witnessed great advances in recent years. Among the LiDAR-based methods, the voxel-based and point-voxel based detection approaches attain superior performances, e.g. (Mao et al., 2021c) attains 82.09% moderate AP_{3D} and (Shi et al., 2020a) obtains 90.25% easy AP_{3D} on the KITTI dataset. The pillar-based detection methods are extremely fast, e.g. (Lang et al., 2019) runs at 60 Hz, but the detection accuracy is generally worse than the voxel-based methods. The range-based and BEV-

Table 16 A comprehensive performance analysis of various categories of 3D object detection methods across different datasets. We report the inference time (ms) originally reported in the papers, and report AP$|_{R_{40}}$ (%) for 3D car detection ($*$ denotes AP$|_{R_{40}}$ for BEV car detection) on the KITTI test benchmark, mAP (%) and NDS scores on the nuScenes test set, Level 1 (L1) mAP and Level 2 (L2) mAP on the Waymo validation set. We group the methods based on the sensor types and the input representations, and sort them by the year of publication

Method	Sensor	Representation	Year	Inference Time (ms)	KITTI Car			nuScenes		Waymo Vehicle	
					Easy	Mod.	Hard	mAP	NDS	L1	L2
IPOD Yang et al. (2018c)	LiDAR	Point	2018	–	71.40	53.46	48.34	–	–	–	–
StarNet Ngiam et al. (2019)	LiDAR	Point	2019	–	81.63	73.99	67.07	–	–	53.7	–
PointRCNN Shi et al. (2019)	LiDAR	Point	2019	100	85.94	75.76	68.32	–	–	–	–
STD Yang et al. (2019)	LiDAR	Point	2019	80	86.61	77.63	76.06	–	-	–	–
3DSSD Yang et al. (2020c)	LiDAR	Point	2020	38	88.36	79.57	74.55	42.6	56.4	–	–
Point-GNN Shi and Rajkumar (2020)	LiDAR	Point	2020	640	88.33	79.47	72.29	–	–	–	–
Pointformer Pan et al. (2021)	LiDAR	Point	2021	–	87.13	77.06	69.25	53.6	–	–	–
Vote3D Wang and Posner (2015)	LiDAR	Voxel	2015	500	–	–	–	–	–	–	–
Vote3Deep Engelcke et al. (2017)	LiDAR	Voxel	2017	1100	–	–	–	–	–	–	–
3D-FCN Li (2017)	LiDAR	Voxel	2017	–	–	–	–	–	–	–	–
VoxelNet Zhou and Tuzel (2018)	LiDAR	Voxel	2018	220	77.47	65.11	57.73	–	–	–	–
SECOND Yan et al. (2018)	LiDAR	Voxel	2018	50	83.13	73.66	66.20	–	–	–	–
CBGS Zhu et al. (2019)	LiDAR	Voxel	2019	–	–	–	–	52.8	63.3	–	–
HVNet Ye et al. (2020a)	LiDAR	Voxel	2020	30	–	–	–	–	–	–	–
DOPS Najibi et al. (2020)	LiDAR	Voxel	2020	–	–	–	–	–	–	56.4	–
MVF Zhou et al. (2020c)	LiDAR	Voxel	2020	–	–	–	–	–	–	62.93	–
AFDet Ge et al. (2020)	LiDAR	Voxel	2020	–	–	–	–	–	–	63.69	–
SSN Zhu et al. (2020)	LiDAR	Voxel	2020	–	–	–	–	46.3	56.9	–	–
CVC-Net Chen et al. (2020a)	LiDAR	Voxel	2020	–	–	–	–	55.8	64.2	65.2	–
Wang et al. (2020b)	LiDAR	Voxel	2020	50	–	–	–	48.5	59.0	–	–
SegVoxelNet Yi et al. (2020)	LiDAR	Voxel	2020	40	84.19	75.81	67.80	–	–	–	–
HotSpotNet Chen et al. (2020b)	LiDAR	Voxel	2020	40	87.60	78.31	73.34	59.3	66.0	–	–
Associate-3Ddet Du et al. (2020)	LiDAR	Voxel	2020	60	85.99	77.40	70.53	–	–	–	–
TANet Liu et al. (2020d)	LiDAR	Voxel	2020	–	83.81	75.38	67.66	–	–	–	–

Table 16 continued

Method	Sensor	Representation	Year	Inference Time (ms)	KITTI Car			nuScenes		Waymo Vehicle	
					Easy	Mod.	Hard	mAP	NDS	L1	L2
Part-A^2 Net Shi et al. (2020b)	LiDAR	Voxel	2020	80	85.94	77.86	72.00	–	–	–	–
CenterPoint Yin et al. (2021a)	LiDAR	Voxel	2021	70	–	–	–	58.0	65.5	76.7	68.8
Object DGCNN Wang and Solomon (2021)	LiDAR	Voxel	2021	–	–	–	–	58.7	66.1	–	-
CIA-SSD Zheng et al. (2021a)	LiDAR	Voxel	2021	30	89.59	80.28	72.87	–	–	–	–
Voxel R-CNN Deng et al. (2021b)	LiDAR	Voxel	2021	40	90.90	81.62	77.06	–	–	75.59	66.59
Voxel Transformer Mao et al. (2021c)	LiDAR	Voxel	2021	140	89.90	82.09	79.14	–	–	74.95	65.91
SST Fan et al. (2022)	LiDAR	Voxel	2022	–	–	–	–	–	–	74.2	65.5
SWFormer Sun et al. (2022)	LiDAR	Voxel	2022	–	–	–	–	–	–	77.8	69.2
PointPillars Lang et al. (2019)	LiDAR	Pillar	2019	16	79.05	74.99	68.30	40.1	55.0	56.62	–
Pillar-OD Wang et al. (2020f)	LiDAR	Pillar	2020	–	–	–	–	–	–	69.8	–
PillarNet Shi et al. (2022)	LiDAR	Pillar	2022	–	–	–	–	66.0	71.4	83.23	76.09
VeloFCN Li et al. (2016)	LiDAR	BEV Image	2016	1000	–	–	–	–	–	-	–
BirdNet Beltrán et al. (2018)	LiDAR	BEV Image	2018	–	75.52*	50.81*	50.00*	–	–	–	–
PIXOR Yang et al. (2018b)	LiDAR	BEV Image	2018	35	81.70*	77.05*	72.95*	–	–	–	–
HDNet Yang et al. (2018a)	LiDAR	BEV Image	2018	50	89.14*	86.57*	78.32*	–	–	–	–
PVCNN Liu et al. (2019b)	LiDAR	Point-Voxel	2019	59	–	–	–	–	–	–	–
Fast Point R-CNN Chen et al. (2019c)	LiDAR	Point-Voxel	2019	65	84.28	75.73	67.39	–	–	–	-
PV-RCNN Shi et al. (2020a)	LiDAR	Point-Voxel	2020	–	90.25	81.43	76.82	–	–	77.51	68.98
SA-SSD He et al. (2020a)	LiDAR	Point-Voxel	2020	40	88.75	79.79	74.16	–	–	–	–
SPVNAS Tang et al. (2020)	LiDAR	Point-Voxel	2020	–	87.8	78.4	74.8	–	–	–	–
InfoFocus Wang et al. (2020a)	LiDAR	Point-Voxel	2020	–	–	–	–	39.5	–	–	–
PVGNet Miao et al. (2021)	LiDAR	Point-Voxel	2021	–	89.94	81.81	77.09	–	–	74.0	–
HVPR Noh et al. (2021)	LiDAR	Point-Voxel	2021	28	86.38	77.92	73.04	–	–	–	–
PV-RCNN++ Shi et al. (2021a)	LiDAR	Point-Voxel	2021	–	90.14	81.88	77.15	–	–	79.25	70.61
CT3D Sheng et al. (2021)	LiDAR	Point-Voxel	2021	–	87.83	81.77	77.16	–	–	76.30	69.04
LiDAR R-CNN Li et al. (2021g)	LiDAR	Point-Voxel	2021	–	–	–	–	–	–	76.0	68.3

Table 16 continued

Method	Sensor	Representation	Year	Inference Time (ms)	KITTI Car			nuScenes		Waymo Vehicle	
					Easy	Mod.	Hard	mAP	NDS	L1	L2
Pyramid R-CNN Mao et al. (2021a)	LiDAR	Point-Voxel	2021	–	88.39	82.08	77.49	–	–	76.30	67.23
LaserNet Meyer et al. (2019b)	LiDAR	Range Image	2019	30	–	–	–	–	–	52.11	–
RCD Bewley et al. (2020)	LiDAR	Range Image	2020	–	–	–	–	–	–	69.59	–
RangeRCNN Liang et al. (2020b)	LiDAR	Range Image	2020	45	88.47	81.33	77.09	–	–	75.43	–
RangeIoUDet Liang et al. (2021c)	LiDAR	Range Image	2021	22	88.60	79.80	76.76	–	–	–	–
PPC Chai et al. (2021)	LiDAR	Range Image	2021	–	–	–	–	–	–	65.2	56.7
RangeDet Fan et al. (2021)	LiDAR	Range Image	2021	–	–	–	–	–	–	72.85	–
RSN Sun et al. (2021)	LiDAR	Range Image	2021	67.5	–	–	–	–	–	78.4	69.5
Mono3D Chen et al. (2016)	Camera	Monocular	2016	–	–	–	–	–	–	–	–
Deep3DBox Mousavian et al. (2017)	Camera	Monocular	2017	–	–	–	–	–	–	–	–
Deep MANTA Chabot et al. (2017)	Camera	Monocular	2017	–	–	–	–	–	–	–	-
SubCNN Xiang et al. (2017)	Camera	Monocular	2017	–	–	–	–	–	–	–	–
3D-RCNN Kundu et al. (2018)	Camera	Monocular	2018	–	–	–	–	–	–	–	–
MultiFusion Xu and Chen (2018)	Camera	Monocular	2018	–	7.08	5.18	4.68	–	–	–	–
Mono3D++ He and Soatto (2019)	Camera	Monocular	2019	–	–	–	–	–	–	–	–
DeepOptics Chang and Wetzstein (2019)	Camera	Monocular	2019	–	–	–	–	–	–	–	–
Weng *et al.* Weng and Kitani (2019)	Camera	Monocular	2019	–	–	–	–	–	–	–	–
CenterNet Zhou et al. (2019b)	Camera	Monocular	2019	–	–	–	–	33.8	40.0	–	–
OFT-Net Roddick et al. (2019)	Camera	Monocular	2019	500	1.32	1.61	1.00	–	–	–	–

Table 17 A comprehensive performance analysis of all branches of 3D object detection methods across different datasets. (Continued)

Method	Sensor	Representation	Year	Inference Time (ms)	KITTI Car			nuScenes		Waymo Vehicle	
					Easy	Mod	Hard	mAP	NDS	L1	L2
FQNet Liu et al. (2019a)	Camera	Monocular	2019	500	2.77	1.51	1.01	–	–	–	–
ROI-10D Manhardt et al. (2019)	Camera	Monocular	2019	200	4.32	2.02	1.46	–	–	–	–
GS3D Li et al. (2019a)	Camera	Monocular	2019	–	4.47	2.90	2.47	–	–	–	–
MonoFENet Bao et al. (2019)	Camera	Monocular	2019	–	8.35	5.14	4.10	–	–	–	–
MonoGRNet Qin et al. (2019a)	Camera	Monocular	2019	60	9.61	5.74	4.25	–	–	–	–
MonoDIS Simonelli et al. (2019)	Camera	Monocular	2019	100	10.37	7.94	6.40	30.4	38.4	–	–
MonoPSR Ku et al. (2019)	Camera	Monocular	2019	–	10.76	7.25	5.85	–	–	–	–
M3D-RPN Brazil and Liu (2019)	Camera	Monocular	2019	160	14.76	9.71	7.42	–	–	–	–
AM3D Ma et al. (2019a)	Camera	Monocular	2019	400	16.50	10.74	9.52	–	–	–	–
SDFLabel Zakharov et al. (2020)	Camera	Monocular	2020	–	–	–	–	–	–	–	–
MonoDR Beker et al. (2020)	Camera	Monocular	2020	–	–	–	–	–	–	–	–
Wang et al. (2020d)	Camera	Monocular	2020	–	–	–	–	–	–	–	–
MoNet3D Zhou et al. (2020b)	Camera	Monocular	2020	–	–	–	–	–	–	–	–
Cai et al. (2020)	Camera	Monocular	2020	–	11.08	7.02	5.63	–	–	–	–
MonoPair Chen et al. (2020f)	Camera	Monocular	2020	60	13.04	9.99	8.65	–	–	–	–
SMOKE Liu et al. (2020c)	Camera	Monocular	2020	30	14.03	9.76	7.84	–	–	–	–
RTM3D Li et al. (2020c)	Camera	Monocular	2020	–	14.41	10.34	8.77	–	–	–	–
MoVi-3D Simonelli et al. (2020)	Camera	Monocular	2020	–	15.19	10.90	9.26	–	–	–	–
UR3D Shi et al. (2020c)	Camera	Monocular	2020	–	15.58	8.61	6.00	–	–	–	–
D^4LCN Ding et al. (2020)	Camera	Monocular	2020	200	16.65	11.72	9.51	–	–	–	–
Ye et al. (2020b)	Camera	Monocular	2020	400	16.77	12.72	9.17	–	–	–	–
Kinematic3D Brazil et al. (2020)	Camera	Monocular	2020	120	19.07	12.72	9.17	–	–	–	–
CaDDN Reading et al. (2021)	Camera	Monocular	2021	–	19.17	13.41	11.46	–	–	–	–
PatchNet Ma et al. (2020)	Camera	Monocular	2021	400	15.68	11.12	10.17	–	–	0.39	0.38
MonoDLE Ma et al. (2021)	Camera	Monocular	2021	–	17.23	12.26	10.29	–	–	–	–
M3DSSD Luo et al. (2021a)	Camera	Monocular	2021	–	17.51	11.46	8.98	–	–	–	–
Kumar et al. (2021)	Camera	Monocular	2021	–	18.10	12.32	9.65	–	–	–	–

Table 17 continued

Method	Sensor	Representation	Year	Inference Time (ms)	KITTI Car			nuScenes		Waymo Vehicle	
					Easy	Mod	Hard	mAP	NDS	L1	L2
MonoRCNN Shi et al. (2021b)	Camera	Monocular	2021	70	18.36	12.65	10.03	–	–	–	–
MonoRUn Chen et al. (2021a)	Camera	Monocular	2021	–	19.65	12.30	10.58	–	–	–	–
DDMP Wang et al. (2021d)	Camera	Monocular	2021	–	19.71	12.78	9.80	–	–	–	–
MonoFlex Zhang et al. (2021c)	Camera	Monocular	2021	–	19.94	13.89	12.07	–	–	–	–
GUP Net Lu et al. (2021)	Camera	Monocular	2021	–	20.11	14.20	11.77	–	–	–	–
PCT Wang et al. (2021e)	Camera	Monocular	2021	–	21.00	13.37	11.31	–	–	0.89	0.66
MonoEF Zhou et al. (2021)	Camera	Monocular	2021	–	21.29	13.87	11.71	–	–	–	–
Liu et al. (2021b)	Camera	Monocular	2021	–	21.65	13.25	9.91	–	–	–	-
DD3D Park et al. (2021)	Camera	Monocular	2021	–	23.22	16.34	14.20	41.8	47.7	–	–
FCOS3D Wang et al. (2021g)	Camera	Monocular	2021	–	–	–	–	35.8	42.8	–	–
PGD Wang et al. (2022a)	Camera	Monocular	2022	28	–	–	–	38.6	44.8	–	–
MonoDTR Huang et al. (2022b)	Camera	Monocular	2022	–	21.99	15.39	12.73	–	–	–	–
3DOP Chen et al. (2015)	Camera	Stereo	2015	–	–	–	–	–	–	–	–
TLNet Qin et al. (2019b)	Camera	Stereo	2019	–	7.64	4.37	3.74	–	–	–	–
Stereo R-CNN Li et al. (2019b)	Camera	Stereo	2019	420	47.58	30.23	23.72	–	–	–	–
Pseudo-LiDAR Wang et al. (2019b)	Camera	Stereo	2019	–	54.53	34.05	28.25	–	–	–	–
IDA-3D Peng et al. (2020)	Camera	Stereo	2020	–	–	–	–	–	–	–	–
OC-Stereo Pon et al. (2020)	Camera	Stereo	2020	350	55.15	37.60	30.25	–	–	–	–
ZoomNet Xu et al. (2020)	Camera	Stereo	2020	300	55.98	38.64	30.97	–	–	–	–
Disp R-CNN Sun et al. (2020b)	Camera	Stereo	2020	420	58.53	37.91	31.93	–	–	–	–
P-LiDAR++ You et al. (2020)	Camera	Stereo	2020	500	61.11	42.43	36.99	–	–	–	–
Qian et al. (2020)	Camera	Stereo	2020	400	64.8	43.9	38.1	–	–	–	–
DSGN Chen et al. (2020e)	Camera	Stereo	2020	670	73.50	52.18	45.14	–	–	–	–
CG-Stereo Li et al. (2020a)	Camera	Stereo	2020	–	74.39	53.58	46.50	–	–	–	–
CDN Garg et al. (2020)	Camera	Stereo	2020	600	74.52	54.22	46.36	–	–	–	–
PLUMENet Wang et al. (2021i)	Camera	Stereo	2021	150	82.97*	66.27*	56.70*	–	–	–	–

Table 17 continued

Method	Sensor	Representation	Year	Inference Time (ms)	KITTI Car			nuScenes		Waymo Vehicle	
					Easy	Mod	Hard	mAP	NDS	L1	L2
LIGA-Stereo Guo et al. (2021)	Camera	Stereo	2021	350	81.39	64.66	57.22	–	–	–	–
Rubino et al. (2017)	Camera	Multi-View	2017	–	–	–	–	–	–	–	–
ImVoxelNet Rukhovich et al. (2022)	Camera	Multi-View	2021	400	17.15	10.97	9.15	–	–	–	–
DETR3D Wang et al. (2022b)	Camera	Multi-View	2021	–	–	–	–	41.2	47.9	–	–
PETR Liu et al. (2022a)	Camera	Multi-View	2022	–	–	–	–	44.5	50.4	–	–
BEVDet Huang et al. (2021)	Camera	Multi-View	2022	–	–	–	–	42.2	52.9	–	–
BEVerse Zhang et al. (2022b)	Camera	Multi-View	2022	–	–	–	–	39.3	53.1	–	–
BEVFormer Li et al. (2022f)	Camera	Multi-View	2022	–	–	–	–	48.1	56.9	–	–
BEVDepth Li et al. (2022c)	Camera	Multi-View	2022	–	–	–	–	52.0	60.9	–	–
BEVStereo Li et al. (2022a)	Camera	Multi-View	2022	–	–	–	–	52.5	61.0	–	–
SOLOFusion Park et al. (2022)	Camera	Multi-View	2022	–	–	–	–	54.0	61.9	–	–
F-PointNet Qi et al. (2018)	Fusion	Early	2018	–	81.20	70.39	62.19	–	–	–	–
RoarNet Shin et al. (2019)	Fusion	Early	2019	100	83.95	75.79	67.88	–	–	–	–
F-ConvNet Wang and Jia (2019)	Fusion	Early	2019	–	85.88	76.51	68.08	–	–	–	–
PointPainting Vora et al. (2020)	Fusion	Early	2020	–	82.11	71.70	67.08	46.4	58.1	–	–
FusionPainting Xu et al. (2021c)	Fusion	Early	2021	–	–	–	–	66.5	70.7	–	–
PointAugmenting Wang et al. (2021a)	Fusion	Early	2021	542	–	–	–	66.8	71.0	67.41	62.70
MVP Yin et al. (2021b)	Fusion	Early	2021	–	–	–	–	66.4	70.5	–	-
MV3D Chen et al. (2017b)	Fusion	Intermediate	2017	240	71.09	62.35	55.12	–	–	–	–
PointFusion Xu et al. (2018)	Fusion	Intermediate	2018	–	–	–	–	–	–	–	–
AVOD Ku et al. (2018)	Fusion	Intermediate	2018	100	81.94	71.88	66.38	–	–	–	–

Table 18 A comprehensive performance analysis of all branches of 3D object detection methods across different datasets. (Continued)

Method	Sensor	Representation	Year	Inference Time (ms)	KITTI Car			nuScenes		Waymo Vehicle	
					Easy	Mod	Hard	mAP	NDS	L1	L2
ContFuse Liang et al. (2018)	Fusion	Intermediate	2018	60	82.54	66.22	64.04	–	–	–	–
MVX-Net Sindagi et al. (2019)	Fusion	Intermediate	2019	–	83.2	72.7	65.2	–	–	–	–
MMF Liang et al. (2019)	Fusion	Intermediate	2019	80	86.81	76.75	68.41	–	–	–	–
3D-CVF Yoo et al. (2020)	Fusion	Intermediate	2020	75	89.20	80.05	73.11	52.7	62.3	–	–
EPNet Huang et al. (2020b)	Fusion	Intermediate	2020	–	89.81	79.28	74.59	–	–	–	–
TransFusion Bai et al. (2022)	Fusion	Intermediate	2022	–	–	–	–	68.9	71.7	–	–
BEVFusion Liang et al. (2022)	Fusion	Intermediate	2022	–	–	–	–	69.2	71.8	–	–
UVTR Li et al. (2022b)	Fusion	Intermediate	2022	–	–	–	–	67.1	71.1	–	–
CLOCs Pang et al. (2020)	Fusion	Late	2020	150	88.94	80.67	77.15	–	–	–	–
Fast-CLOCs Pang et al. (2022)	Fusion	Late	2022	125	89.11	80.34	76.98	63.1	68.7	–	–

based approaches are also quite efficient, e.g. (Yang et al., 2018b) and (Meyer et al., 2019b) only requires 30 ms for one-pass inference. The point-based detectors can obtain a good performance, but their inference speeds are greatly influenced by the choices of sampling and operators.

For point-based 3D object detectors, moderate AP has been increasing from 53.46% (Shi et al., 2019) to 79.57% (Shi & Rajkumar, 2020) on the KITTI benchmark. The performance improvements are mainly owing to two factors: more robust point cloud samplers and more powerful point cloud operators. The development of point cloud samplers starts with Farthest Point Sampling (FPS) (Shi et al., 2019; Yang et al., 2019), and many following point cloud detectors have been improving point cloud samplers based on FPS, including fusion-based FPS (Yang et al., 2020c), target-based FPS (Ngiam et al., 2019), FPS with coordinates refinement (Pan et al., 2021). A good point cloud sampler could produce candidate points that have better coverage of the whole scene, so it avoids missing detections when the point cloud is sparse, which helps improve the detection performance. Besides point cloud samplers, point cloud operators have also progressed rapidly, from the standard set abstraction (Shi et al., 2019; Yang et al., 2018c, 2019, 2020c) to graph operators (Shi & Rajkumar, 2020; Ngiam et al., 2019) and Transformers (Pan et al., 2021). Point cloud operators are crucial for extracting powerful feature representations from point clouds. Hence powerful point cloud operators can help detectors better obtain semantic information about 3D objects and improve performance.

For grid-based 3D object detectors, moderate AP has been increasing from 50.81% (Beltrán et al., 2018) to 82.09% (Mao et al., 2021c) on the KITTI benchmark. The performance improvements are mainly driven by better backbone networks and detection heads. The development of backbone networks has experienced four stages: (1) 2D networks to process BEV images that are generated by point cloud projection (Beltrán et al., 2018; Yang et al., 2018b), (2) 2D networks to process pillars that are generated by PointNet encoding (Lang et al., 2019), (3) 3D sparse convolutional networks to process voxelized point clouds (Zhou & Tuzel, 2018), (4) Transformer-based architectures (Mao et al., 2021c; Fan et al., 2022; Sun et al., 2022). The trend of backbone designs is to encode more 3D information from point clouds, which leads to more powerful BEV representations and better detection performance, but those early designs are still popular due to efficiency. Detection head designs have experienced the transition from anchor-based heads (Yan et al., 2018) to center-based heads (Yin et al., 2021a), and the object localization ability has been improved with the development of detection heads. Other head designs such as IoU rectification (Zheng et al., 2021a) and sequential head (Xue et al., 2022) can further boost performance.

For point-voxel based 3D object detectors, moderate AP has been increasing from 75.73% (Chen et al., 2019c) to 82.08% (Mao et al., 2021a) on the KITTI benchmark. The performance improvements come from more power operators (Liu et al., 2019b; He et al., 2020a) and modules (Shi et al., 2020a, 2021a; Mao et al., 2021a; Sheng et al., 2021) that can effectively fuse point and voxel features.

For range-based 3D object detectors, L1 mAP has been increasing from 52.11% (Meyer et al., 2019b) to 78.4% (Sun et al., 2021) on the Waymo Open dataset. The performance improvements come from designs of specialized operators (Bewley et al., 2020; Fan et al., 2021; Chai et al., 2021) that can handle range images more effectively, as well as view transforms and multi-view aggregation (Liang et al., 2020b, 2021c; Sun et al., 2021).

10.1.4 Trends of the Camera-Based Methods

Camera-based 3D object detection has shown rapid progress recently. Among the camera-based methods, the stereo-based detection methods generally outperform the monocular detection approaches by a large margin. For example, the state-of-the-art stereo-based method (Guo et al., 2021) attains 64.66% moderate AP_{3D}, while the state-of-the-art monocular method (Park et al., 2021) only achieves 16.34% moderate AP_{3D}. This is mainly because depth and disparity estimated from stereo images are much more accurate than those estimated from monocular images, and accurate depth estimation is the most important factor in camera-based 3D object detection. Multi-camera 3D object detection has been progressing fast with the emergence of BEV perception and Transformers. State-of-the-art method (Park et al., 2022) attains 54.0% mAP and 61.9 NDS on nuScenes, which has outperformed some prestigious LiDAR-based 3D object detectors (Lang et al., 2019).

For monocular 3D object detectors, moderate AP has been increasing from 1.51% (Liu et al., 2019a) to 16.34% (Park et al., 2021) on the KITTI benchmark. The major challenge of monocular 3D object detection is how to obtain accurate 3D information from a single 2D image, as localization errors dominate detection errors. The performance improvements are driven by more accurate depth prediction, which can be achieved by better network architecture designs (Brazil & Liu, 2019; Wang et al., 2021g; Zhou et al., 2019b; Manhardt et al., 2019), leveraging depth images (Xu & Chen, 2018) or pseudo-LiDAR point clouds (Wang et al., 2019b; You et al., 2020), introducing geometry constraints (Mousavian et al., 2017; Cai et al., 2020; Chen et al., 2016, 2020f), and 3D object reconstruction (Chen et al., 2021a; Zakharov et al., 2020; Xiang et al., 2015; Kundu et al., 2018).

For stereo-based 3D object detectors, moderate AP has been increasing from 4.37% (Qin et al., 2019b) to 64.66% (Guo et al., 2021) on the KITTI benchmark. The performance improvements mainly come from better network designs and data representations. Early works (Li et al., 2019b) rely on stereo-based 2D detection networks to produce paired object bounding boxes and then predict object-centric stereo/depth information with a sub-network. However, those object-centric methods generally lack global disparity information which hampers accurate 3D detection in a scene. Later on, pseudo-LiDAR based approaches (Wang et al., 2019b) generate disparity maps from stereo images and then transform disparity maps into 3D pseudo-LiDAR point clouds that are finally passed to a LiDAR detector to perform 3D detection. The transformation from 2D disparity maps to 3D point clouds is crucial and can significantly boost 3D detection performance. Many following papers are based on the pseudo-LiDAR paradigm and improve it with stronger stereo matching network (You et al., 2020) and end-to-end training of stereo matching and LiDAR detection (Qian et al., 2020). Recent methods (Wang et al., 2021i; Chen et al., 2020e) transforms disparity maps into 3D volumes and apply grid-based detectors on the volumes, which results in better performance.

For multi-view 3D object detection, mAP has been increasing from 41.2% (Wang et al., 2022b) to 54.0% (Park et al., 2022) on the nuScenes dataset. For BEV-based approaches, the performance improvements are mainly from better depth prediction (Li et al., 2022c, a). More accurate depth information results in more accurate camera-to-BEV transformation so detection performance can be improved. For query-based methods, the performance improvements come from better designs of 3D object queries (Li et al., 2022f), more powerful image features (Liu et al., 2022a), and new attention mechanisms (Doll et al., 2022).

10.1.5 Trends of the Multi-modal Methods

The multi-modal methods generally exhibit a performance improvement over the single-modal baselines but at the cost of introducing additional inference time. For instance, the multi-modal detector (Wang et al., 2021a) outperforms the LiDAR baseline (Yin et al., 2021a) by 8.8% mAP on nuScenes, but the inference time of Wang et al. (2021a) also increases to 542 ms compared to the baseline 70 ms. The problem can be more severe in the early-fusion based approaches, where the 2D networks and the 3D detection networks are connected in a sequential manner. Most multi-modal detection methods are designed and tested on the KITTI dataset, in which only a front-view image and the corresponding point cloud are utilized. Recently more and more methods are proposed and evaluated on the nuScenes dataset, in which multi-view images, point clouds, and high-definition maps are provided.

For early-fusion based methods, moderate AP increases from 70.39% (Qi et al., 2018) to 76.51% (Wang & Jia,

2019) on the KITTI benchmark, and mAP increases from 46.4% (Vora et al., 2020) to 66.8% (Wang et al., 2021a) on nuScenes dataset. There are two crucial factors that contribute to the performance increase: knowledge fusion and data augmentation. From the results, we can observe that point-level knowledge fusion (Vora et al., 2020; Xu et al., 2021c) is generally more effective than region-level fusion (Qi et al., 2018; Wang & Jia, 2019). This is because region-level knowledge fusion simply reduces the detection range, while point-level knowledge fusion can provide fine-grained semantic information which is more beneficial in 3D detection. Besides, consistent data augmentations between point clouds and images (Wang et al., 2021a) can also significantly boost detection performance.

For intermediate and late fusion based methods, moderate AP increases from 62.35% (Chen et al., 2017b) to 80.67% (Pang et al., 2020) on the KITTI benchmark, and mAP increases from 52.7% (Yoo et al., 2020) to 69.2% (Liu et al., 2022b) on the nuScenes dataset. Most methods focus on three critical problems: where to fuse different data representations, how to fuse these representations, and how to build reliable alignments between points and image pixels. For the where-to-fuse problem, different approaches try to fuse image and LiDAR features at different places, e.g. 3D backbone networks, BEV feature maps, RoI heads, and outputs. From the results we can observe that fusion at any place can boost detection performance over single-modality baselines, and fusion in the BEV space (Liu et al., 2022b; Liang et al., 2022; Bai et al., 2022) is more popular recently for its performance and efficiency. For the how-to-fuse problem, the development of fusion operators has experienced simple concatenation (Ku et al., 2018), continuous convolutions (Liang et al., 2019, 2018), attention (Yoo et al., 2020; Huang et al., 2020b), and Transformers (Bai et al., 2022; Li et al., 2022b; Chen et al., 2022a), and fusion with Transformers exhibit prominent performance on all benchmarks. For the point-to-pixel alignment problem, most papers reply on fixed extrinsics and intrinsics to construct point-to-pixel correspondences. However, due to occlusion and calibration errors, those correspondences can be noisy and misalignment will harm performance. Recent works (Liu et al., 2022b) circumvent this problem by directly fusing camera and LiDAR BEV feature maps, which is more robust to noise.

10.1.6 Systematic Comparisons

Considering all the input sensors and modalities, LiDAR-based detection is the best solution to the 3D object detection problem, in terms of both speed and accuracy. For instance, (Yin et al., 2021a) achieves 80.28% moderate AP_{3D} and still runs at 30 FPS on KITTI. Multi-modal detection is built upon LiDAR-based detection, and can obtain a better detection performance compared to the LiDAR baselines, becoming state-of-the-art in terms of accuracy. Camera-based 3D object detection is a much cheaper and quite efficient solution in contrast to LiDAR and multi-modal detection. Nevertheless, the camera-based methods generally have a worse detection performance due to inaccurate depth predictions from images. The state-of-the-art monocular (Park et al., 2021) and stereo (Guo et al., 2021) detection approach only obtain 16.34% and 64.66% moderate AP_{3D} respectively on KITTI. Recent advances in multi-view 3D object detection are quite promising. The state-of-the-art (Park et al., 2022) achieves 54.0% mAP on nuScenes, which could perform on par with some classic LiDAR detectors (Yan et al., 2018). In conclusion, LiDAR-based and multi-modal detectors are the best solutions considering speed and accuracy as the dominant factors, while camera-based detectors can be the best choice considering cost as the most important factor, and multi-view 3D detectors are becoming promising and may outperform LiDAR detectors in the future.

10.2 Future Outlooks

With all the reviewed literature and the analysis of research trends over the past years, we can now make some predictions on the future research directions of 3D object detection.

10.2.1 Open-Set 3D Object Detection

Nearly all existing works are proposed and evaluated on close datasets, in which the data only covers limited driving scenarios and the annotations only include basic classes, e.g. cars, pedestrians, cyclists. Although those datasets can be large and diverse, they are still not sufficient for real-world applications, in which critical scenarios like traffic accidents and rare classes like unknown obstacles are important but not covered by the existing datasets. Therefore, existing 3D object detectors that are trained on the close sets have a limited capacity of dealing with those critical scenarios and cannot identify the unknown categories. To overcome the above limitations, designing 3D object detectors that can learn from the open world and recognize a wide range of object categories will be a promising research direction. Cen et al. (2021) is a good start for open-set 3D object detection and hopefully more methods will be proposed to tackle this problem.

10.2.2 Detection with Stronger Interpretability

Deep learning based 3D object detection models generally lack interpretability. Namely, some important questions on how the networks can identify 3D objects in point clouds, how occlusion and noise of 3D objects can affect the model outputs, and how much context information is needed for detecting a 3D object, have not been properly answered due to the black-box property of deep neural networks. On the

other hand, understanding the behaviors of 3D detectors and answering these questions are quite important if we want to perform 3D object detection in a more robust manner and avoid those unexpected cases brought by black-box detectors. Therefore, the methods that can understand and interpret the existing 3D object detection models will be appealing in future research.

10.2.3 Efficient Hardware Design for 3D Object Detection

Most existing works focus on designing algorithms to tackle the 3D object detection problem, and their models generally run on GPUs. Nevertheless, unlike image operators that are highly optimized for GPU devices, point clouds and voxels are sparse and irregular, and the commonly adopted 3D operators like set abstraction or 3D sparse convolutions are not well suited for GPUs. Hence those LiDAR object detectors cannot run as efficiently as the image detectors on the existing hardware devices. To handle this challenge, designing novel devices where the hardware architectures are optimized for 3D operators as well as the task of 3D object detection will be an important research direction and will be beneficial for real-world deployment. Lin et al. (2021) is a pioneering hardware work to accelerate point cloud processing, and we believe more and more papers will come in this field. In addition, new sensors, e.g. solid-state LiDARs, LiDARs with doppler, 4D radars, will also inspire the design of 3D object detectors.

10.2.4 Detection in End-to-End Self-Driving Systems

Most existing works treat 3D object detection as an independent task and try to maximize the detection metrics such as average precision. Nevertheless, 3D object detection is closely correlated with other perception tasks as well as downstream tasks such as prediction and planning, so simply pursuing high average precision for 3D object detection may not be optimal when considering the autonomous driving system as a whole. Therefore, conducting 3D object detection and other tasks in an end-to-end manner, and learning 3D detectors from the feedback of planners, will be the future research trends of 3D object detection.

11 Conclusion

In this paper, we comprehensively review and analyze various aspects of 3D object detection for autonomous driving. We start from the problem definition, datasets, and evaluation metrics for 3D object detection, and then we introduce various kinds of sensor-based 3D object detection approaches, including LiDAR-based, camera-based, and multi-modal 3D object detection methods. We further investigate 3D object detection leveraging temporal data, with label-efficient learning, as well as its applications in autonomous driving systems. Finally, we summarize the research trends in recent years and prospect the future research directions of 3D object detection.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s11263-023-01790-1.

Acknowledgements This project is funded in part by the National Key R&D Program of China Project 2022ZD0161100, by the Centre for Perceptual and Interactive Intelligence (CPII) Ltd under the Innovation and Technology Commission (ITC)'s InnoHK, by General Research Fund Project 14204021 and Research Impact Fund Project R5001-18 of Hong Kong RGC. Hongsheng Li and Xiaogang Wang are PIs of CPII under the InnoHK.

References

Abu Alhaija, H., Mustikovela, S. K., Mescheder, L., Geiger, A., & Rother, C. (2018). Augmented reality meets computer vision: Efficient data generation for urban driving scenes. *IJCV, 126*, 961–972.

Aghdam, H. H., Heravi, E. J., Demilew, S. S., & Laganiere, R. (2021). Rad: Realtime and accurate 3D object detection on embedded systems. In *CVPR*.

Ali, W., Abdelkarim, S., Zidan, M., Zahran, M., & El Sallab. A. (2018). YOLO3D: End-to-end real-time 3D oriented object bounding box detection from lidar point cloud. In *ECCVW*.

Amini, A., Gilitschenski, I., Phillips, J., Moseyko, J., Banerjee, R., Karaman, S., & Rus, D. (2020). Learning robust control policies for end-to-end autonomous driving from data-driven simulation. *IEEE RA-L, 5*, 1143–1150.

Arnold, E., Al-Jarrah, O. Y., Dianati, M., Fallah, S., Oxtoby, D., & Mouzakitis, A. (2019). A survey on 3D object detection methods for autonomous driving applications. *IEEE T-ITS, 20*, 3782–3795.

Bai, X., Hu, Z., Zhu, X., Huang, Q., Chen, Y., Fu, H., & Tai, C.-L. (2022). Transfusion: Robust lidar-camera fusion for 3d object detection with transformers. In *CVPR*.

Bao, W., Xu, B., & Chen, Z. (2019). Monofenet: Monocular 3D object detection with feature enhancement networks. *IEEE T-IP, 29*, 2753–2765.

Barrera, A., Guindel, C., Beltrán, J., & García, F. (2020). Birdnet+: End-to-end 3D object detection in lidar bird's eye view. In *ITSC*.

Beker, D., Kato, H., Morariu, M. A., Ando, T., Matsuoka, T., Kehl, W., & Gaidon, A. (2020). Monocular differentiable rendering for self-supervised 3d object detection. In *ECCV*.

Beltrán, J., Guindel, C., Moreno, F. M., Cruzado, D., Garcia, F., & De La Escalera, A. (2018). Birdnet: A 3d object detection framework from lidar information. In *ITSC*.

Bewley, A., Sun, P., Mensink, T., Anguelov, D., & Sminchisescu, C. (2020). Range conditioned dilated convolutions for scale invariant 3d object detection. arXiv preprint arXiv:2005.09927

Bojarski, M., Del Testa, D., Dworakowski, D., Firner, B., Flepp. B., Goyal, P., Jackel, L. D., Monfort, M., Muller, U., & Zhang. J., et al. (2016). End to end learning for self-driving cars. arXiv preprint arXiv:1604.07316

Brazil, G., & Liu, X. (2019). M3d-rpn: Monocular 3d region proposal network for object detection. In *ICCV*.

Brazil, G., Pons-Moll, G., Liu, X., & Schiele, B. (2020). Kinematic 3d object detection in monocular video. In *ECCV*.

Caesar, H., Bankiti, V., Lang, A. H., Vora, S., Liong, V. E., Xu, Q., Krishnan, A., Pan, Y., Baldan, G., & Beijbom, O. (2020). nuscenes: A multimodal dataset for autonomous driving. In *CVPR*.

Caesar, H., Kabzan, J., Tan, K. S., Fong, W. K., Wolff, E., Lang, A., Fletcher, L., Beijbom, O., & Omari, S. (2021). nuplan: A closed-loop ml-based planning benchmark for autonomous vehicles. arXiv preprint arXiv:2106.11810

Cai, Y., Li, B., Jiao, Z., Li, H., Zeng, X., & Wang, X. (2020). Monocular 3d object detection with decoupled structured polygon estimation and height-guided depth estimation. In *AAAI*.

Caine, B., Roelofs, R., Vasudevan, V., Ngiam, J., Chai, Y., Chen, Z., & Shlens, J. (2021). Pseudo-labeling for scalable 3d object detection. arXiv preprint arXiv:2103.02093

Cao, Y., Xiao, C., Cyr, B., Zhou, Y., Park, W., Rampazzi, S., Chen, Q. A., Fu, K., & Mao, Z. M. (2019). Adversarial sensor attack on lidar-based perception in autonomous driving. In *ACM SIGSAC*.

Cao, Y., Wang, N., Xiao, C., Yang, D., Fang, J., Yang, R., Chen, Q. A., Liu, M., & Li, B. (2021). Invisible for both camera and lidar: Security of multi-sensor fusion based perception in autonomous driving under physical-world attacks. In *IEEE Symposium on Security and Privacy*.

Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., & Zagoruyko, S. (2020). End-to-end object detection with transformers. In *ECCV*.

Casas, S., Luo, W., & Urtasun, R. (2018). Intentnet: Learning to predict intention from raw sensor data. In *CoRL*.

Casas, S., Sadat, A., & Urtasun. R. (2021). Mp3: A unified model to map, perceive, predict and plan. In *CVPR*.

Cen, J., Yun, P., Cai, J., Wang, M. Y., & Liu, M. (2021). Open-set 3d object detection. In *3DV*.

Chabot, F., Chaouch, M., Rabarisoa, J., Teuliere, C., & Chateau, T. (2017). Deep manta: A coarse-to-fine many-task network for joint 2d and 3d vehicle analysis from monocular image. In *CVPR*.

Chadwick, S., Maddern, W., & Newman, P. (2019). Distant vehicle detection using radar and vision. In *ICRA*.

Chai, Y., Sun, P., Ngiam, J., Wang, W., Caine, B., Vasudevan, V., Zhang, X., & Anguelov, D. (2021). To the point: Efficient 3d object detection in the range image with graph convolution kernels. In *CVPR*.

Chang, J., & Wetzstein, G. (2019). Deep optics for monocular depth estimation and 3d object detection. In *ICCV*.

Chang, J.-R., & Chen, Y.-S. (2018). Pyramid stereo matching network. In *CVPR*.

Chang, M.-F., Lambert, J., Sangkloy, P., Singh, J., Bak, S., Hartnett, A., Wang, D., Carr, P., Lucey, S., & Ramanan, D., et al. (2019). Argoverse: 3d tracking and forecasting with rich maps. In *CVPR*.

Chen, H., Huang, Y., Tian, W., Gao, Z., & Xiong, L. (2021a). Monorun: Monocular 3d object detection by reconstruction and uncertainty propagation. In *CVPR*.

Chen, L., Sun, J., Xie, Y., Zhang, S., Shuai, Q., Jiang, Q., Zhang, G., Bao, H., & Zhou, X. (2021b). Shape prior guided instance disparity estimation for 3d object detection. *IEEE T-PAMI*.

Chen, Q., Ma, X., Tang, S., Guo, J., Yang, Q., & Fu, S. (2019a). F-cooper: Feature based cooperative perception for autonomous vehicle edge computing system using 3d point clouds. In *ACM/IEEE symposium on edge computing*.

Chen, Q., Tang, S., Yang, Q., & Fu, S. (2019b). Cooper: Cooperative perception for connected autonomous vehicles based on 3d point clouds. In *ICDCS*.

Chen, Q., Sun, L., Cheung, E., & Yuille, A. L. (2020a). Every view counts: Cross-view consistency in 3d object detection with hybrid-cylindrical-spherical voxelization. *NeurIPS*.

Chen, Q., Sun, L., Wang, Z., Jia, K., & Yuille, A. (2020b). Object as hotspots: An anchor-free 3d object detection approach via firing of hotspots. In *ECCV*.

Chen, Q., Vora, S., & Beijbom, O. (2021c). Polarstream: Streaming lidar object detection and segmentation with polar pillars. arXiv preprint arXiv:2106.07545

Chen, X., Kundu, K., Zhu, Y., Berneshawi, A. G., Ma, H., Fidler, S., & Urtasun, R. (2015). 3d object proposals for accurate object class detection. *NeurIPS*.

Chen, X., Kundu, K., Zhang, Z., Ma, H., Fidler, S., & Urtasun, R. (2016). Monocular 3d object detection for autonomous driving. In *CVPR*.

Chen, X., Kundu, K., Zhu, Y., Ma, H., Fidler, S., & Urtasun, R. (2017a). 3d object proposals using stereo imagery for accurate object class detection. *IEEE T-PAMI*.

Chen, X., Ma, H., Wan, J., Li, B., & Xia, T. (2017b). Multi-view 3d object detection network for autonomous driving. In *CVPR*.

Chen, X., Fan, H., Girshick, R., & He, K. (2020c). Improved baselines with momentum contrastive learning. arXiv preprint arXiv:2003.04297

Chen, X., Zhang, T., Wang, Y., Wang, Y., & Zhao, H. (2022a). Futr3d: A unified sensor fusion framework for 3d detection. arXiv preprint arXiv:2203.10642

Chen, Y., Liu, S., Shen, X., & Jia, J. (2019c). Fast point R-CNN. In *ICCV*.

Chen, Y., Li, H., Gao, R., & Zhao, D. (2020d). Boost 3-d object detection via point clouds segmentation and fused 3-d giou-l1 loss. *IEEE T-NNLS*.

Chen, Y., Liu, S., Shen, X., & Jia, J. (2020e). Dsgn: Deep stereo geometry network for 3d object detection. In *CVPR*.

Chen, Y., Tai, L., Sun, K., & Li, M. (2020f). Monopair: Monocular 3d object detection using pairwise spatial relationships. In *CVPR*.

Chen, Y., Rong, F., Duggal, S., Wang, S., Yan, X., Manivasagam, S., Xue, S., Yumer, E., & Urtasun, R. (2021d). Geosim: Realistic video simulation via geometry-aware composition for self-driving. In *CVPR*.

Chen, Y., Li, Y., Zhang, X., Sun, J., & Jia, J. (2022b). Focal sparse convolutional networks for 3d object detection. In *CVPR*.

Chen, Z., Li, Z., Zhang, S., Fang, L., Jiang, Q., Zhao, F., Zhou, B., & Zhao, H. (2022c). Autoalign: Pixel-instance feature aggregation for multi-modal 3d object detection. In *IJCAI*.

Choi, Y., Kim, N., Hwang, S., Park, K., Yoon, J. S., An, K., & Kweon, I. S. (2018). Kaist multi-spectral day/night data set for autonomous and assisted driving. *T-ITS*.

Codevilla, F., Müller, M., López, A., Koltun, V., & Dosovitskiy, A. (2018). End-to-end driving via conditional imitation learning. In *ICRA*.

Cui, A., Casas, S., Sadat, A., Liao, R., & Urtasun, R. (2021). Lookout: Diverse multi-future prediction and planning for self-driving. In *ICCV*.

Dai, A., Chang, A. X., Savva, M., Halber, M., Funkhouser, T., & Nießner, M. (2017). Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *CVPR*.

DeBortoli, R., Fuxin, L., Kapoor, A., & Hollinger, G. A. (2021). Adversarial training on point clouds for sim-to-real 3d object detection. *IEEE RA-L*.

Deng, B., Qi, C. R., Najibi, M., Funkhouser, T., Zhou, Y., & Anguelov, D. (2021a). Revisiting 3d object detection from an egocentric perspective. *NeurIPS*.

Deng, J., Shi, S., Li, P., Zhou, W., Zhang, Y., & Li, H. (2021b). Voxel r-cnn: Towards high performance voxel-based 3d object detection. In *AAAI*.

Deng, J., Zhou, W., Zhang, Y., & Li, H. (2021c). From multi-view to hollow-3d: Hallucinated hollow-3d r-CNN for 3d object detection. *IEEE T-CSVT*.

Deng, S., Liang, Z., Sun, L., & Jia, K. (2022). Vista: Boosting 3d object detection via dual cross-view spatial attention. In *CVPR*.

Ding, M., Huo, Y., Yi, H., Wang, Z., Shi, J., Lu, Z., & Luo, P. (2020). Learning depth-guided convolutions for monocular 3d object detection. In *CVPRW*.

Doll, S., Schulz, R., Schneider, L., Benzin, V., Enzweiler, M., & Lensch, H. P. (2022). Spatialdetr: Robust scalable transformer-based 3d

object detection from multi-view camera images with global cross-sensor attention. In *ECCV*.
Dosovitskiy, A., Ros, G., Codevilla, F., Lopez, A., & Koltun, V. (2017). Carla: An open urban driving simulator. In *CoRL*.
Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., & Gelly, S., et al. (2021). An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*.
Dou, J., Xue, J., & Fang, J. (2019). Seg-voxelnet for 3d vehicle detection from rgb and lidar data. In *ICRA*.
Du, L., Ye, X., Tan, X., Feng, J., Xu, Z., Ding, E., & Wen, S. (2020). Associate-3ddet: Perceptual-to-conceptual association for 3d point cloud object detection. In *CVPR*.
Du, L., Ye, X., Tan, X., Johns, E., Chen, B., Ding, E., Xue, X., & Feng, J. (2021). Ago-net: Association-guided 3d point cloud object detection network. *IEEE T-PAMI*.
Du, X., Ang, M. H., Karaman, S., & Rus, D. (2018). A general pipeline for 3d detection of vehicles. In *ICRA*.
Engelcke, M., Rao, D., Wang, D. Z., Tong, C. H., & Posner, I. (2017). Vote3deep: Fast object detection in 3d point clouds using efficient convolutional neural networks. In *ICRA*.
Fan, L., Xiong, X., Wang, F., Wang, N., & Zhang, Z. (2021). Rangedet: In defense of range view for lidar-based 3d object detection. In *ICCV*.
Fan, L., Pang, Z., Zhang, T., Wang, Y.-X., Zhao, H., Wang, F., Wang, N., & Zhang, Z. (2022). Embracing single stride 3d object detector with sparse transformer. In *CVPR*.
Fang, J., Zhou, D., Yan, F., Zhao, T., Zhang, F., Ma, Y., Wang, L., & Yang, R. (2020). Augmented lidar simulator for autonomous driving. *IEEE RA-L*.
Fang, J., Zhou, D., Song, X., & Zhang, L. (2021a). Mapfusion: A general framework for 3d object detection with hdmaps. In *IROS*.
Fang, J., Zuo, X., Zhou, D., Jin, S., Wang, S., & Zhang, L. (2021b). Lidar-aug: A general rendering-based augmentation framework for 3d object detection. In *CVPR*.
Feng, M., Gilani, S. Z., Wang, Y., Zhang, L., & Mian, A. (2020). Relation graph network for 3d object detection in point clouds. *IEEE T-IP*.
Fernandes, D., Silva, A., Névoa, R., Simões, C., Gonzalez, D., Guevara, M., Novais, P., Monteiro, J., & Melo-Pinto, P. (2021). Point-cloud based 3d object detection and classification methods for self-driving applications: A survey and taxonomy. *Information Fusion*.
Frossard, D., Da Suo, S., Casas, S., Tu, J., & Urtasun, R. (2021). Strobe: Streaming object detection from lidar packets. In *CoRL*.
Fruhwirth-Reisinger, C., Opitz, M., Possegger, H., & Bischof, H. (2021). Fast3d: Flow-aware self-training for 3d object detectors. In *BMVC*.
Fu, H., Gong, M., Wang, C., Batmanghelich, K., & Tao, D. (2018). Deep ordinal regression network for monocular depth estimation. In *CVPR*.
Gählert, N., Jourdan, N., Cordts, M., Franke, U., & Denzler, J. (2020). Cityscapes 3d: Dataset and benchmark for 9 dof vehicle detection. arXiv preprint arXiv:2006.07864
Garg, D., Wang, Y., Hariharan, B., Campbell, M., Weinberger, K. Q., & Chao, W.-L. (2020). Wasserstein distances for stereo disparity estimation. *NeurIPS*.
Ge, R., Ding, Z., Hu, Y., Wang, Y., Chen, S., Huang, L., & Li, Y. (2020). Afdet: Anchor free one stage 3d object detection. arXiv preprint arXiv:2006.12671
Geiger, A., Lenz, P., & Urtasun, R. (2012). Are we ready for autonomous driving? The kitti vision benchmark suite. In *CVPR*.
Geiger, A., Lenz, P., Stiller, C., & Urtasun, R. (2013). Vision meets robotics: The Kitti dataset. *IJRR*.
Geyer, J., Kassahun, Y., Mahmudi, M., Ricou, X., Durgesh, R., Chung, A. S., Hauswald, L., Pham, V. H., Mühlegg, M., & Dorn, S., et al. (2020). A2d2: Audi autonomous driving dataset. arXiv preprint arXiv:2004.06320
Godard, C., Mac Aodha, O., & Brostow, G. J. (2017). Unsupervised monocular depth estimation with left-right consistency. In *CVPR*.
Graham, B., Engelcke, M., & Van Der Maaten, L. (2018). 3d semantic segmentation with submanifold sparse convolutional networks. In *CVPR*.
Gu, Q., Zhou, Q., Xu, M., Feng, Z., Cheng, G., Lu, X., Shi, J., & Ma, L. (2021). Pit: Position-invariant transform for cross-fov domain adaptation. In *ICCV*.
Guan, T., Wang, J., Lan, S., Chandra, R., Wu, Z., Davis, L., & Manocha, D. (2022). M3detr: Multi-representation, multi-scale, mutual-relation 3d object detection with transformers. In *WACV*.
Guo, X., Shi, S., Wang, X., & Li, H. (2021). Liga-stereo: Learning lidar geometry aware representations for stereo-based 3d detector. In *ICCV*.
Guo, Y., Wang, H., Hu, Q., Liu, H., Liu, L., & Bennamoun, M. (2020). Deep learning for 3d point clouds: A survey. *IEEE T-PAMI*.
Hahner, M., Sakaridis, C., Dai, D., & Van Gool, L. (2021). Fog simulation on real lidar point clouds for 3d object detection in adverse weather. In *ICCV*.
Han, W., Zhang, Z., Caine, B., Yang, B., Sprunk, C., Alsharif, O., Ngiam, J., Vasudevan, V., Shlens, J., & Chen, Z. (2020). Streaming object detection for 3-d point clouds. In *ECCV*.
Hartley, R., & Zisserman, A. (2003). *Multiple view geometry in computer vision*. Cambridge: Cambridge University Press.
He, C., Zeng, H., Huang, J., Hua, X.-S., & Zhang, L. (2020a). Structure aware single-stage 3d object detection from point cloud. In *CVPR*.
He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *CVPR*.
He, K., Fan, H., Wu, Y., Xie, S., & Girshick, R. (2020b). Momentum contrast for unsupervised visual representation learning. In *CVPR*.
He, Q., Wang, Z., Zeng, H., Zeng, Y., Liu, S., & Zeng, B. (2020c). Svga-net: Sparse voxel-graph attention network for 3d object detection from point clouds. arXiv preprint arXiv:2006.04043
He, T., & Soatto, S. (2019). Mono3d++: Monocular 3d vehicle detection with two-scale 3d hypotheses and task priors. In *AAAI*.
Heylen, J., De Wolf, M., Dawagne, B., Proesmans, M., Van Gool, L., Abbeloos, W., Abdelkawy, H., & Reino, D. O. (2021). Monocinis: Camera independent monocular 3d object detection using instance segmentation. In *ICCV*.
Hu, H.-N., Cai, Q.-Z., Wang, D., Lin, J., Sun, M., Krahenbuhl, P., Darrell, T., & Yu, F. (2019). Joint monocular 3d vehicle detection and tracking. In *ICCV*.
Hu, J. S., Kuai, T., & Waslander, S. L. (2022). Point density-aware voxels for lidar 3d object detection. In *CVPR*.
Hu, P., Ziglar, J., Held, D., & Ramanan, D. (2020). What you see is what you get: Exploiting visibility for 3d object detection. In *CVPR*.
Hu, Y., Ding, Z., Ge, R., Shao, W., Huang, L., Li, K., & Liu, Q. (2021). Afdetv2: Rethinking the necessity of the second stage for object detection from point clouds. arXiv preprint arXiv:2112.09205
Huang, B., Li, Y., Xie, E., Liang, F., Wang, L., Shen, M., Liu, F., Wang, T., Luo, P., & Shao, J. (2022a). Fast-bev: Towards real-time on-vehicle bird's-eye view perception. In *NeurIPS*.
Huang, J., & Huang, G. (2022). Bevdet4d: Exploit temporal cues in multi-camera 3d object detection. arXiv preprint arXiv:2203.17054
Huang, J., Huang, G., Zhu, Z., & Du, D. (2021). Bevdet: High-performance multi-camera 3d object detection in bird-eye-view. arXiv preprint arXiv:2112.11790
Huang, K.-C., Wu, T.-H., Su, H.-T., & Hsu, W. H. (2022b). Monodtr: Monocular 3d object detection with depth-aware transformer. In *CVPR*.
Huang, R., Zhang, W., Kundu, A., Pantofaru, C., Ross, D. A., Funkhouser, T., & Fathi, A. (2020a). An lstm approach to temporal 3d object detection in lidar point clouds. In *ECCV*.
Huang, T., Liu, Z., Chen, X., & Bai, X. (2020b). Epnet: Enhancing point features with image semantics for 3d object detection. In *ECCV*.

Huang, X., Wang, P., Cheng, X., Zhou, D., Geng, Q., & Yang, R. (2019). The apolloscape open dataset for autonomous driving and its application. *IEEE T-PAMI*.

Jiang, B., Chen, S., Wang, X., Liao, B., Cheng, T., Chen, J., Zhou, H., Zhang, Q., Liu, W., & Huang, C. (2022). Perceive, interact, predict: Learning dynamic and static clues for end-to-end motion prediction. arXiv preprint arXiv:2212.02181

Jörgensen, E., Zach, C., & Kahl, F. (2019). Monocular 3d object detection and box fitting trained end-to-end using intersection-over-union loss. arXiv preprint arXiv:1906.08070

Kendall, A., Hawke, J., Janz, D., Mazur, P., Reda, D., Allen, J.-M., Lam, V.-D., Bewley, A., & Shah, A. (2019). Learning to drive in a day. In *ICRA*.

Kesten, R., Usman, M., Houston, J., Pandya, T., Nadhamuni, K., Ferreira, A., Yuan, M., Low, B., Jain, A., Ondruska, P., Omari, S., Shah, S., Kulkarni, A., Kazakova, A., Tao, C., Platinsky, L., Jiang, W., & Shet, V. (2019). Lyft level 5 av dataset 2019. https://level5.lyft.com/dataset/

Kim, S. W., Philion, J., Torralba, A., & Fidler, S. (2021). Drivegan: Towards a controllable high-quality neural simulation. In *CVPR*.

Königshof, H., Salscheider, N. O., & Stiller, C. (2019). Realtime 3d object detection for automated driving using stereo vision and semantic information. In *ITSC*.

Ku, J., Mozifian, M., Lee, J., Harakeh, A., & Waslander, S. L. (2018). Joint 3d proposal generation and object detection from view aggregation. In *IROS*.

Ku, J., Pon, A. D., & Waslander, S. L. (2019). Monocular 3d object detection leveraging accurate proposals and shape reconstruction. In *CVPR*.

Kuang, H., Wang, B., An, J., Zhang, M., & Zhang, Z. (2020). Voxel-fpn: Multi-scale voxel feature aggregation for 3d object detection from lidar point clouds. *Sensors*.

Kuhn, H. W. (1955). The Hungarian method for the assignment problem. *Naval Research Logistics Quarterly, 2*, 83–97.

Kumar, A., Brazil, G., & Liu, X. (2021). Groomed-nms: Grouped mathematically differentiable nms for monocular 3d object detection. In *CVPR*.

Kundu, A., Li, Y., & Rehg, J. M. (2018). 3d-rcnn: Instance-level 3d object reconstruction via render-and-compare. In *CVPR*.

Laddha, A., Gautam, S., Meyer, G. P., Vallespi-Gonzalez, C., & Wellington, C. K. (2020). Rv-fusenet: Range view based fusion of time-series lidar data for joint 3d object detection and motion forecasting. In *IROS*.

Lang, A. H., Vora, S., Caesar, H., Zhou, L., Yang, J., & Beijbom, O. (2019). Pointpillars: Fast encoders for object detection from point clouds. In *CVPR*.

Li, B. (2017). 3d fully convolutional network for vehicle detection in point cloud. In *IROS*.

Li, B., Zhang, T., & Xia, T. (2016). Vehicle detection from 3d lidar using fully convolutional network. arXiv preprint arXiv:1608.07916

Li, B., Ouyang, W., Sheng, L., Zeng, X., & Wang, X. (2019a). Gs3d: An efficient 3d object detection framework for autonomous driving. In *CVPR*.

Li, C., Ku, J., & Waslander, S. L. (2020a). Confidence guided stereo 3d object detection with split depth estimation. In *IROS*.

Li, F., Jin, W., Fan, C., Zou, L., Chen, Q., Li, X., Jiang, H., & Liu, Y. (2021a). Psanet: Pyramid splitting and aggregation network for 3d object detection in point cloud. *Sensors*.

Li, J., Dai, H., Shao, L., & Ding, Y. (2021b). Anchor-free 3d single stage detector with mask-guided attention for point cloud. In *ACM multimedia*.

Li, J., Dai, H., Shao, L., & Ding, Y. (2021c). From voxel to point: Iou-guided 3d object detection for point cloud with voxel-to-point decoder. In *ACM multimedia*.

Li, L. L., Yang, B., Liang, M., Zeng, W., Ren, M., Segal, S., & Urtasun, R. (2020b). End-to-end contextual perception and prediction with interaction transformer. In *IROS*.

Li, P., & Zhao, H. (2021). Monocular 3d detection with geometric constraint embedding and semi-supervised training. *IEEE RA-L*.

Li, P., Chen, X., & Shen, S. (2019b). Stereo r-cnn based 3d object detection for autonomous driving. In *CVPR*.

Li, P., Zhao, H., Liu, P., & Cao, F. (2020c). Rtm3d: Real-time monocular 3d detection from object keypoints for autonomous driving. In *ECCV*.

Li, Y., Ren, S., Wu, P., Chen, S., Feng, C., & Zhang, W. (2021d). Learning distilled collaboration graph for multi-agent perception. *NeurIPS*.

Li, Y., Wen, C., Juefei-Xu, F., Feng, C. (2021e). Fooling lidar perception via adversarial trajectory perturbation. In *ICCV*.

Li, Y., Bao, H., Ge, Z., Yang, J., Sun, J., & Li, Z. (2022a). Bevstereo: Enhancing depth estimation in multi-view 3d object detection with dynamic temporal stereo. arXiv preprint arXiv:2209.10248

Li, Y., Chen, Y., Qi, X., Li, Z., Sun, J., & Jia, J. (2022b). Unifying voxel-based representation with transformer for 3d object detection. In *NeurIPS*.

Li, Y., Ge, Z., Yu, G., Yang, J., Wang, Z., Shi, Y., Sun, J., & Li, Z. (2022c). Bevdepth: Acquisition of reliable depth for multi-view 3d object detection. arXiv preprint arXiv:2206.10092

Li, Y., Qi, X., Chen, Y., Wang, L., Li, Z., Sun, J., & Jia, J. (2022d). Voxel field fusion for 3d object detection. In *CVPR*.

Li, Y., Yu, A. W., Meng, T., Caine, B., Ngiam, J., Peng, D., Shen, J., Wu, B., Lu, Y., & Zhou, D., et al. (2022e). Deepfusion: Lidar-camera deep fusion for multi-modal 3d object detection. In *CVPR*.

Li, Z., Chen, Z., Li, A., Fang, L., Jiang, Q., Liu, X., Jiang, J., Zhou, B., & Zhao, H. (2021f). Simipu: Simple 2d image and 3d point cloud unsupervised pre-training for spatial-aware visual representations. In *AAAI*.

Li, Z., Wang, F., & Wang, N. (2021g). Lidar r-cnn: An efficient and universal 3d object detector. In *CVPR*.

Li, Z., Wang, W., Li, H., Xie, E., Sima. C., Lu. T., Yu. Q., & Dai. J. (2022f). Bevformer: Learning bird's-eye-view representation from multi-camera images via spatiotemporal transformers. In *ECCV*.

Liang, H., Jiang, C., Feng, D., Chen, X., Xu, H., Liang, X., Zhang, W., Li, Z., & Van Gool, L. (2021a). Exploring geometry-aware contrast and clustering harmonization for self-supervised 3d object detection. In *ICCV*.

Liang, M., Yang, B., Wang, S., & Urtasun, R. (2018). Deep continuous fusion for multi-sensor 3d object detection. In *ECCV*.

Liang, M., Yang, B., Chen, Y., Hu, R., & Urtasun, R. (2019). Multi-task multi-sensor fusion for 3d object detection. In *CVPR*.

Liang, M., Yang, B., Zeng, W., Chen, Y., Hu, R., Casas, S., & Urtasun, R. (2020a), Pnpnet: End-to-end perception and prediction with tracking in the loop. In *CVPR*.

Liang, T., Xie, H., Yu, K., Xia, Z., Lin, Z., Wang, Y., Tang, T., Wang, B., & Tang, Z. (2022). Bevfusion: A simple and robust lidar-camera fusion framework. In *NeurIPS*.

Liang, W., Xu, P., Guo, L., Bai, H., Zhou, Y., & Chen, F. (2021b). A survey of 3d object detection. *Multimedia Tools and Applications*.

Liang, Z., Zhang, M., Zhang, Z., Zhao, X., & Pu, S. (2020b). Rangercnn: Towards fast and accurate 3d object detection with range image representation. arXiv preprint arXiv:2009.00206

Liang, Z., Zhang, Z., Zhang, M., Zhao, X., & Pu, S. (2021c). Rangeioudet: Range image based real-time 3d object detector optimized by intersection over union. In *CVPR*.

Liao, Y., Xie, J., & Geiger, A. (2021). Kitti-360: A novel dataset and benchmarks for urban scene understanding in 2d and 3d. arXiv preprint arXiv:2109.13410

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., & Zitnick, C. L. (2014). Microsoft coco: Common objects in context. In *ECCV*.

Lin, T.-Y., Dollár, P., Girshick, R., He, K., Hariharan, B., & Belongie S. (2017a). Feature pyramid networks for object detection. In *CVPR*.
Lin, T.-Y., Goyal, P., Girshick, R., He, K., & Dollár, P. (2017b). Focal loss for dense object detection. In *ICCV*.
Lin, Y., Zhang, Z., Tang, H., Wang, H., & Han, S. (2021). Pointacc: Efficient point cloud accelerator. In *MICRO*.
Liu, L., Lu, J., Xu, C., Tian, Q., & Zhou, J. (2019a). Deep fitting degree scoring network for monocular 3d object detection. In *CVPR*.
Liu, Y., Wang, L., & Liu, M. (2021a). Yolostereo3d: A step back to 2d for efficient stereo 3d detection. In *ICRA*.
Liu, Y., Yixuan, Y., & Liu, M. (2021b). Ground-aware monocular 3d object detection for autonomous driving. *IEEE RA-L*.
Liu, Y., Wang, T., Zhang, X., & Sun, J. (2022a). Petr: Position embedding transformation for multi-view 3d object detection. In *ECCV*.
Liu, Y.-C., Tian, J., Glaser, N., & Kira, Z. (2020a). When2com: Multi-agent perception via communication graph grouping. In *CVPR*.
Liu, Y.-C., Tian, J., Ma, C.-Y., Glaser, N., Kuo, C.-W., & Kira, Z. (2020b). Who2com: Collaborative perception via learnable handshake communication. In *ICRA*.
Liu, Z., Tang, H., Lin, Y., & Han, S. (2019b). Point-voxel cnn for efficient 3d deep learning. *NeurIPS*.
Liu, Z., Wu, Z., & Tóth, R. (2020c). Smoke: Single-stage monocular 3d object detection via keypoint estimation. In *CVPRW*.
Liu, Z., Zhao, X., Huang, T., Hu, R., Zhou, Y., & Bai, X. (2020d). Tanet: Robust 3d object detection from point clouds with triple attention. In *AAAI*.
Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., & Guo, B. (2021c). Swin transformer: Hierarchical vision transformer using shifted windows. In *ICCV*.
Liu, Z., Zhang, Z., Cao, Y., Hu, H., & Tong, X. (2021d). Group-free 3d object detection via transformers. In *ICCV*.
Liu, Z., Tang, H., Amini, A., Yang, X., Mao, H., Rus, D., & Han, S. (2022b). Bevfusion: Multi-task multi-sensor fusion with unified bird's-eye view representation. arXiv preprint arXiv:2205.13542
Long, J., Shelhamer, E., & Darrell, T. (2015). Fully convolutional networks for semantic segmentation. In *CVPR*.
Lu, Y., Ma, X., Yang, L., Zhang, T., Liu, Y., Chu, Q., Yan, J., & Ouyang, W. (2021). Geometry uncertainty projection network for monocular 3d object detection. In *ICCV*.
Luo, S., Dai, H., Shao, L., & Ding, Y. (2021a). M3dssd: Monocular 3d single stage object detector. In *CVPR*.
Luo, W., Yang, B., & Urtasun, R. (2018). Fast and furious: Real time end-to-end 3d detection, tracking and motion forecasting with a single convolutional net. In *CVPR*.
Luo, Z., Cai, Z., Zhou, C., Zhang, G., Zhao, H., Yi, S., Lu, S., Li, H., Zhang, S., & Liu, Z. (2021b). Unsupervised domain adaptive 3d detection with multi-level consistency. In *ICCV*.
Ma, X., Wang, Z., Li, H., Zhang, P., Ouyang, W., & Fan, X. (2019a). Accurate monocular 3d object detection via color-embedded 3d reconstruction for autonomous driving. In *ICCV*.
Ma, X., Liu, S., Xia, Z., Zhang, H., Zeng, X., & Ouyang, W. (2020). Rethinking pseudo-lidar representation. In *ECCV*.
Ma, X., Zhang, Y., Xu, D., Zhou, D., Yi, S., Li, H., & Ouyang, W. (2021). Delving into localization errors for monocular 3d object detection. In *CVPR*.
Ma, X., Ouyang, W., Simonelli, A., & Ricci, E. (2022). 3d object detection from images for autonomous driving: A survey. arXiv preprint arXiv:2202.02980
Ma, Y., Zhu, X., Zhang, S., Yang, R., Wang, W., & Manocha, D. (2019b). Trafficpredict: Trajectory prediction for heterogeneous traffic-agents. In *AAAI*.
Major, B., Fontijne, D., Ansari, A., Teja Sukhavasi, R., Gowaikar, R., Hamilton, M., Lee, S., Grzechnik, S., & Subramanian, S. (2019). Vehicle detection with automotive radar using deep learning on range-azimuth-doppler tensors. In *ICCVW*.
Manhardt, F., Kehl, W., & Gaidon, A. (2019). Roi-10d: Monocular lifting of 2d detection to 6d pose and metric shape. In *CVPR*.
Manivasagam, S., Wang, S., Wong, K., Zeng, W., Sazanovich, M., Tan, S., Yang, B., Ma, W.-C., & Urtasun, R. (2020). Lidarsim: Realistic lidar simulation by leveraging the real world. In *CVPR*.
Mao, J., Wang, X., & Li, H. (2019). Interpolated convolutional networks for 3d point cloud understanding. In *ICCV*.
Mao, J., Niu, M., Bai, H., Liang, X., Xu, H., & Xu, C. (2021a). Pyramid r-cnn: Towards better performance and adaptability for 3d object detection. In *ICCV*.
Mao, J., Niu, M., Jiang, C., Liang, H., Chen, J., Liang, X., Li, Y., Ye, C., Zhang, W., & Li, Z., et al. (2021b). One million scenes for autonomous driving: Once dataset. In *NeurIPS*.
Mao, J., Xue, Y., Niu, M., Bai, H., Feng, J., Liang, X., Xu, H., & Xu, C. (2021c). Voxel transformer for 3d object detection. In *ICCV*.
Mayer, N., Ilg, E., Hausser, P., Fischer, P., Cremers, D., Dosovitskiy, A., & Brox, T. (2016). A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation. In *CVPR*.
Meng, Q., Wang, W., Zhou, T., Shen, J., Gool, L. V., & Dai, D. (2020). Weakly supervised 3d object detection from lidar point cloud. In *ECCV*.
Meng, Q., Wang, W., Zhou, T., Shen, J., Jia, Y., & Van Gool, L. (2021). Towards a weakly supervised framework for 3d point cloud object detection and annotation. *IEEE T-PAMI*.
Meyer, G. P., Charland, J., Hegde, D., Laddha, A., & Vallespi-Gonzalez, C. (2019a). Sensor fusion for joint 3d object detection and semantic segmentation. In *CVPRW*.
Meyer, G. P., Laddha, A., Kee, E., Vallespi-Gonzalez, C., & Wellington, C. K. (2019b). Lasernet: An efficient probabilistic 3d object detector for autonomous driving. In *CVPR*.
Meyer, G. P., Charland, J., Pandey, S., Laddha, A., Gautam, S., Vallespi-Gonzalez, C., & Wellington, C. K. (2020). Laserflow: Efficient and probabilistic object detection and motion forecasting. *IEEE RA-L*.
Meyer, M., Kuschk, G., & Tomforde, S. (2021). Graph convolutional networks for 3d object detection on radar data. In *ICCV*.
Miao, Z., Chen, J., Pan, H., Zhang, R., Liu, K., Hao, P., Zhu, J., Wang, Y., & Zhan, X. (2021). Pvgnet: A bottom-up one-stage 3d object detector with integrated multi-level features. In *CVPR*.
Misra, I., Girdhar, R., & Joulin, A. (2021). An end-to-end transformer model for 3d object detection. In *ICCV*.
Mousavian, A., Anguelov, D., Flynn, J., & Kosecka, J. (2017). 3d bounding box estimation using deep learning and geometry. In *CVPR*.
Nabati, R., & Qi, H. (2019). Rrpn: Radar region proposal network for object detection in autonomous vehicles. In *ICIP*.
Nabati, R., & Qi, H. (2021). Centerfusion: Center-based radar and camera fusion for 3d object detection. In *WACV*.
Naiden, A., Paunescu, V., Kim, G., Jeon, B., & Leordeanu, M. (2019). Shift r-cnn: Deep monocular 3d object detection with closed-form geometric constraints. In *ICIP*.
Najibi, M., Lai, G., Kundu, A., Lu, Z., Rathod, V., Funkhouser, T., Pantofaru, C., Ross, D., Davis, L. S., & Fathi, A. (2020). Dops: Learning to detect 3d objects and predict their 3d shapes. In *CVPR*.
Nakashima, K., & Kurazume, R. (2021). Learning to drop points for lidar scan synthesis. In *IROS*.
Ngiam, J., Caine, B., Han, W., Yang, B., Chai, Y., Sun, P., Zhou, Y., Yi, X., Alsharif, O., & Nguyen, P., et al. (2019). Starnet: Targeted computation for object detection in point clouds. arXiv preprint arXiv:1908.11069
Noh, J., Lee, S., & Ham, B. (2021). Hvpr: Hybrid voxel-point representation for single-stage 3d object detection. In *CVPR*.
Paigwar, A., Erkent, O., Wolf, C., & Laugier, C. (2019). Attentional pointnet for 3d-object detection in point clouds. In *CVPRW*.
Paigwar, A., Sierra-Gonzalez, D., Erkent, Ö., & Laugier, C. (2021). Frustum-pointpillars: A multi-stage approach for 3d object detection using rgb camera and lidar. In *ICCV*.

Palffy, A., Pool, E., Baratam, S., Kooij, J. F., & Gavrila, D. M. (2022). Multi-class road user detection with 3+ 1d radar in the view-of-delft dataset. *IEEE RA-L*.
Pan, X., Xia, Z., Song, S., Li, L. E., & Huang, G. (2021). 3d object detection with pointformer. In *CVPR*.
Pang, S., Morris, D., & Radha, H. (2020). Clocs: Camera-lidar object candidates fusion for 3d object detection. In *IROS*.
Pang, S., Morris, D., & Radha, H. (2022). Fast-clocs: Fast camera-lidar object candidates fusion for 3d object detection. In *WACV*.
Park, D., Ambrus, R., Guizilini, V., Li, J., & Gaidon, A. (2021). Is pseudo-lidar needed for monocular 3d object detection? In *ICCV*.
Park, J., Xu, C., Yang, S., Keutzer, K., Kitani, K., Tomizuka, M., & Zhan, W. (2022). Time will tell: New outlooks and a baseline for temporal multi-view 3d object detection. arXiv preprint arXiv:2210.02443
Park, J. J., Florence, P., Straub, J., Newcombe, R., & Lovegrove, S. (2019). Deepsdf: Learning continuous signed distance functions for shape representation. In *CVPR*.
Patil, A., Malla, S., Gang, H., & Chen, Y.-T. (2019). The h3d dataset for full-surround 3d multi-object detection and tracking in crowded urban scenes. In *ICRA*.
Peng, L., Yan, S., Wu, B., Yang, Z., He, X., & Cai, D. (2021). Weakm3d: Towards weakly supervised monocular 3d object detection. In *ICLR*.
Peng, W., Pan, H., Liu, H., & Sun, Y. (2020). Ida-3d: Instance-depth-aware 3d object detection from stereo vision for autonomous driving. In *CVPR*.
Peng, X., Zhu, X., Wang, T., & Ma, Y. (2022). Side: Center-based stereo 3d detector with structure-aware instance depth estimation. In *WACV*.
Pham, Q.-H., Sevestre, P., Pahwa, R. S., Zhan, H., Pang, C. H., Chen, Y., Mustafa, A., Chandrasekhar, V., & Lin, J. (2020). A* 3d dataset: Towards autonomous driving in challenging environments. In *ICRA*.
Philion, J., & Fidler, S. (2020). Lift, splat, shoot: Encoding images from arbitrary camera rigs by implicitly unprojecting to 3d. In *ECCV*.
Philion, J., Kar, A., & Fidler, S. (2020). Learning to evaluate perception models using planner-centric metrics. In *CVPR*.
Phillips, J., Martinez, J., Bârsan, I. A., Casas, S., Sadat, A., & Urtasun, R. (2021). Deep multi-task learning for joint localization, perception, and prediction. In *CVPR*.
Piergiovanni, A., Casser, V., Ryoo, M. S., & Angelova, A. (2021). 4d-net for learned multi-modal alignment. In *ICCV*.
Pon, A. D., Ku, J., Li, C., & Waslander, S. L. (2020). Object-centric stereo matching for 3d object detection. In *ICRA*.
Qi, C. R., Su, H., Mo, K., & Guibas, L. J. (2017a). Pointnet: Deep learning on point sets for 3d classification and segmentation. In *CVPR*.
Qi, C. R., Yi, L., Su, H., & Guibas, L. J. (2017b). Pointnet++ deep hierarchical feature learning on point sets in a metric space. In *NeurIPS*.
Qi, C. R., Liu, W., Wu, C., Su, H., & Guibas, L. J. (2018). Frustum pointnets for 3d object detection from rgb-d data. In *CVPR*.
Qi, C. R., Litany, O., He, K., & Guibas, L. J. (2019). Deep hough voting for 3d object detection in point clouds. In *ICCV*.
Qi, C. R., Chen, X., Litany, O., & Guibas, L. J. (2020). Imvotenet: Boosting 3d object detection in point clouds with image votes. In *CVPR*.
Qi, C. R., Zhou, Y., Najibi, M., Sun, P., Vo, K., Deng, B., & Anguelov, D. (2021). Offboard 3d object detection from point cloud sequences. In *CVPR*.
Qian, K., Zhu, S., Zhang, X., & Li, L. E. (2021a). Robust multimodal vehicle detection in foggy weather using complementary lidar and radar signals. In *CVPR*.
Qian, R., Garg, D., Wang, Y., You, Y., Belongie, S., Hariharan, B., Campbell, M., Weinberger, K. Q., & Chao, W.-L. (2020). End-to-end pseudo-lidar for image-based 3d object detection. In *CVPR*.
Qian, R., Lai, X., & Li, X. (2021b). 3d object detection for autonomous driving: A survey. *Pattern Recognition*.
Qin, Z., Wang, J., & Lu, Y. (2019a). Monogrnet: A geometric reasoning network for monocular 3d object localization. In *AAAI*.
Qin, Z., Wang, J., & Lu, Y. (2019b). Triangulation learning network: from monocular to stereo 3d object detection. In *CVPR*.
Qin, Z., Wang, J., & Lu, Y. (2020). Weakly supervised 3d object detection from point clouds. In *ACM Multimedia*.
Rapoport-Lavie, M., & Raviv, D. (2021). It's all around you: Range-guided cylindrical network for 3d object detection. In *ICCV*.
Reading, C., Harakeh, A., Chae, J., & Waslander, S. L. (2021). Categorical depth distribution network for monocular 3d object detection. In *CVPR*.
Ren, S., He, K., Girshick, R., & Sun, J. (2015a). Faster r-cnn: Towards real-time object detection with region proposal networks. *NeurIPS*.
Ren, S., He, K., Girshick, R., & Sun, J. (2015b). Faster r-cnn: Towards real-time object detection with region proposal networks. *NeurIPS*.
Rist, C. B., Enzweiler, M., & Gavrila, D. M. (2019). Cross-sensor deep domain adaptation for lidar detection and segmentation. In *IV*.
Roddick, T., Kendall, A., & Cipolla, R. (2019). Orthographic feature transform for monocular 3d object detection. In *BMVC*.
Ronneberger, O., Fischer, P., & Brox, T. (2015). U-net: Convolutional networks for biomedical image segmentation. In *MICCAI*.
Ros, G., Sellart, L., Materzynska, J., Vazquez, D., & Lopez, A. M. (2016). The synthia dataset: A large collection of synthetic images for semantic segmentation of urban scenes. In *CVPR*.
Rubino, C., Crocco, M., & Del Bue, A. (2017). 3d object localisation from multi-view image detections. *IEEE T-PAMI*.
Rukhovich, D., Vorontsova, A., & Konushin, A. (2022). Imvoxelnet: Image to voxels projection for monocular and multi-view general-purpose 3d object detection. In *WACV*.
Sadat, A., Casas, S., Ren, M., Wu, X., Dhawan, P., & Urtasun, R. (2020). Perceive, predict, and plan: Safe motion planning through interpretable semantic representations. In *ECCV*.
Saleh, K., Abobakr, A., Attia, M., Iskander, J., Nahavandi, D., Hossny, M., & Nahvandi, S. (2019). Domain adaptation for vehicle detection from bird's eye view lidar point cloud data. In *ICCVW*.
Saltori, C., Lathuiliére, S., Sebe, N., Ricci, E., & Galasso, F. (2020). Sf-uda 3d: Source-free unsupervised domain adaptation for lidar-based 3d object detection. In *3DV*.
Shah, S., Dey, D., Lovett, C., & Kapoor, A. (2018). Airsim: High-fidelity visual and physical simulation for autonomous vehicles. In Field and service robotics.
Sheng, H., Cai, S., Liu, Y., Deng, B., Huang, J., Hua, X.-S., & Zhao, M.-J. (2021). Improving 3d object detection with channel-wise transformer. In *ICCV*.
Shi, G., Li, R., & Ma, C. (2022). Pillarnet: Real-time and high-performance pillar-based 3d object detection. In *ECCV*.
Shi, S., Wang, X., & Li, H. (2019). Pointrcnn: 3d object proposal generation and detection from point cloud. In *CVPR*.
Shi, S., Guo, C., Jiang, L., Wang, Z., Shi, J., Wang, X., & Li, H. (2020a). Pv-rcnn: Point-voxel feature set abstraction for 3d object detection. In *CVPR*.
Shi, S., Wang, Z., Shi, J., Wang, X., & Li, H. (2020b). From points to parts: 3d object detection from point cloud with part-aware and part-aggregation network. *IEEE T-PAMI*.
Shi, S., Jiang, L., Deng, J., Wang, Z., Guo, C., Shi, J., Wang, X., & Li, H. (2021a). Pv-rcnn++: Point-voxel feature set abstraction with local vector representation for 3d object detection. arXiv preprint arXiv:2102.00463
Shi, W., & Rajkumar, R. (2020). Point-gnn: Graph neural network for 3d object detection in a point cloud. In *CVPR*.
Shi, X., Chen, Z., & Kim, T.-K. (2020c). Distance-normalized unified representation for monocular 3d object detection. In *ECCV*.

Shi, X., Ye, Q., Chen, X., Chen, C., Chen, Z., & Kim, T.-K. (2021b). Geometry-based distance decomposition for monocular 3d object detection. In *ICCV*.
Shin, K., Kwon, Y. P., & Tomizuka, M. (2019). Roarnet: A robust 3d object detection based on region approximation refinement. In *IV*.
Simon, M., Amende, K., Kraus, A., Honer, J., Samann, T., Kaulbersch, H., Milz, S., & Michael Gross, H. (2019). Complexer-yolo: Real-time 3d object detection and tracking on semantic point clouds. In *CVPRW*.
Simonelli, A., Bulo, S. R., Porzi, L., López-Antequera, M., & Kontschieder, P. (2019). Disentangling monocular 3d object detection. In *ICCV*.
Simonelli, A., Bulo, S. R., Porzi, L., Ricci, E., & Kontschieder, P. (2020). Towards generalization across depth for monocular 3d object detection. In *ECCV*.
Simony, M., Milzy, S., Amendey, K., & Gross, H.-M. (2018). Complex-yolo: An euler-region-proposal for real-time 3d object detection on point clouds. In *ECCVW*.
Sindagi, V. A., Zhou, Y., & Tuzel, O. (2019). Mvx-net: Multimodal voxelnet for 3d object detection. In *ICRA*.
Song, S., Lichtenberg, S. P., & Xiao, J. (2015). Sun rgb-d: A rgb-d scene understanding benchmark suite. In *CVPR*.
Sun, J., Cao, Y., Chen, Q. A., & Mao, Z. M. (2020a). Towards robust {LiDAR-based} perception in autonomous driving: General black-box adversarial sensor attack and countermeasures. In *USENIX security*.
Sun, J., Chen, L., Xie, Y., Zhang, S., Jiang, Q., Zhou, X., & Bao, H. (2020b). Disp r-cnn: Stereo 3d object detection via shape prior guided instance disparity estimation. In *CVPR*.
Sun, P., Kretzschmar, H., Dotiwalla, X., Chouard, A., Patnaik, V., Tsui, P., Guo, J., Zhou, Y., Chai, Y., & Caine, B., et al. (2020c). Scalability in perception for autonomous driving: Waymo open dataset. In *CVPR*.
Sun, P., Wang, W., Chai, Y., Elsayed, G., Bewley, A., Zhang, X., Sminchisescu, C., & Anguelov, D. (2021). Rsn: Range sparse net for efficient, accurate lidar 3d object detection. In *CVPR*.
Sun, P., Tan, M., Wang, W., Liu, C., Xia, F., Leng, Z., & Anguelov, D. (2022). Swformer: Sparse window transformer for 3d object detection in point clouds. In *ECCV*.
Suo, S., Regalado, S., Casas, S., & Urtasun, R. (2021). Trafficsim: Learning to simulate realistic multi-agent behaviors. In *CVPR*.
Tan, S., Wong, K., Wang, S., Manivasagam, S., Ren, M., & Urtasun, R. (2021). Scenegen: Learning to generate realistic traffic scenes. In *CVPR*.
Tang, H., Liu, Z., Zhao, S., Lin, Y., Lin, J., Wang, H., & Han, S. (2020). Searching efficient 3d architectures with sparse point-voxel convolution. In *ECCV*.
Tarvainen, A., & Valpola, H. (2017). Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. *NeurIPS*.
Tian, Z., Shen, C., Chen, H., & He, T. (2019). Fcos: Fully convolutional one-stage object detection. In *ICCV*.
Tu, J., Ren, M., Manivasagam, S., Liang, M., Yang, B., Du, R., Cheng, F., & Urtasun, R. (2020). Physically realizable adversarial examples for lidar object detection. In *CVPR*.
Tu, J., Wang, T., Wang, J., Manivasagam, S., Ren, M., & Urtasun, R. (2021). Adversarial attacks on multi-agent communication. In *ICCV*.
Tu, J., Li, H., Yan, X., Ren, M., Chen, Y., Liang, M., Bitar, E., Yumer, E., & Urtasun, R. (2022). Exploring adversarial robustness of multi-sensor perception systems in self driving. In *CoRL*.
Vadivelu, N., Ren, M., Tu, J., Wang, J., & Urtasun, R. (2021). Learning to communicate and correct pose errors. In *CoRL*.
Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. In *NeurIPS*.
Vora, S., Lang, A. H., Helou, B., & Beijbom, O. (2020). Pointpainting: Sequential fusion for 3d object detection. In *CVPR*.
Wang, C., Ma, C., Zhu, M., & Yang, X. (2021a). Pointaugmenting: Cross-modal augmentation for 3d object detection. In *CVPR*.
Wang, D. Z., & Posner, I. (2015). Voting for voting in online point cloud object detection. In *RSS*.
Wang, H., Cong, Y., Litany, O., Gao, Y., & Guibas, L. J. (2021b). 3dioumatch: Leveraging iou prediction for semi-supervised 3d object detection. In *CVPR*.
Wang, J., Lan, S., Gao, M., & Davis, L. S. (2020a). Infofocus: 3d object detection for autonomous driving with dynamic information modeling. In *ECCV*.
Wang, J., Pun, A., Tu, J., Manivasagam, S., Sadat, A., Casas, S., Ren, M., & Urtasun, R. (2021c). Advsim: Generating safety-critical scenarios for self-driving vehicles. In *CVPR*.
Wang, L., & Goldluecke, B. (2021). Sparse-pointnet: See further in autonomous vehicles. *IEEE RA-L*.
Wang, L., Du, L., Ye, X., Fu, Y., Guo, G., Xue, X., Feng, J., & Zhang, L. (2021d). Depth-conditioned dynamic message propagation for monocular 3d object detection. In *CVPR*
Wang, L., Zhang, L., Zhu, Y., Zhang, Z., He, T., Li, M., & Xue, X. (2021e). Progressive coordinate transforms for monocular 3d object detection. *NeurIPS*.
Wang, Q., Chen, J., Deng, J., & Zhang, X. (2021f). 3d-centernet: 3d object detection network for point clouds with center estimation priority. *Pattern Recognition*.
Wang, S., Suo, S., Ma, W.-C., Pokrovsky, A., & Urtasun, R. (2018). Deep parametric continuous convolutional neural networks. In *CVPR*.
Wang, T., Zhu, X., & Lin, D. (2020b). Reconfigurable voxels: A new representation for lidar-based point clouds. arXiv preprint arXiv:2004.02724
Wang, T., Zhu, X., Pang, J., & Lin, D. (2021g). Fcos3d: Fully convolutional one-stage monocular 3d object detection. In *ICCV*.
Wang, T., Xinge, Z., Pang, J., & Lin, D. (2022a). Probabilistic and geometric depth: Detecting objects in perspective. In *CoRL*.
Wang, T.-H., Manivasagam, S., Liang, M., Yang, B., Zeng, W., & Urtasun, R. (2020c). V2vnet: Vehicle-to-vehicle communication for joint perception and prediction. In *ECCV*.
Wang, X., Yin, W., Kong, T., Jiang, Y., Li, L., & Shen, C. (2020d). Task-aware monocular depth estimation for 3d object detection. In *AAAI*.
Wang, Y., & Solomon, J. M. (2021). Object dgcnn: 3d object detection using dynamic graphs. *NeurIPS*.
Wang, Y., Chao, W.-L., Garg, D., Hariharan, B., Campbell, M., & Weinberger, K. Q. (2019a). Pseudo-lidar from visual depth estimation: Bridging the gap in 3d object detection for autonomous driving. In *CVPR*.
Wang, Y., Chao, W.-L., Garg, D., Hariharan, B., Campbell, M., & Weinberger, K. Q. (2019b). Pseudo-lidar from visual depth estimation: Bridging the gap in 3d object detection for autonomous driving. In *CVPR*.
Wang, Y., Sun, Y., Liu, Z., Sarma, S. E., Bronstein, M. M., & Solomon, J. M. (2019c). Dynamic graph cnn for learning on point clouds. *ACM TOG*.
Wang, Y., Chen, X., You, Y., Li, L. E., Hariharan, B., Campbell, M., Weinberger, K. Q., & Chao, W.-L. (2020e). Train in germany, test in the usa: Making 3d object detectors generalize. In *CVPR*.
Wang, Y., Fathi, A., Kundu, A., Ross, D. A., Pantofaru, C., Funkhouser, T., & Solomon, J. (2020f). Pillar-based object detection for autonomous driving. In *ECCV*.
Wang. Y., Mao. Q., Zhu. H., Zhang, Y., Ji, J., & Zhang, Y. (2021h). Multi-modal 3d object detection in autonomous driving: a survey. arXiv preprint arXiv:2106.12735
Wang, Y., Yang, B., Hu, R., Liang, M., & Urtasun, R. (2021i). Plumenet: Efficient 3d object detection from stereo images. In *IROS*.

Wang, Y., Guizilini, V. C., Zhang, T., Wang, Y., Zhao, H., & Solomon, J. (2022b). Detr3d: 3d object detection from multi-view images via 3d-to-2d queries. In *CoRL*.

Wang, Z., & Jia, K. (2019). Frustum convnet: Sliding frustums to aggregate local point-wise features for amodal 3d object detection. In *IROS*.

Wang, Z., Ding, S., Li, Y., Fenn, J., Roychowdhury, S., Wallin, A., Martin, L., Ryvola, S., Sapiro, G., & Qiu, Q. (2021j). Cirrus: A long-range bi-pattern lidar dataset. In *ICRA*.

Wang, Z., Zhao, Z., Jin, Z., Che, Z., Tang, J., Shen, C., & Peng, Y. (2021k). Multi-stage fusion for multi-class 3d lidar detection. In *ICCVW*.

Wang, Z., Min, C., Ge, Z., Li, Y., Li, Z., Yang, H., & Huang, D. (2022c). Sts: Surround-view temporal stereo for multi-view 3d detection. arXiv preprint arXiv:2208.10145

Wei, B., Ren, M., Zeng, W., Liang, M., Yang, B., & Urtasun, R. (2021a). Perceive, attend, and drive: Learning spatial attention for safe self-driving. In *ICRA*.

Wei, Y., Su, S., Lu, J., & Zhou, J. (2021b). Fgr: Frustum-aware geometric reasoning for weakly supervised 3d vehicle detection. In *ICRA*.

Weng, X., & Kitani, K. (2019). Monocular 3d object detection with pseudo-lidar point cloud. In *ICCVW*.

Weng, X., Man, Y., Cheng, D., Park, J., O'Toole, M., Kitani, K., Wang, J., & Held, D. (2020). All-in-one drive: A large-scale comprehensive perception dataset with high-density long-range point clouds.

Wicker, M., & Kwiatkowska, M. (2019). Robustness of 3d deep learning in an adversarial setting. In *CVPR*.

Wilson, B., Qi, W., Agarwal, T., Lambert, J., Singh, J., Khandelwal, S., Pan, B., Kumar, R., Hartnett, A., & Pontes, J. K., et al. (2021). Argoverse 2: Next generation datasets for self-driving perception and forecasting. In *NeurIPS*.

Wong, K., Zhang, Q., Liang, M., Yang, B., Liao, R., Sadat, A., & Urtasun, R. (2020). Testing the safety of self-driving vehicles by simulating perception and prediction. In *ECCV*.

Wu, J., Yin, D., Chen, J., Wu, Y., Si, H., & Lin, K. (2020a). A survey on monocular 3d object detection algorithms based on deep learning. *Journal of Physics: Conference Series*.

Wu, P., Chen, S., & Metaxas, D. N. (2020b). Motionnet: Joint perception and motion prediction for autonomous driving based on bird's eye view maps. In *CVPR*.

Xiang, Y., Choi, W., Lin, Y., & Savarese, S. (2015). Data-driven 3d voxel patterns for object category recognition. In *CVPR*.

Xiang, Y., Choi, W., Lin, Y., & Savarese, S. (2017). Subcategory-aware convolutional neural networks for object proposals and detection. In *WACV*.

Xiao, P., Shao, Z., Hao, S., Zhang, Z., Chai, X., Jiao, J., Li, Z., Wu, J., Sun, K., & Jiang, K., et al. (2021). Pandaset: Advanced sensor suite dataset for autonomous driving. In *ITSC*.

Xiao, Y., Codevilla, F., Gurram, A., Urfalioglu, O., & López, A. M. (2020). Multimodal end-to-end autonomous driving. *IEEE T-ITS*.

Xie, E., Yu, Z., Zhou, D., Philion, J., Anandkumar, A., Fidler, S., Luo, P., & Alvarez, J. M. (2022). M^2bev: Multi-camera joint 3d detection and segmentation with unified birds-eye view representation. arXiv preprint arXiv:2204.05088

Xie, L., Xiang, C., Yu, Z., Xu, G., Yang, Z., Cai, D., & He, X. (2020a). Pi-rcnn: An efficient multi-sensor 3d object detector with point-based attentive cont-conv fusion module. In *AAAI*.

Xie, S., Gu, J., Guo, D., Qi, C. R., Guibas, L., & Litany, O. (2020b). Pointcontrast: Unsupervised pre-training for 3d point cloud understanding. In *ECCV*.

Xu, B., & Chen, Z. (2018). Multi-level fusion based 3d object detection from monocular images. In *CVPR*.

Xu, D., Anguelov, D., & Jain, A. (2018). Pointfusion: Deep sensor fusion for 3d bounding box estimation. In *CVPR*.

Xu, Q., Zhong, Y., & Neumann, U. (2021a). Behind the curtain: Learning occluded shapes for 3d object detection. arXiv preprint arXiv:2112.02205

Xu, Q., Zhou, Y., Wang, W., Qi, C. R., & Anguelov, D. (2021b). Spg: Unsupervised domain adaptation for 3d object detection via semantic point generation. In *ICCV*.

Xu, S., Zhou, D., Fang, J., Yin, J., Bin, Z., & Zhang, L. (2021c). Fusionpainting: Multimodal fusion with adaptive attention for 3d object detection. In *ITSC*.

Xu, Z., Zhang, W., Ye, X., Tan, X., Yang, W., Wen, S., Ding, E., Meng, A., & Huang, L. (2020). Zoomnet: Part-aware adaptive zooming neural network for 3d object detection. In *AAAI*.

Xue, Y., Mao, J., Niu, M., Xu, H., Mi, M. B., Zhang, W., Wang, X., & Wang, X. (2022). Point2seq: Detecting 3d objects as sequences. In *CVPR*.

Yan, Y., Mao, Y., & Li, B. (2018). Second: Sparsely embedded convolutional detection. *Sensors*.

Yang, B., Liang, M., & Urtasun, R. (2018a). Hdnet: Exploiting hd maps for 3d object detection. In *CoRL*.

Yang, B., Luo, W., & Urtasun, R. (2018b). Pixor: Real-time 3d object detection from point clouds. In *CVPR*.

Yang, B., Guo, R., Liang, M., Casas, S., & Urtasun, R. (2020a). Radarnet: Exploiting radar for robust perception of dynamic objects. In *ECCV*.

Yang, B., Bai, M., Liang, M., Zeng, W., & Urtasun, R. (2021a). Auto4d: Learning to label 4d objects from sequential point clouds. arXiv preprint arXiv:2101.06586

Yang, J., Shi, S., Wang, Z., Li, H., & Qi, X. (2021b). St3d: Self-training for unsupervised domain adaptation on 3d object detection. In *CVPR*.

Yang, Z., Sun, Y., Liu, S., Shen, X., & Jia, J. (2018c). Ipod: Intensive point-based object detector for point cloud. arXiv preprint arXiv:1812.05276

Yang, Z., Sun, Y., Liu, S., Shen, X., & Jia, J. (2019). Std: Sparse-to-dense 3d object detector for point cloud. In *ICCV*.

Yang, Z., Chai, Y., Anguelov, D., Zhou, Y., Sun, P., Erhan, D., Rafferty, S., & Kretzschmar, H. (2020b). Surfelgan: Synthesizing realistic sensor data for autonomous driving. In *CVPR*.

Yang, Z., Sun, Y., Liu, S., & Jia, J. (2020c). 3dssd: Point-based 3d single stage object detector. In *CVPR*.

Yang, Z., Zhou, Y., Chen, Z., & Ngiam, J. (2021c). 3d-man: 3d multi-frame attention network for object detection. In *CVPR*.

Ye, M., Xu, S., & Cao, T. (2020a). Hvnet: Hybrid voxel network for lidar based 3d object detection. In *CVPR*.

Ye, X., Du, L., Shi, Y., Li, Y., Tan, X., Feng, J., Ding, E., & Wen, S. (2020b). Monocular 3d object detection via feature domain adaptation. In *ECCV*.

Ye, Y., Chen, H., Zhang, C., Hao, X., & Zhang, Z. (2020c). Sarpnet: Shape attention regional proposal network for lidar-based 3d object detection. *Neurocomputing*.

Yi, H., Shi, S., Ding, M., Sun, J., Xu, K., Zhou, H., Wang, Z., Li, S., & Wang, G. (2020). Segvoxelnet: Exploring semantic context and depth-aware features for 3d vehicle detection from point cloud. In *ICRA*.

Yihan, Z., Wang, C., Wang, Y., Xu, H., Ye, C., Yang, Z., & Ma, C. (2021). Learning transferable features for point cloud detection via 3d contrastive co-training. *NeurIPS*.

Yin, J., Shen, J., Guan, C., Zhou, D., & Yang, R. (2020). Lidar-based online 3d video object detection with graph-based message passing and spatiotemporal transformer attention. In *CVPR*.

Yin, T., Zhou, X., & Krahenbuhl, P. (2021a). Center-based 3d object detection and tracking. In *CVPR*.

Yin, T., Zhou, X., & Krähenbühl, P. (2021b). Multimodal virtual point 3d detection. *NeurIPS*.

Yogamani, S., Hughes, C., Horgan, J., Sistu, G., Varley, P., O'Dea, D., Uricár, M., Milz, S., Simon, M., & Amende, K., et al.

(2019). Woodscape: A multi-task, multi-camera fisheye dataset for autonomous driving. In *ICCV*.
Yoo, J. H., Kim, Y., Kim, J., & Choi, J. W. (2020). 3d-cvf: Generating joint camera and lidar features using cross-view spatial feature fusion for 3d object detection. In *ECCV*.
You, Y., Wang, Y., Chao, W.-L., Garg, D., Pleiss, G., Hariharan, B., Campbell, M., & Weinberger, K. Q. (2020). Pseudo-lidar++: Accurate depth for 3d object detection in autonomous driving. In *ICLR*.
You, Y., Diaz-Ruiz, C. A., Wang, Y., Chao, W.-L., Hariharan, B., Campbell, M., & Weinberger, K. Q. (2021). Exploiting playbacks in unsupervised domain adaptation for 3d object detection. arXiv preprint arXiv:2103.14198
Yu, F., Wang, D., Shelhamer, E., & Darrell, T. (2018). Deep layer aggregation. In *CVPR*.
Yuan, Z., Song, X., Bai, L., Wang, Z., & Ouyang, W. (2021). Temporal-channel transformer for 3d lidar-based video object detection for autonomous driving. *IEEE T-CSVT*.
Yun, P., Tai, L., Wang, Y., Liu, C., & Liu, M. (2019). Focal loss in 3d object detection. *IEEE RA-L*.
Zakharov, S., Kehl, W., Bhargava, A., & Gaidon, A. (2020). Autolabeling 3d objects with differentiable rendering of sdf shape priors. In *CVPR*.
Zamanakos, G., Tsochatzidis, L., Amanatiadis, A., & Pratikakis, I. (2021). A comprehensive survey of lidar-based 3d object detection methods with deep learning for autonomous driving. *Computers and Graphics*.
Zarzar, J., Giancola, S., & Ghanem, B. (2019). Pointrgcn: Graph convolution networks for 3d vehicles detection refinement. arXiv preprint arXiv:1911.12236
Zeeshan Zia, M., Stark, M., & Schindler, K. (2014). Are cars just 3d boxes?-jointly estimating the 3d shape of multiple objects. In *CVPR*.
Zeng, W., Wang, S., Liao, R., Chen, Y., Yang, B., & Urtasun, R. (2020). Dsdnet: Deep structured self-driving network. In *ECCV*.
Zeng, Y., Hu, Y., Liu, S., Ye, J., Han, Y., Li, X., & Sun, N. (2018). Rt3d: Real-time 3-d vehicle detection in lidar point cloud for autonomous driving. *IEEE RA-L*.
Zeng, Y., Zhang, D., Wang, C., Miao, Z., Liu, T., Zhan, X., Hao, D., & Ma, C. (2022). Lift: Learning 4d lidar image fusion transformer for 3d object detection. In *CVPR*.
Zhang, W., Li, W., & Xu, D. (2021a). Srdan: Scale-aware and range-aware domain adaptation network for cross-dataset 3d object detection. In *CVPR*.
Zhang, X., Zhang, A., Sun, J., Zhu, X., Guo, Y. E., Qian, F., & Mao, Z. M. (2021b). Emp: edge-assisted multi-vehicle perception. In *MobiCom*.
Zhang, Y., Xiang, Z., Qiao, C., & Chen, S. (2019). Accurate and real-time object detection based on bird's eye view on 3d point clouds. In *3DV*.
Zhang, Y., Lu, J., & Zhou, J. (2021c). Objects are different: Flexible monocular 3d object detection. In *CVPR*.
Zhang, Y., Chen, J., & Huang, D. (2022a). Cat-det: Contrastively augmented transformer for multi-modal 3d object detection. In *CVPR*.
Zhang, Y., Zhu, Z., Zheng, W., Huang, J., Huang, G., Zhou, J., & Lu, J. (2022b). Beverse: Unified perception and prediction in birds-eye-view for vision-centric autonomous driving. arXiv preprint arXiv:2205.09743
Zhang, Z., Gao, J., Mao, J., Liu, Y., Anguelov, D., & Li, C. (2020a). Stinet: Spatio-temporal-interactive network for pedestrian detection and trajectory prediction. In *CVPR*.
Zhang, Z., Gao, J., Mao, J., Liu, Y., Anguelov, D., & Li, C. (2020b). Stinet: Spatio-temporal-interactive network for pedestrian detection and trajectory prediction. In *CVPR*.
Zhang, Z., Girdhar, R., Joulin, A., & Misra, I. (2021d). Self-supervised pretraining of 3d features on any point-cloud. In *ICCV*.
Zheng, W., Tang, W., Chen, S., Jiang, L., & Fu, C.-W. (2021a). Cia-ssd: Confident iou-aware single-stage object detector from point cloud. In *AAAI*.
Zheng, W., Tang, W., Jiang, L., & Fu, C.-W. (2021b). Se-ssd: Self-ensembling single-stage object detector from point cloud. In *CVPR*.
Zheng, W., Tang, W., Jiang, L., & Fu, C.-W. (2021c). Se-ssd: Self-ensembling single-stage object detector from point cloud. In *CVPR*.
Zhou, D., Fang, J., Song, X., Guan, C., Yin, J., Dai, Y., & Yang, R. (2019a). Iou loss for 2d/3d object detection. In *3DV*.
Zhou, D., Fang, J., Song, X., Liu, L., Yin, J., Dai, Y., Li, H., & Yang, R. (2020a). Joint 3d instance segmentation and object detection for autonomous driving. In *CVPR*.
Zhou, X., Wang, D., & Krähenbühl, P. (2019b). Objects as points. arXiv preprint arXiv:1904.07850
Zhou, X., Peng, Y., Long, C., Ren, F., & Shi, C. (2020b). Monet3d: Towards accurate monocular 3d object localization in real time. In *ICML*.
Zhou, Y., & Tuzel, O. (2018). Voxelnet: End-to-end learning for point cloud based 3d object detection. In *CVPR*.
Zhou, Y., Sun, P., Zhang, Y., Anguelov, D., Gao, J., Ouyang, T., Guo, J., Ngiam, J., & Vasudevan, V. (2020c). End-to-end multi-view fusion for 3d object detection in lidar point clouds. In *CoRL*.
Zhou, Y., He, Y., Zhu, H., Wang, C., Li, H., & Jiang, Q. (2021). Monocular 3d object detection: An extrinsic parameter free approach. In *CVPR*.
Zhu, B., Jiang, Z., Zhou, X., Li, Z., & Yu, G. (2019). Class-balanced grouping and sampling for point cloud 3d object detection. arXiv preprint arXiv:1908.09492
Zhu, J.-Y., Park, T., Isola, P., & Efros, A. A. (2017). Unpaired image-to-image translation using cycle-consistent adversarial networks. In *ICCV*.
Zhu, M., Ma, C., Ji, P., & Yang, X. (2021a). Cross-modality 3d object detection. In *WACV*.
Zhu, X., Ma, Y., Wang, T., Xu, Y., Shi, J., & Lin, D. (2020). Ssn: Shape signature networks for multi-class object detection from point clouds. In *ECCV*.
Zhu, Y., Miao, C., Zheng, T., Hajiaghajani, F., Su, L., & Qiao, C. (2021b). Can we use arbitrary objects to attack lidar perception in autonomous driving? In *ACM SIGSAC*.
Zou, Z., Ye, X., Du, L., Cheng, X., Tan, X., Zhang, L., Feng, J., Xue, X., & Ding, E. (2021). The devil is in the task: Exploiting reciprocal appearance-localization features for monocular 3d object detection. In *ICCV*.