Journal of Artificial Intelligence Research 82 (2025) 973-1015 Submitted 07/2024; published 02/2025

A New Literature Review of 3D Object Detection on Autonomous Driving

Peng Zhang 52205901027@stu.ecnu.edu.cn
Xin Li *(Corresponding author)* 51194506020@stu.ecnu.edu.cn
Xin Lin *(Corresponding author)* xlin@cs.ecnu.edu.cn
Liang He lhe@cs.ecnu.edu.cn
East China Normal University, China

Abstract

In recent years, the realm of computer vision has experienced a significant surge in the importance of 3D object detection, especially in the context of autonomous driving. The capability to precisely identify the locations, dimensions, and types of key 3D objects surrounding an autonomous vehicle is crucial, rendering 3D object detection a vital component of any advanced perception system. This review delivers an extensive overview of the emerging technologies in 3D object detection tailored for autonomous vehicles. It encompasses a thorough examination, evaluation, and integration of the current research landscape in this domain, staying up-to-date with the latest advancements in 3D object detection and suggesting prospective avenues for future research. Our survey begins by clarifying the principles of 3D object detection and addressing its present challenges in the 3D domain. We then introduce three distinct taxonomies: camera-based, point cloud-based, and multi-modality-based approaches, providing a comprehensive classification of contemporary 3D object detection methodologies from various angles. Diverging from previous reviews, this paper also highlights and scrutinizes common issues and solutions for specific scenarios (such as pedestrian detection, lane lines, roadside cameras, and weather conditions) in object detection. Furthermore, we conduct an in-depth analysis and comparison of different classifications and methods, utilizing various datasets and experimental outcomes. Conclusively, we suggest several potential research directions, offering valuable insights for the ongoing evolution of 3D object detection technology. This review aims to serve as a comprehensive resource for researchers and practitioners in the field, guiding future innovations in 3D object detection for autonomous driving.

1. Introduction

The emergence of autonomous driving technology heralds a transformative era in the automotive industry, fundamentally altering the interaction between vehicles and their environment. Central to this technological marvel is the role of object detection, a multifaceted domain that includes multi-object tracking, target segmentation, and multi-camera detection. Within this spectrum, the significance of 3D object detection is paramount, especially for its role in comprehending the environment in a three-dimensional context, which is vital for the functionality of autonomous vehicles. 3D object detection, a cornerstone of autonomous driving, is a complex and nuanced technology. It equips autonomous vehicles with the capability to identify and pinpoint the location of various objects such as other vehicles, pedestrians, and obstacles within their immediate surroundings. This capability is realized through the amalgamation and sophisticated interpretation of data derived from an array of sensors, including, but not limited to, LiDAR, cameras, and radar systems. The intricacy and significance of 3D object detection in the context of autonomous driving are profound, as they play a direct role in influencing the vehicle's decision-making processes and ensuring safety. Initially, the focus was predominantly on enhancing the accuracy and reliability of detection in clear weather conditions.

However, recent developments have expanded the scope to include challenging scenarios such as varying weather conditions, diverse lighting environments, and complex urban landscapes. This expansion necessitates the integration of more robust and adaptive algorithms capable of handling the dynamic and often unpredictable nature of real-world driving scenarios. Moreover, the field has seen a shift towards the integration of multi-modal sensor data, combining the strengths of various sensing technologies to achieve more accurate and reliable detection. For instance, the fusion of LiDAR and camera data leverages the high-resolution color information from cameras and the precise distance measurements from LiDAR, leading to a more comprehensive understanding of the vehicle's surroundings. This multi-modal approach not only enhances object detection capabilities but also significantly improves the system's ability to function reliably under a wide range of conditions. In light of these advancements and challenges, this paper aims to provide a comprehensive overview of the current state and future directions of 3D object detection in autonomous driving. We delve into the latest technological developments, explore the integration of multi-modal sensor data, and discuss the implications of these technologies in terms of safety, ethics, and regulation. Our goal is to present a holistic view of 3D object detection, highlighting its critical role in the advancement of autonomous driving technologies and its potential to reshape our transportation systems.

The organization of the paper is structured to guide the reader through the various facets of 3D object detection in autonomous driving. We begin by exploring the theoretical foundations and technical aspects of 3D object detection, followed by an in-depth analysis of the latest advancements and innovations in the field. We then examine the integration of multi-modal sensor data and its implications for the accuracy and reliability of detection systems. Finally, we discuss the challenges and future directions of 3D object detection and conclude with a summary of our findings and suggestions for future research directions in this rapidly evolving field.

The key contributions of this work are as follows:

- This review highlights research breakthroughs in 3D object detection over the last decade. We have carefully curated over 200 methods and studies from a comprehensive collection of more than 2000 papers, covering camera-based, LiDAR-based, radar-based, and multi-modal detection methods.

- Our analysis offers a detailed examination of the critical elements of 3D object detection, encompassing datasets, metrics, and methodologies. We introduce new classifications on research methodologies to provide clarity in the rapidly advancing field.

- The paper explores diverse environments and novel detection categories, examines intricate problem-solving approaches in 3D object detection, and suggests potential directions for future research.

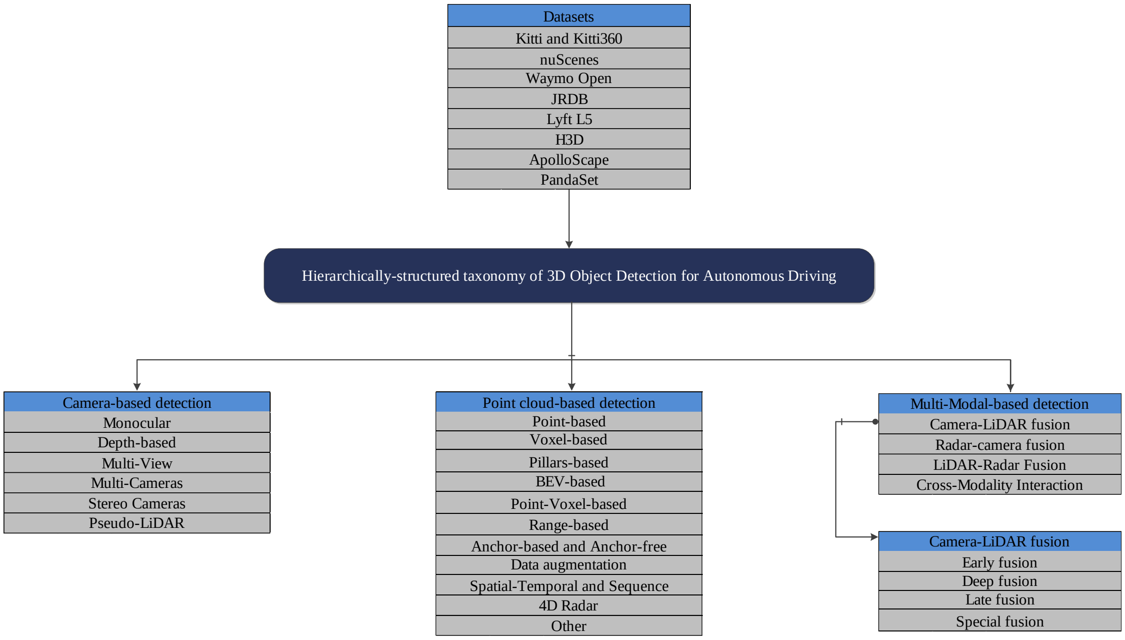

Figure 1: 3D Object Detection general structure figure.

Despite these numerous studies (Lu et al., 2022a; Ma et al., 2022; Brenner et al., 2023; Wang & Kitani, 2023; Elharrouss et al., 2023; Wang et al., 2023b; Huang et al., 2022a; Yao et al., 2023), there has been a lack of comprehensive reviews that explore 3D object detection in its entirety. This paper aims to fill that gap by providing the first exhaustive and up-to-date review of 3D object detection. It delves into the theoretical foundations, experimental aspects, and future developmental trajectories of this field. As shown in Figure 1, we provide a clear visual representation of the paper's structure, which outlines the main types, methods, and datasets related to 3D object detection. The organization of the paper is as follows:

Initially, we explore the background of 3D object detection in Section 2. Subsequently, we conduct a comprehensive review and analysis of various types and approaches to 3D object detection in Section 3. In Section 4, we will conduct a detailed exploration of the comparative results and conclusions drawn from the experimental evaluation of 3D object detection. Concluding the paper, Section 5 presents a conclusion and provides insightful suggestions for future research directions in the field of 3D object detection.

2. Background

3D object detection serves as a critical foundation for autonomous driving systems. It currently faces several challenges that need resolution. And what are the key sensors and datasets involved in this process? Let's begin with a more intuitive understanding of this concept.

2.1 Challenges of 3D Object Detection

Despite significant advancements in 3D object detection research, practical applications continue to face numerous challenges. These include issues related to object occlusion, truncation, detection of small targets, crowded scenes, adverse weather conditions, and detecting safety. Furthermore, the majority of existing methods depend heavily on surface texture or structural features of objects, which can lead to confusion or even failure in detection. Lastly, striking a balance between algorithm efficiency and accuracy remains a substantial hurdle. Addressing these complex issues forms the crux of our ongoing research endeavors.

Here are some prevalent challenges in 3D object detection:

- **Occlusion (Xu et al., 2022a):** This is categorized into two scenarios: objects occluding each other and objects being occluded by the background.

- **Truncation:** In some instances, objects are truncated by the image frame, resulting in only a portion of the objects being displayed.

- **Small and Distant Objects:** Similar to distant targets, the sparsity of the point cloud can result in missed detection.

- **Crowded Scenes (Zheng et al., 2022a):** The presence of multiple targets nearby can lead to missed detection and also necessitates higher hardware requirements.

- **Adverse Weather Conditions (Paek et al., 2022a):** Special weather phenomena like snow reflections can increase the rate of false detection, while scattering can reduce radar visibility.

2.2 Sensors

Self-driving cars typically incorporate five main types of sensors: **Camera**, **Long-range radar**, **Short-range/Medium-range radar**, **LiDAR**, and **Ultrasonic**. These sensors can be further categorized into passive sensors (like cameras) and active sensors (such as LiDAR, radar, and ultrasonic transceivers). In this article, we will focus on introducing and analyzing the three primary sensors used in self-driving cars: the camera, radar, and LiDAR. We will also explore their multi-modality applications.

Each sensor type possesses its unique strengths and can often complement one another. For instance, cameras, being high-resolution and cost-effective sensors, are limited by their lack of depth information and sensitivity to light conditions. On the other hand, LiDAR points, despite their ability to provide three-dimensional spatial data about the surrounding environment, are sometimes constrained by the capture of sparse points and come with a relatively high cost. Radar, operating on a longer radio band, delivers reliable measurements even under adverse weather conditions. In particular, 4D radar stands out as a robust sensor capable of withstanding harsh weather conditions. However, the existing radar datasets are relatively small in size compared to those of cameras and LiDAR, leaving the potential of 4D imaging radars largely untapped and unexplored.

To leverage the complementary features of these sensors and enhance overall performance, an increasing number of methods are being developed to design fusion networks

Dataset	Year	LiDAR scans	Image	Classes	Stereo	Temporal	LiDAR	Night/Rain
KITTI	2012	15K	15K	8	Yes	Yes	Yes	No/No
NuScenes	2020	400K	1.4M	23	No	Yes	Yes	Yes/Yes
Waymo open	2020	230K	1.0M	4	No	Yes	Yes	Yes/Yes
Argoverse	2019	44K	490K	15	Yes	Yes	Yes	Yes/Yes
Argoverse v2	2021	–	–	30	Yes	Yes	Yes	Yes/Yes
Lyft L5	2019	46K	323K	9	No	Yes	Yes	No/No
H3D	2019	27K	83K	8	No	Yes	Yes	No/No
ApolloScape	2019	20K	144K	6	Yes	Yes	Yes	–
PandaSet	2021	8.2K	49K	37	–	–	Yes	Yes/Yes

Table 1: Autonomous driving datasets that are used for 3D object detection.

that integrate images with point clouds. These methods have demonstrated superior performance in 3D object detection tasks compared to methods that rely on a single sensor. We will delve deeper into this topic in Section 3.

2.3 Datasets

The primary datasets used for traditional 2D object detection and tracking include MOT17, BDD100K, among others. The leading 3D object detection datasets relevant to autonomous driving currently include KITTI (Geiger et al., 2013), nuScenes (Caesar et al., 2020), Waymo Open (Sun et al., 2020), Argoverse (Chang et al., 2019), ApolloScape (Huang et al., 2019), PandaSet (Xiao et al., 2021), H3D (Patil et al., 2019), and more. We have compiled the most recent datasets for 3D object detection, as detailed in Table 1.

KITTI Dataset KITTI dataset provides annotations for eight distinct classes, each of which is further categorized into ”easy”, ”moderate”, and ”hard” cases. It offers comprehensive RGB and point cloud data (in Bin format), along with a range of tools, calibration data, and labeling information. To further advance multi-modal detection methods in autonomous driving, the KITTI development team introduced KITTI360. This dataset contains richer sensor information and 360-degree annotations.

nuScenes Dataset nuScenes comprises 700 training scenes, 150 validation scenes, and 150 testing scenes. In addition to this, it provides annotations for object-level attributes such as visibility, activity, pose, and more. It includes a large volume of RGB and point-cloud data (in PCD format). Given the substantial size of the complete nuScenes dataset, users often prefer the nuScenes-mini dataset.

Waymo Open Dataset Waymo includes a total of 798 training scenes and 202 validation scenes, each with 2D and 3D annotated labels. The annotations in the Waymo open dataset categorize objects into four groups: cars, pedestrians, cyclists, and signs. Many recent research papers have started to utilize it for their studies and evaluations.

ApolloScape Dataset ApolloScape supports a variety of autonomous driving tasks such as scene parsing, lane segmentation, trajectory prediction, object detection, and tracking. It contains over 140,000 annotated images with lane line annotations. For 3D object detection, apolloScape provides over 6,000 point cloud frames with annotated 3D bounding boxes.

H3D Dataset H3D is a comprehensive 3D object detection and tracking dataset, specifically designed for crowded urban traffic scenes. The dataset comprises over 27,000 frames in 160 scenes, with more than 1 million objects. For evaluation purposes, H3D follows a protocol similar to KITTI, with a 0.5 IOU threshold for cars and a 0.25 IOU threshold for pedestrians.

Argoverse Dataset Argoverse boasts extensive semantic annotations for maps, providing detailed insights into road infrastructure and traffic regulations. In addition to this, Argoverse stands out for its inclusion of high-definition (HD) maps, facilitating automatic map creation.

2.4 Evaluation Metrics

Just as in 2D object detection, the Average Precision (AP) and mean Average Precision (mAP) are the primary evaluation metrics used in 3D object detection. In this context, we'll first review the original AP metric, followed by an introduction to its variants. These variants are commonly adopted in widely used benchmarks, including the KITTI 3D, nuScenes, and Waymo open datasets.

2.4.1 AP and mAP

Before delving into the AP and mAP, it's crucial to grasp a few fundamental concepts. Intersection over Union (IOU) is a metric that measures the overlap between predicted values and the ground truth. Additionally, precision and recall are two vital concepts. Precision can be viewed as a function of recall, denoted as $p(r)$. To mitigate the effect of "wiggles" in the precision-recall curve, interpolated precision values are utilized to calculate the AP. The formula for AP is as follows:

$$\mathrm{AP} = \frac{1}{|\mathbb{R}|} \sum_{r \in \mathbb{R}} p_{interp}(r). \tag{1}$$

The calculation of AP is specific to one category. Once the AP is obtained, the calculation of mAP becomes straightforward. It involves calculating the AP for all categories and then taking the average. The mAP metric assesses the proficiency of the trained model in detecting all categories. Assuming that there are K categories and it is greater than 1, the formula for mAP is as follows:

$$mAP = \frac{\sum_{i=1}^{K} AP_i}{K} \tag{2}$$

2.4.2 Datasets Specific Metrics

KITTI Benchmark The KITTI 3D benchmark uses AP and mAP as primary metrics. However, for object orientation prediction in object detection tasks, KITTI introduces a unique approach known as Average Orientation Similarity (AOS). This metric, which ranges from 0-100%, is used to gauge the similarity between the detected object's orientation and the ground truth. In essence, AOS measures how closely the predicted direction of an object matches the actual direction. The formula for AOS is as follows:

$$\mathrm{AOS} = \frac{1}{|\mathbb{R}|} \sum_{r \in \mathbb{R}} \max_{r^{\prime}: r^{\prime} \geq r} s\left(r^{\prime}\right) \tag{3}$$

Here, r is the object detection recall. Define the directional similarity $s(r) \in [0, 1]$:

$$s(r) = \frac{1}{|D(r)|} \sum_{i \in D(r)} \frac{1 + \cos \Delta_\theta^{(i)}}{2} \delta_i \tag{4}$$

where $D(r)$ represents the set of all object detections at recall r, while $\Delta_\theta^{(i)}$ denotes the angular difference between the estimated and true directions of the detection.

nuScenes Benchmark Contrary to KITTI, nuScenes employs several metrics beyond AP and mAP to evaluate the performance of True Positive (TP) detections. These include five TP metrics designed as positive scalars: Average Translation Error (ATE), Average Scale Error (ASE), Average Orientation Error (AOE), Average Velocity Error (AVE), and Average Attribute Error (AAE). ATE is a measure of the Euclidean distance between 2D centers in meters. ASE is the 3D IoU between the predicted and real labels after aligning orientation and position, calculated as (1 - IoU). AOE is the difference in minimum yaw angle between predicted and real tags in radians. AVE is the absolute velocity error as the L2 norm of the 2D velocity difference in meters per second. AAE is defined as 1 minus the attribute classification accuracy, i.e., 1 - accuracy. Furthermore, the final metric, known as the nuScenes Detection Score (NDS), is a weighted sum of mAP and these errors. This provides a more comprehensive assessment of detection performance.

Waymo Open Benchmark In order to evaluate the performance of the algorithm, WOD employs both AP and an AP-weighted metric (APH). The latter takes into account heading errors, thus providing a more comprehensive assessment of the algorithm's capabilities. In contrast to the approach taken with the KITTI dataset, where AP is calculated using 11-point and 40-point interpolation methods, the WOD algorithm employs a different methodology based on the area under the precision-recall (P-R) curve. Furthermore, WOD introduces an APH, which considers heading errors to provide a more comprehensive evaluation of the algorithm's performance. The formula for APH is presented below:

$$APH = \frac{1}{C} \sum_{i=1}^{C} \frac{\mathrm{TP}_{i,\mathrm{heading}}}{\mathrm{TP}_{i,\mathrm{heading}} + \mathrm{FP}_{i,\mathrm{heading}} + \mathrm{FN}_{i,\mathrm{heading}}} \cdot \mathrm{AP}_i \tag{5}$$

where C is the total number of object categories, and $\mathrm{TP}_{i,\mathrm{heading}}$, $\mathrm{FP}_{i,\mathrm{heading}}$, and $\mathrm{FN}_{i,\mathrm{heading}}$ denote the true instances where the prediction of the category i is towards accuracy, respectively, false positive examples and false negative examples. The AP_i represents the average accuracy of each category, calculated from the area under the P-R curve.

Furthermore, the WOD defines two difficulty levels, designated as LEVEL 1 (L1) and LEVEL 2 (L2), to distinguish between detection targets with varying degrees of complexity. Initially, all targets lacking LiDAR points are disregarded, and targets classified as LEVEL 2 are deemed more challenging, typically those with a maximum of 5 LiDAR points or those manually designated as difficult (hard). LEVEL 1 encompasses the remaining targets. Furthermore, WOD establishes distinct evaluation ranges contingent on the distance between the object and the sensor, encompassing three distance intervals: 0-30 metres, 30-50 metres, and over 50 metres.

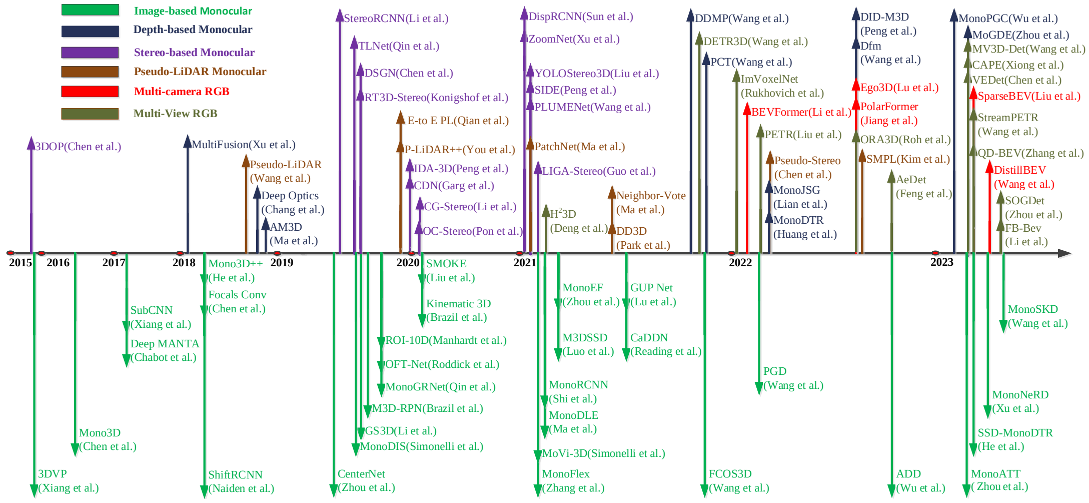

Figure 2: The diagram for sequential classification in 3D object detection (Camera-based) meticulously delineates the primary models and methods. For camera-based methods, the classification includes a variety of techniques such as monocular(below the timeline), depth-based, multi-camera, multi-view, stereo, and pseudo-LiDAR classifications.

3. 3D Object Detection Approaches

In this section, we delve into an exploration and examination of typical sensors used in the field of autonomous driving. We place a special emphasis on 3D object detection methods that rely on a variety of sensor types. These include, but are not limited to, cameras, LiDAR, radar, and multi-modal sensors. Additionally, we provide an overview of other classifications and strategic approaches that can be effectively utilized for 3D object detection.

3.1 Types of 3D Object Detection

Current classification methods in autonomous driving are typically based on sensor data and can be categorized into camera-based, point-cloud-based, and multi-modal-based classifications (Qian et al., 2022). Image and depth maps are the most common types of data used in object detection. Datasets in RGB-D format include Pascal VOC (Hoiem et al., 2009), COCO (Veit et al., 2016), and ImageNet (Deng et al., 2009). In recent years, LiDAR-based 3D object detection has become a popular research topic in the field of 3D computer vision. Relevant datasets include KITTI (Geiger et al., 2013), nuScenes (Caesar et al., 2020), and waymo open (Sun et al., 2020). Radar data also plays a crucial role in object detection. However, radar alone does not provide sufficient information for detection and classification, making the fusion of different types of data essential.

A thorough review of numerous conferences and journals in recent years has led to the chronological classification of 3D object detection methods. Figure 2 refers to the camera-based model and Figure 3 refers to the point-cloud-based model, while Figure 4

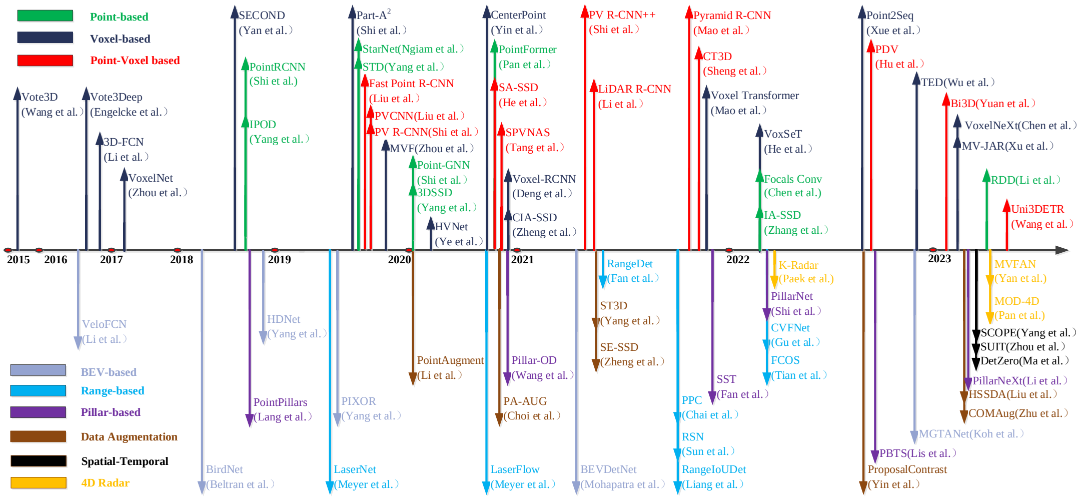

Figure 3: The diagram for sequential classification in 3D object detection (Point Cloud-based) meticulously delineates the primary models and methods. Point cloud-based methods comprise classifications like point-based, voxel-based, Bird's Eye View (BEV), Pillar-based, Range-based, Spatial-Temporal, 4D Radar, and data augmentation, among others.

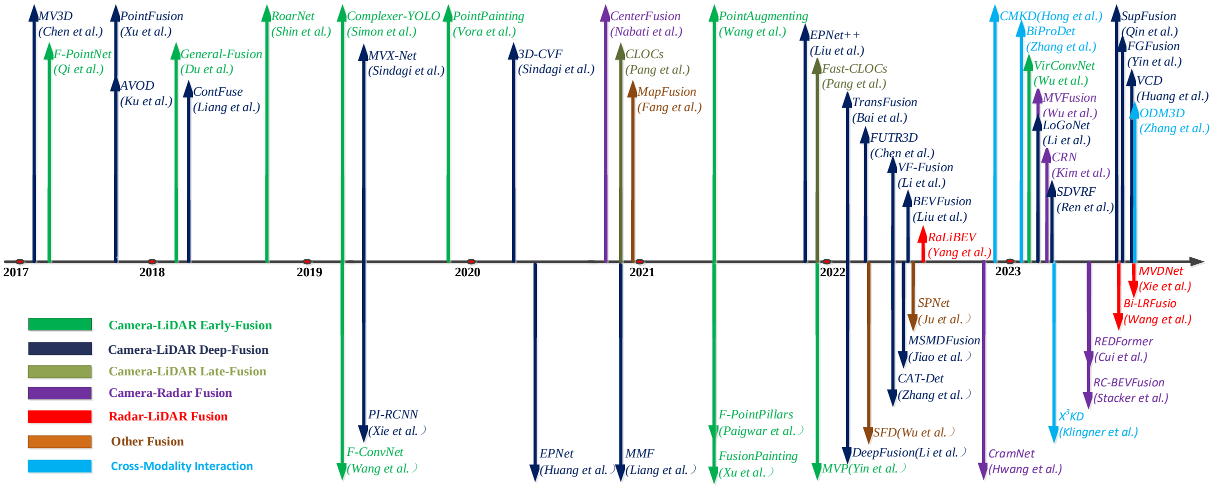

Figure 4: The sequential classification diagram for 3D object detection (multi-modal and cross-modality interaction) systematically categorizes the primary models and methods. Multi-modality primarily encompasses camera-LiDAR fusion, camera-radar fusion, radar-LiDAR fusion, and cross-modality interaction. The camera-LiDAR fusion category includes early fusion, deep fusion, late fusion, and other fusion methods.

represents multi-modality fusion and cross-modality interaction. The examination of Figure 2, Figure 3, and Figure 4 reveals that the majority of models and methods are based on cameras, radar and LiDAR, while there are relatively few multi-modal approaches. Recent research trends include cameras-based methods (such as monocular, multi-camera, multi-view, BEV-based camera, and pseudo-LiDAR techniques), point cloud-based methods (including voxel-based, point-voxel, BEV, 4d radar, and spatial-temporal and sequence, and data augmentation techniques), and multi-modality-based methods (like deep-fusion and cross-modality interaction). These findings highlight the diverse approaches being explored in this field.

3.2 Camera-based Approaches

Camera-based 3D object detection plays a pivotal role in numerous tasks, including depth estimation, 3D object detection, 3D multi-object tracking, and 3D segmentation. However, the inherent lack of depth information in images poses a challenge as it does not provide sufficient 3D information. Despite this limitation, camera-based algorithms are cost-effective and can bolster the robustness of the system. The array of camera-based methods encompasses monocular, multi-camera, multi-view, stereo-camera, and pseudo-LiDAR techniques. Beyond these, there are also BEV-based Camera methods and the increasingly popular BEVFormer.

Monocular-based It involves using a single camera to capture data. The image is then fed into the model to predict 3D bounding boxes and category labels for each object of interest. The principal advantages of monocular-based 3D object detection are reflected in the low cost and high universality. However, the loss of depth information inevitably results in a significant degree of inaccuracy in the detection results, with the upper limit of detection accuracy being constrained. Furthermore, monocular-based detection is subject to significant limitations in terms of illumination and occlusion, which further restrict its practical applications. Early monocular methods include 3DVP (Xiang et al., 2015), Mono3D (Yan & Salman, 2017), Deep MANTA (Chabot et al., 2017), SubCNN (Xiang et al., 2017), 3D-RCNN (Kundu, Li, & Rehg, 2018), Mono3D++ (He & Soatto, 2019), Shift R-CNN (Naiden et al., 2019), CenterNet (Duan et al., 2019), GS3D (Li et al., 2019), M3D-RPN (Brazil & Liu, 2019), SMOKE (Liu et al., 2020), Kinematic 3D (Brazil et al., 2020), etc. After 2021, there are newer papers: MonoEF (Zhou et al., 2021), M3DSSD (Luo et al., 2021), CaDDN (Reading et al., 2021), GUP Net (Lu et al., 2021), FCOS3D (Wang et al., 2021), PGD (Wang et al., 2022b), ADD (Wu et al., 2022), MonoEdge (Zhu et al., 2023a), MonoATT (Zhou et al., 2023a), etc. The most recent papers are SSD-MonoDTR (He et al., 2023), MonoNeRD (Xu et al., 2023a), and MonoSKD (Wanga & Zheng, 2023).

Depth-based Depth-based methods estimate the depth map corresponding to each pixel in the image using a depth estimation network. The depth map is then used directly as input, combined with the original image, or converted into 3D point cloud data (Pseudo-LiDAR) for 3D object detection tasks. A drawback of this method is the separate training structure for depth and object detection, which could result in the loss of some implicit information. Some notable methods and models include AM3D (Ma et al., 2019), DDMP (Wang et al., 2021a), PCT (Wang et al., 2021b), MonoJSG (Lian et al., 2022), MonoDTR (Huang et al., 2022b), DID-M3D (Peng et al., 2022), Dfm (Wang et al., 2022a), MonoPGC (Wu et al., 2023b), and MoGDE (Zhou et al., 2022a). In the context of 3D object detection, the process of generating depth maps is a complex and pivotal aspect. AM3D is an illustrative case in point, as its emphasis is on the utilization of depth information, rather than the acquisition thereof. One of the primary reasons for the suboptimal performance of previous image-based 3D detectors is their inadequate utilization of depth maps. The approach of simply using a depth map as an additional channel of an RGB image and then expecting a neural network to automatically extract effective features is not optimal. This work employs a methodology whereby the estimated depth is transformed into a point cloud with the assistance of a camera calibration file provided by the KITTI. This transformed data is then utilized as the input form for subsequent processing.

Stereo Camera Stereo Camera methods use two or more cameras to capture two or more images of an object from different locations. The 3D coordinates of the points are calculated using the triangulation principle by calculating the position deviation of the corresponding points. These methods are cost-effective, suitable for both indoor and outdoor use, and sensitive to ambient light. Some notable models include Stereo R-CNN (Li et al., 2019), DSGN (Chen et al., 2020), RT3D-Stereo (Königshof et al., 2019), IDA-3D (Peng et al., 2020), CDN (Garg et al., 2020), CG-Stereo (Li et al., 2020), OC-Stereo (Pon et al., 2020), ZoomNet (Xu et al., 2020), YOLOStereo3D (Liu et al., 2021), SIDE (Peng et al., 2022), and LIGA-Stereo (Guo et al., 2021).

Pseudo-LiDAR Pseudo-LiDAR methods generally work by first obtaining the corresponding depth map from monocular or stereo images. The original image is then combined with the depth information to obtain the pseudo-LiDAR point cloud after projection transformation. Finally, the raw point cloud is replaced by pseudo-LiDAR to complete the 3D object detection. The Pseudo-LiDAR approach is not contingent on a particular depth estimation algorithm. In comparison to alternative camera-based techniques, pseudo-LiDAR is capable of markedly enhancing the precision of object detection over extended distances. Furthermore, the financial outlay required for pseudo-LiDAR is comparatively modest in comparison to that of LiDAR. Nevertheless, when compared to a genuine LiDAR-based, the Pseudo-LiDAR approach still exhibits a discernible discrepancy in the precision of 3D object detection. This shortfall can be attributed primarily to the deficiency in depth estimation accuracy, particularly the depth estimation error in the vicinity of the object. Methods related to Pseudo-LiDAR include Pseudo-LiDAR (Wang et al., 2019), Pseudo-LiDAR++ (You et al., 2019), PatchNet (Ma et al., 2020), Neighbor-Vote (Chu et al., 2021), DD3D (Park et al., 2021), Pseudo-Stereo (Chen et al., 2022b), and SMPL (Kim et al., 2022).

Multi-View Multi-View methods utilize a series of images from various perspectives, captured by either monocular or multi-camera systems, as raw data for 3D detection. These methods are cost-effective, efficient, and have wide-ranging application prospects. However, the lack of depth information makes it extremely challenging to accurately detect objects through perspective views. Notable applications include DETR3D (Wang et al., 2022), ImVoxelNet (Rukhovich et al., 2022), PETR (Liu et al., 2022b), and ORA3D (Roh et al., 2022b). Post-2023, several newer papers have emerged, including CAPE (Xiong et al., 2023), StreamPETR (Wang et al., 2023a), and VEDet (Chen et al., 2023). Multi-camera object detection is primarily applied to crowded scenes or object tracking in autonomous driving, serving as an important foundation for 3D multi-object tracking. While there are fewer papers on Multi-Camera methods, the more representative article is PolarFormer (Jiang et al., 2022) and Vampire (Xu et al., 2023b). The fundamental concept of BEV is to transform the traditional 2D image view and ranging perception associated with autonomous driving into a 3D perception from a bird's eye perspective. BEV methods are widely utilized in camera-based, LiDAR-based, and multi-modal fields. Representative papers include BEVDepth (Li et al., 2023), MV3D-Det (Wang et al., 2023b), QD-BEV (Zhang et al., 2023), Ego3D (Lu et al., 2022b), SparseBEV (Liu et al., 2023b), DistillBEV (Wang et al., 2023a), FB-BEV (Li et al., 2023a), SOGDet (Zhou et al., 2024), and IA-BEV (Jiao et al., 2023). The integration of BEV and transformer methods represents a significant research direction in the realm of BEV camera technology. BEVFormer (Li et al., 2022) represents a noteworthy study in the field. It effectively leverages both spatial and temporal information by interacting with spatial and temporal spaces through predefined grid-shaped BEV queries. This approach significantly enhances the accuracy of velocity estimation and the recall of objects under conditions of low visibility. BEVFormer v2 (Yang et al., 2023a) introduces an innovative BEV detector that incorporates perspective supervision. This two-stage BEV detector operates by feeding proposals from the perspective head into the BEV head for final predictions. Notably, BEVFormer v2 converges more rapidly and is better suited to contemporary image backbones. This advancement represents a significant step forward in the field, demonstrating the potential of integrating perspective supervision into BEV detection systems. BEVFormer, along with subsequent research findings (Jiang et al., 2023; Li et al., 2023b; Qin et al., 2023b) based on it, represents one of the most promising avenues of exploration in the field of camera-based studies.

Conclusion Presently, camera-based perception and detection systems offer a rich source of environmental information and hold a significant cost advantage over radar and LiDAR systems. However, it's important to acknowledge that the reliability of images can be compromised under certain conditions(e.g. tunnels at night). Weather conditions can also significantly influence the effectiveness of camera-based methods. To mitigate these limitations, point cloud or multi-modal based methods are often employed in addition to the currently popular BEVFormer methods. These approaches provide a more robust solution, ensuring reliable perception and detection across a variety of environmental conditions and scenarios.

3.3 Point Cloud-based Approaches

Point cloud data, with its abundant geometric information, tends to be more stable compared to other single-modal data. However, it does present its own set of challenges, such as higher costs, vulnerability to weather conditions, and missed detection due to sparsity. Methods for point cloud-based applications can be broadly classified into three main categories: those based on learning objectives, those based on data representations, and other methods. The category based on learning objectives primarily consists of anchor-based (Liu et al., 2022a) and anchor-free (Ge et al., 2020a) methods. The category based on data representations includes methods that are point-based, grid-based, point-voxel based, range-based, 4D radar methods, and more. Additionally, some other methods also play crucial roles in point cloud applications. These include data augmentation (Hahner et al., 2020), spatial-temporal sequences, and pseudo-labeling treatments, among others.

Anchor-based and Anchor-free Anchor-based methods utilize pre-defined boxes for bounding box encoding. However, the use of dense anchors can lead to an overwhelming number of potential target objects, necessitating the use of Non-Maximum Suppression (NMS). On the other hand, the Anchor-free method eliminates the need for a complex anchor design phase and can be flexibly applied to various views such as BEV, point view, depth view, etc. Some previous works have mentioned anchor-free concepts. For instance, PointRCNN (Shi et al., 2019) proposes a 3D proposal generation sub-network that doesn't rely on anchor boxes and is based on whole-scene point cloud segmentation. VoteNet (Qi et al., 2019) constructs 3D bounding boxes from voted interest points instead of predefined anchor boxes. However, these methods are not NMS-free, which makes them less efficient and less suitable for embedded systems. Moreover, PIXOR (Yang et al., 2018b) is a BEV detector rather than a 3D detector. Other notable works in this field include AFDet (Ge et al., 2020b), CenterPoint (Yin et al., 2021a), 3DSSD (Yang et al., 2020), and MGAF-3DSSD (Li et al., 2021).

Point-based Point-based methods are among the most common LiDAR-based approaches. These methods operate directly on the raw point cloud, which allows for more complete preservation of information and richer semantics. However, this approach can lead to increased memory access, reducing the efficiency of data operations. It also presents challenges for direct spatial convolution for feature extraction. Examples of LiDAR-based methods include Point R-CNN (Shi et al., 2019), IPOD (Yang et al., 2018), Point-GNN (Shi & Rajkumar, 2020), StarNet (Ngiam et al., 2019), STD (Yang et al., 2019), PointFormer (Pan et al., 2021), IA-SSD (Zhang et al., 2022), and RRD (Li et al., 2023a).

Voxel-based The process of voxelizing point cloud data involves extracting features from each voxel, which are then combined for global feature extraction. This method is highly efficient in terms of data computation, offering fast speeds and the ability to handle large volumes of data. However, it does have its drawbacks. These include the potential for data loss, the occurrence of numerous meaningless operations, and high memory requirements. Voxel-based methods are among the earlier applications of LiDAR technology. Numerous studies have been conducted in this area, including Vote3D (Wang & Posner, 2015), shortcite (Engelcke, Rao, Wang, Tong, & Posner, 2017), 3D-FCN (Li, 2017), VoxelNet (Zhou & Tuzel, 2018), MVF (Zhou et al., 2020), HVNet (Ye et al., 2020), CIA-SSD (Zheng et al., 2021a), Voxel R-CNN (Deng et al., 2021), Voxel Transformer (Mao et al., 2021b), TED (Wu et al., 2022), VoxelNeXt (Chen et al., 2023b), and MV-JAR (Xu et al., 2023). More recent contributions to this field include Diff3Det (Zhou et al., 2023), and SCP (Shan et al., 2023).

Pillars-based This method can be seen as a unique type of voxel that is unrestricted on the vertical scale. It was first introduced by PointPillars (Lang et al., 2019). Related studies in this area include Pillar-OD (Wang et al., 2020), SST (Fan et al., 2022a), PillarNet (Shi et al., 2022), PBTS (Lis & Kryjak, 2023), and Pillar-NeXt (Li et al., 2023).

Point-voxel-based The Point-voxel-based(Hybrid) method strikes a balance between accuracy and efficiency by combining the speed advantage of the voxel-based method and the high performance and accuracy of the point-based method. Relevant papers include PVCNN (Liu et al., 2019), PV-RCNN (Shi et al., 2020), Fast Point R-CNN (Chen et al., 2019), SPVNAS (Tang et al., 2020), SA-SSD (He et al., 2020), LiDAR R-CNN (Li et al., 2021), PV-RCNN++ (Shi et al., 2023), Pyramid R-CNN (Mao et al., 2021a), CT3D (Sheng et al., 2021), PDV (Hu et al., 2022), Bi3D (Yuan et al., 2023), PVT-SSD (Yang et al., 2023b), and Uni3DETR2023 (Wang et al., 2023b).

Range-based The method involves detecting objects from the range view (also known as range image) of the LiDAR. Papers related to this method include LaserNet (Meyer et al., 2019), LaserFlow (Meyer et al., 2020), RangeDet (Fan et al., 2021), PPC (Chai et al., 2021), RSN (Sun et al., 2021), RangeIoUDet (Liang et al., 2021), CVFNet (Gu et al., 2022), and FCOS (Tian et al., 2019).

BEV-based In autonomous driving and robotics, data from sensors like LiDAR and cameras are often converted into BEV representations for improved object detection, path planning, and other tasks. BEV simplifies complex three-dimensional environments into two-dimensional images, which is crucial for efficient computation in real-time systems. Recently, there has been an increase in LiDAR-based applications of BEV (point-based or voxel-based view transformation). Related studies include VeloFCN (Li et al., 2016), HDNet (Yang et al., 2018a), BirdNet (Beltrán et al., 2018), PIXOR (Yang et al., 2018b), BEVDetNet (Mohapatra et al., 2021), MGTANet (Koh et al., 2022), and GPA-3D (Li et al., 2023c).

4D Radar The development of radar-based 3D object detection is not only enhancing the capabilities of autonomous vehicles but also providing economic benefits. Recognized for its resilience and cost-effectiveness under adverse weather conditions, 4D radar has become a crucial component in autonomous driving. Here are some notable papers in this field. CHR4D (Palmer et al., 2023) analyzes the detection performance of existing models on the new data modality, evaluates them in depth, and conducts cross-model validation as well as cross-data set validation. MOD-4D (Pan et al., 2023a), an innovative solution tailored for radar-based tracking, showcases superior tracking precision of moving objects, largely surpassing the performance of the state-of-the-art. Multi-View Feature Assisted Network(MVFAN) (Pan et al., 2023b) is an end-to-end, anchor-free, and single-stage framework for 4D-radar-based 3D object detection for autonomous vehicles. CenterRadarNet (Cheng et al., 2023a) is an efficient joint architecture, designed to facilitate high-resolution representation learning from 4D radar data for 3D object detection and re-identification (re-ID) tasks. There are also some datasets related to 4D radar, such as TJ4DRadSet (Zheng et al., 2022b), K-Radar (Paek et al., 2022b), and Dual Radar (Zhang et al., 2023b). These datasets include a variety of scenes and improved annotations.

Data Augmentation Data augmentation can enhance the model's ability to generalize and improve its robustness. In the realm of object detection tasks, data augmentation can be implemented in two distinct manners: point cloud augmentation and label augmentation. It is a crucial technique for enhancing the efficiency of 3D point cloud detection and reducing annotation costs. Common data augmentation methods in point cloud include rotation, noise addition, downsampling, and varying degrees of masking. Data augmentation has been a focal point in LiDAR research, with related papers including PointAugmentation (Wang et al., 2021a), PA-AUG (Choi et al., 2021), ST3D (Yang et al., 2021), SE-SSD (Zheng et al., 2021b), ProposalContrast (Yin et al., 2022), HSSDA (Liu et al., 2023a), COMAug (Zhu et al., 2023b), and PG-RCNN (Koo et al., 2023).

Spatial-temporal and Sequence Spatial-temporal and sequence methods have also been hot topics in point cloud research. Papers related to these methods include MPP-Net (Chen et al., 2022a), DetZero (Ma et al., 2023), SUIT (Zhou et al., 2023b), and

SCOPE (Yang et al., 2023a). MPPNet is a flexible and high-performance 3D detection framework designed for 3D temporal object detection with point cloud sequences. DetZero is an offline tracker that works in conjunction with a multi-frame detector. It focuses on the completeness of generated object tracks. SUIT simplifies temporal information into sparse features for information fusion across frames. This not only significantly reduces the memory and computation cost of temporal fusion, but also performs well over state-of-the-art baselines. SCOPE is a novel collaborative perception framework that aggregates the spatial-temporal awareness characteristics across on-road agents in an end-to-end manner.

Pseudo-labeling Treatments Pseudo-labeling treatments include several generalized frameworks such as SSAPL (Xu et al., 2021), CL3D (Peng et al., 2023), and ReDB (Chen et al., 2023c). SSAPL, a novel semi-supervised framework, utilizes pseudo-labeling for outdoor 3D object detection tasks. It introduces the Adaptive Class Confidence Selection module (ACCS) to generate high-quality pseudo-labels. CL3D employs the Spatial Geometry Alignment module and Temporal Motion Alignment module to leverage motion features in sequential data frames, facilitating the alignment of two fields. Lastly, ReDB generates Reliable, Diverse, and class-balanced pseudo 3D boxes, iteratively guiding the self-training on a target domain with a different distribution. This method ensures the production of high-quality results in 3D object detection.

Conclusion 3D LiDAR is frequently utilized in autonomous driving due to its superior accuracy, although it comes with a higher price tag. LiDAR provides practical and precise 3D sensing capabilities in both daylight and nighttime conditions. LiDAR's performance can be compromised in adverse weather conditions such as fog, snow, and rain. Thus, the integration of different modal information becomes increasingly crucial in the task of autonomous driving scene perception.

3.4 Multi-modality-based Approaches

Perception technology in autonomous driving has seen rapid advancements in recent years, largely driven by deep learning techniques. To achieve accurate and robust sensing capabilities, self-driving cars are typically equipped with multiple sensors, making sensor fusion a critical component of the sensing system. The fusion of camera, millimeter-wave radar, and LiDAR data can reduce information redundancy and provide reliable sensing capabilities, but the system cost remains high. Multi-modal approaches primarily include multi-modal fusion and cross-modality interaction. Multi-modal fusion encompasses various combinations such as camera-radar (Fu et al., 2020), camera-LiDAR, and LiDAR-radar fusions.

Camera-LiDAR Fusion For camera-LiDAR, the most prevalent multi-modal fusion for 3D object detection is currently categorized into the following three primary approaches:

- **Early Fusion:** This approach involves fusing sensor data before feature extraction, maximizing the use of multi-modal information. Notable papers include F-PointNet (Qi et al., 2018), Complexer-YOLO (Simon et al., 2019), PointPainting (Vora et al., 2020), FusionPainting (Xu et al., 2021), F-PointPillars (Paigwar et al., 2021), PointAugmenting (Wang et al., 2021b), MVP (Yin et al., 2021b), VirConvNet (Wu et al., 2023), and others. Of the aforementioned methods, F-PointNet is the most representative, with the PointNet (Qi et al., 2017a) and PointNet++ (Qi et al., 2017b) models forming its foundation. The PointNet model is capable of directly processing point clouds, extracting global features that are effective for classification tasks. However, its local feature extraction capability is limited, which presents challenges in analysing complex scenes. Although PointNet itself does not run on the self-driving car dataset, it represents a significant milestone in the field of 3D object detection. PointNet++ is an enhanced version of PointNet, with the core of its design comprising a multilevel feature extraction structure that is capable of efficiently extracting both local features and global features. F-PointNet extends the application of PointNet to 3D target detection, offering the option of utilising either PointNet or PointNet++ for point cloud processing. Prior to point cloud processing, the image information is used to establish a series of a priori search ranges, thereby enhancing efficiency and accuracy.

- **Deep Fusion:** Also known as feature-level fusion, this approach involves the fusion of image and LiDAR features at an intermediate stage of a LiDAR-based 3D target detector, such as in the backbone network, proposal generation stage, or RoI refinement stage. Deep fusion is further divided into **RoI-level** fusion and **Point/Voxel-level** fusion. It is currently the most widely utilized fusion method. Typical methods include MV3D (Chen et al., 2017), AVOD (Ku et al., 2018), PointFusion (Xu et al., 2018), ContFuse (Liang et al., 2018), PI-RCNN (Xie et al., 2020), EPNet (Huang et al., 2020), 3D-CVF (Yoo et al., 2020), MMF (Liang et al., 2019), TransFusion (Bai et al., 2022), BEVFusion (Liu et al., 2022b), Deep-Fusion (Li et al., 2022), CAT-Det (Zhang et al., 2022), HMFI (Li et al., 2022), MSMDFusion (Jiao et al., 2022), LoGoNet (Li et al., 2023), SDVRF (Ren & Yin, 2023), EP-Net++ (Liu et al., 2022a), UniTR (Wang et al., 2023), SupFusion (Qin et al., 2023a), FGFusion (Yin et al., 2023), VCD (Huang et al., 2023) and others.

- **Late Fusion:** Also known as decision-level fusion, this approach requires joint alignment and labeling of the data only at the final fusion step. The advantage of decision-level fusion is its efficiency, as it only performs multi-modal fusion on the outputs of different modalities, avoiding complex interactions on intermediate features or input point clouds. However, since these methods do not rely on depth features from camera and LiDAR sensors, they are unable to integrate rich semantic information from different modalities, which limits their potential. Representative papers include CLOCs (Pang et al., 2020), and Fast-CLOCs (Pang et al., 2022).

Camera-Radar Fusion Among all sensor combinations, the fusion of Radar and Camera solutions offers the advantage of being complementary and cost-effective, regardless of lighting and weather conditions. However, fusing camera and millimeter-wave radar data presents a challenge due to each sensor's lack of information along a three-dimensional coordinate axis. The key to their fusion lies in resolving the ambiguity in the geometric correspondence between camera features and radar features. Notable studies in this area include CenterFusion (Nabati & Qi, 2021), CramNet (Hwang et al., 2022), MVFusion (Wu et al., 2023a), CRN (Kim et al., 2023), and REDFormer (Cui et al., 2023).

CenterFusion is one of the earlier approaches. It employs a center point detection network to detect objects by identifying their center points in the image. It then addresses the crucial data association problem using a novel frustum-based method to associate radar detections with their corresponding object's center point. CramNet is an efficient approach that fuses sensor readings from the camera and radar in a joint 3D space, leading to robust 3D object detection, even when a camera or radar sensor suddenly malfunctions on a vehicle. MVFusion introduces a novel Multi-View radar-camera Fusion method to achieve semantic-aligned radar features and enhance cross-modal information interaction. CRN is a novel camera-radar fusion framework that generates a semantically rich and spatially accurate BEV feature map for various tasks. The latest paper is REDFormer, which proposes a novel transformer-based 3D object detection model to tackle low visibility conditions by leveraging BEV camera-radar fusion.

LiDAR-Radar Fusion LiDAR and radar are integral components in autonomous driving systems, playing crucial roles in perceiving the surrounding environment. LiDAR offers precise 3D spatial sensing information, but its functionality is compromised in adverse weather conditions such as fog. Radar signals can diffract when encountering raindrops or mist particles due to their wavelength, but they are prone to significant noise. Recent state-of-the-art research suggests that the fusion of radar and LiDAR can result in robust detection even in adverse weather conditions. Notable approaches include RaLiBEV (Yang et al., 2022), Bi-LRFusion (Wang et al., 2023a), and MVDNet (Xie et al., 2023).

RaLiBEV is a learning-based anchor box-free object detection system that operates in BEV. It fuses features derived from the radar range-azimuth heat-map and the LiDAR point cloud to estimate potential objects. Bi-LRFusion is a bi-directional LiDAR-Radar fusion framework designed to tackle challenges and improve 3D detection for dynamic objects. It also alleviates problems caused by the absence of height information and extreme sparsity. MVDNet's performance can be enhanced by improving the training program to tolerate temporal misalignment of input data. This results in higher output frequencies with less loss of accuracy.

Other Fusion In addition to the aforementioned methods, several other fusion techniques also significantly contribute to the research of object detection for autonomous driving.

MapFusion (Fang et al., 2021) is a highly specific yet effective solution for object detection. It is detector-independent and can be seamlessly integrated into various detectors. Dual point cloud fusion is a relatively new method that offers numerous advantages, such as reducing external perturbations, handling complex weather conditions, and processing point cloud sparsity. Notable papers in this area include SPNet (Ju et al., 2022) and SSDA3D (Wang et al., 2022). Sparse Fuse Dense(SFD) (Wu et al., 2022) is an application that fuses point clouds and pseudo-point clouds. It utilizes pseudo point clouds generated from depth completion to address certain issues and proposes a new RoI fusion strategy, 3D-GAF, to make fuller use of information from different types of point clouds.

Cross-Modality Interaction Cross-modality interaction encompasses a variety of methods, including the fusion of multi-camera and LiDAR data, as well as the integration of monocular and LiDAR data, among others. This approach primarily employs the cross-modal knowledge distillation method. It concentrates on two key issues: multi-level information fusion, and information extraction and enhancement. Noteworthy methods in this field include CMKD (Hong et al., 2022), BiProDet (Zhang et al., 2023), X^3KD (Klingner et al., 2023), and ODM3D (Zhang et al., 2023a).

Conclusion Multi-modality 3D object detection is a current research hot-spot in 3D target detection. It primarily involves the use of cross-modal data to enhance the model's detection accuracy. Multi-modal data typically includes image data, point cloud data, millimeter-wave radar data, binocular depth data, and more. The multi-modal approaches of camera-radar, radar-LiDAR, and cross-modality interaction are increasingly being utilized.

3.5 Other Classifications

Other classifications encompass pedestrian detection, lane detection, and roadside cameras.

Pedestrian At present, there are not many research results on the 3D object detection of pedestrians. Notable works in this area include PiFeNet (Le et al., 2022) and CrossDTR (Tseng et al., 2023). PiFeNet incorporates a stack-able Pillar Aware Attention (PAA) module and a compact yet effective feature network (Mini-BiFPN) that facilitates bidirectional information flow and multi-level cross-scale feature fusion to better integrate multi-resolution features. CrossDTR employs cross-view and depth-guided transformers for 3D Object Detection. Extensive experiments have shown that this method significantly outperforms existing multi-camera methods by 10 percent in pedestrian detection and about 3 percent in overall mAP and NDS metrics.

Lane Detection An early application of lane lines was 3DLaneNet (Garnett et al., 2019), which predicts the 3D layout of lanes in a road scene from a single image. The approach explicitly handles complex situations such as lane merges and splits. 3D-LaneNet+ (Efrat, Bluvstein, Oron, Levi, Garnett, & Shlomo, 2020) is a camera-based DNN method for anchor free 3D lane detection which is able to detect 3d lanes of any arbitrary topology such as splits, merges, as well as short and perpendicular lanes. Two notable papers presented at CVPR2022, CLRNet (Zheng et al., 2022c) and GANet (Wang et al., 2022), have made significant advancements in the field. CLRNet aims to fully utilize both high-level and low-level features in lane detection. It first employs high-level semantic features for initial lane detection, followed by a refinement process based on low-level features. GANet introduces a new perspective to the problem of lane detection. Instead of extending the lane line point-by-point, each keypoint is directly regressed to the starting point of the lane line. This parallel processing approach greatly enhances efficiency. Furthermore, a specialized dataset for lane detection, ONCE-3DLanes (Yan et al., 2022), has been introduced. This real-world autonomous driving dataset provides lane layout annotations in 3D space. The dataset has been benchmarked and a novel evaluation metric has been provided. Extensive experiments have been conducted using both existing approaches and the proposed method. DCM-PRLD (Han & Shen, 2023) is the latest paper on lane detection. It presents a novel approach to the lane detection task by decomposing it into two parts: curve modeling and ground height regression. Additionally, it has unified the 2D and 3D lane detection tasks by designing a new framework and a series of losses.

Roadside Cameras A number of papers on roadside cameras were published in 2023, including BEVHeight (Yang et al., 2023d), BEVHeight++ (Yang et al., 2023b), Mono-GAE (Yang et al., 2023c) and CoBEV (Shi et al., 2023), and a related dataset A9 Intersection Dataset (Zimmer et al., 2023). BEVHeight leverages intelligent roadside cameras to extend perception ability beyond the visual range. BEVHeight++ is an enhancement of the

BEVHeight method, incorporating both height and depth encoding techniques to achieve a more accurate and robust projection from 2D to BEV spaces. MonoGAE introduces a novel framework for roadside monocular 3D object detection with ground-aware embedding. CoBEV is an innovative end-to-end monocular 3D object detection framework that integrates depth and height to construct robust BEV representations. The A9 Intersection Dataset consists of labeled LiDAR point clouds and synchronized camera images, featuring 4.8k images and point clouds with over 57.4k manually labeled 3D boxes. With ten object classes, it captures a high diversity of road users engaged in complex driving maneuvers such as left and right turns, overtaking, and U-turns.

3.6 Problem-solving Approaches

At their core, various 3D object detection methods are designed to address specific challenges encountered during object detection. These challenges include small and distant objects, occlusion and overlap, special weather conditions, detecting safety, among others. The issues and representative articles are discussed in the following sections.

- **Small and Distant Objects:** When sensors capture fewer point clouds for long-range objects, the target may blend with the background, leading to false detection. Solutions include dual point cloud fusion, data augmentation, perspective-aware aggregation, effective 3D detectors, and annotated datasets. Notable solutions include FSD (Fan et al., 2022b), Far3Det (Gupta et al., 2023), MoGDE (Zhou et al., 2022b), Far3D (Jiang et al., 2023), and RangeFSD (Khoche et al., 2023).

- **Occlusion and Overlap:** Occlusion can be inter-class (target is obscured by objects of the same class) or among class (target is obscured by objects of other classes). Solutions include fine-tuning the GT bounding box of occluded targets, data enhancement, adding attention mechanisms, or positive and negative sample matching. Representative works include BtcDet (Xu et al., 2022b), Real3D-Aug (Šebek et al., 2022), and ORA3D (Roh et al., 2022a).

- **Distant Objects and Occlusion:** Works that address both distant objects and occlusion issues include PC-RGNN (Zhang et al., 2021), MoDAR (Li et al., 2023b), PG-RCNN (Koo et al., 2023), among others.

- **Special Weather Conditions:** Adverse conditions like snow, fog, and rain can introduce noise into measurements, impacting LiDAR-based perception systems. Solutions often involve filter networks, camera-radar fusion, specialized datasets, etc. Specialized datasets include K-Radar (Paek et al., 2022b) and Ithaca365 (Diaz-Ruiz et al., 2022). Key research in this area includes LSS (Hahner et al., 2022), REDFormer (Cui et al., 2023), and TROD (Piroli et al., 2023).

- **Detecting Safety:** Safety detection has often been an overlooked aspect of target detection in the past. Factors such as false negatives, image attacks, and frustum attacks can potentially lead to hazardous situations in autonomous driving. Consequently, there has been a surge in research papers on this topic in the last two to three years. In response to the camera mode attack, the MDE (Cheng et al., 2022) method can generate stealthy, effective, and robust adversarial patches for different target objects and models. There are several security methods for detecting adversarial attacks in camera-LiDAR fusion, including FNE (Cheng et al., 2023b), SecurityFABA (Hallyburton et al., 2022), and FocalFormer3D (Chen et al., 2023a). FNE proposes an attack framework that targets advanced camera-LiDAR fusion models with adversarial patches. SecurityFABA illustrates the damaging effects of the ”frustum” attack on sensor fusion with multi-frame tracking through comprehensive experiments. FocalFormer3D, a simple yet effective detector, excels at identifying difficult objects and improving prediction recall, thereby addressing the issue of false negatives more effectively. Deep neural networks (DNNs) are increasingly integrated into LiDAR-based perception systems for autonomous vehicles, necessitating robust performance under adversarial conditions. Some of the more notable papers in this area include ADoPT (Cho et al., 2023), Transcender-MC (Răduţoiu et al., 2023), and EAR (Yang & Ji, 2023).

3.7 Analysis in 3D Object Detection Classifications and Approaches

It’s clear that camera-based methods are more affordable while point cloud methods are more efficient. Millimeter wave radar serves as an effective auxiliary tool in 3D object detection. The integration of LiDAR and cameras can yield image data with depth. Some studies have even combined LiDAR and millimeter-wave radar with cameras to enhance safety redundancy. While the integration of data from multiple sensors, known as sensor fusion, offers substantial advantages, it also presents considerable challenges in terms of system design. One of the primary obstacles is the lack of synchronization among various types of sensors in both temporal and spatial domains. Additionally, deploying sensors from diverse perspectives in the spatial domain can further complicate the fusion process. Moreover, designing fusion methods requires careful consideration of several issues. These include potential information loss, difficulties with multi-sensor calibration and data alignment, challenges related to cross-modality data enhancement, and a limited number of assessment indicators coupled with mislabeling in the dataset. Each of these factors must be meticulously addressed to ensure efficient implementation of sensor fusion. Despite the various challenges associated with multi-modality fusion and interaction methods, these techniques hold immense potential and bear significant implications for research.

In addition to sensors and multi-modal approaches, classifications such as pedestrians, lane lines, and roadside cameras are also crucial for object detection. There’s still much work to be done in specialized pedestrian detection and lane detection as there are fewer papers in these areas. Research on roadside cameras has been gaining traction since 2023, but most papers are yet to be included in conferences and top journals.

3D object detection aims to accurately localize a object. This process involves addressing challenges such as small or distant objects, occlusion, crowded scenarios, adverse weather conditions, safe detection, etc. Between 2022 and 2023, many papers were published addressing the long-range problem of object detection, focusing on point cloud sparsity and monocular depth information. Most experiments were performed on KITTI and Argoverse2 datasets. However, there’s a noticeable lack of research in data enhancement and multi-modality. The main solutions to the occlusion problem are data enhancement and effective

				KITTI car AP			KITTI all mAP	nuScenes all mAP		Waymo car mAP and mAPH			
Methods	Classification	Year	FPS (ms)	Easy	Mod.	Hard	mAP	mAP	NDS	L1 mAP	L1 mAPH	L2 mAP	L2 mAPH
CenterNet (Kundu et al., 2018)	Camera-Monocular	2019	-	-	-	-	-	33.80	40.00	-	-	-	-
ROI-10D (Manhardt et al., 2019)	Camera-Monocular	2019	**200**	4.32	2.02	1.46	-	-	-	-	-	-	-
PCT (Wang et al., 2021b)	Camera-Depth	2021	-	21.00	13.37	11.31	-	-	-	14.70	-	14.67	-
LIGA-Stereo (Guo et al., 2021)	Camera-Stereo	2021	**350**	**81.39**	**64.66**	**57.22**	47.13	-	-	-	-	-	-
PatchNet (Ma et al., 2020)	Camera-Pseudo-LiDAR	2021	**400**	15.68	11.12	10.07	-	-	-	39.00	-	38.00	-
DETR3D (Wang et al., 2022)	Camera-Multi-View	2022	-	-	-	-	-	41.20	47.90	-	-	-	-
CAPE (Xiong et al., 2023)	Camera-Multi-View	2023	-	-	-	-	-	44.00	61.00	-	-	-	-
PolarFormer (Jiang et al., 2022)	Camera-Multi-Camera	2022	-	-	-	-	-	49.30	57.20	-	-	-	-
StarNet (Ngiam et al., 2019)	LiDAR-Point	2019	-	81.63	73.99	67.07	-	-	-	53.70	-	-	-
Point-GNN (Shi & 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	-	53.60	-	-	-	-	-
IA-SSD (Zhang et al., 2022)	LiDAR-Point	2022	83	88.87	80.32	75.04	64.34	-	-	70.53	69.67	61.55	60.8
VoxelNet (Zhou & Tuzel, 2018)	LiDAR-Voxel	2018	**220**	77.47	65.11	57.73	-	-	-	-	-	-	-
CenterPoint (Yin et al., 2021a)	LiDAR-Voxel	2021	70	-	-	-	-	58.00	65.50	76.70	-	68.80	-
Voxel R-CNN (Deng et al., 2021)	LiDAR-Voxel	2021	40	**90.9**	81.62	77.06	-	-	-	75.59	-	66.59	-
TED-M (Wu et al., 2022)	LiDAR-Voxel	2022	18	**91.61**	**85.28**	**80.68**	**70.99**	-	-	**79.26**	**78.73**	**70.50**	**70.07**
VoxelNeXt (Chen et al., 2023b)	LiDAR-Voxel	2023	-	-	-	-	-	**66.20**	**71.40**	-	-	-	-
PointPillars (Lang et al., 2019)	LiDAR-Pillars	2019	16	79.05	74.99	68.30	66.19	40.10	55.00	-	-	-	-
PillarNet (Shi et al., 2022)	LiDAR-Pillars	2022	16	-	-	-	-	**66.00**	**71.40**	**79.09**	**78.59**	**70.92**	**70.46**
RangeIoUDet (Liang et al., 2021)	LiDAR-Range	2021	45	88.60	79.80	76.76	**73.79**	-	-	-	-	-	-
RSN (Sun et al., 2021)	LiDAR-Range	2021	67	-	-	-	-	-	-	78.40	78.10	69.50	69.10
CVF-Net (Gu et al., 2022)	LiDAR-Range	2022	28	88.75	70.70	71.95	-	54.90	63.30	-	-	-	-
HDNet (Yang et al., 2018a)	LiDAR-BEV	2018	50	89.14	86.57	78.32	-	-	-	-	-	-	-
MGTANet (Koh et al., 2022)	LiDAR-BEV	2022	-	-	-	-	-	**67.50**	**72.70**	-	-	-	-
PV-RCNN (Shi et al., 2020)	LiDAR-Point-Voxel	2020	-	**90.25**	81.43	76.82	-	-	-	77.51	76.89	68.98	68.41
PV-RCNN++ (Shi et al., 2023)	LiDAR-Point-Voxel	2021	-	**90.14**	81.88	77.15	65.47	-	-	**79.25**	**78.78**	**70.61**	**70.18**
PG-RCNN (Koo et al., 2023)	Data Augmentation	2023	-	**92.73**	85.26	82.83	**76.01**	-	-	-	-	-	-
DetZero (Ma et al., 2023)	Spatial-Temporal and Sequence	2023	**200**	-	-	-	-	-	-	**92.17**	-	**87.32**	**85.15**
PointAugmenting (Wang et al., 2021b)	Fusion-Early	2021	**542**	-	-	-	-	**66.80**	**71.00**	67.41	-	62.70	-
3D-CVF (Yoo et al., 2020)	Fusion-Deep	2020	75	89.20	80.05	73.11	-	42.17	49.78	-	-	-	-
EPNet (Huang et al., 2020)	Fusion-Deep	2020	-	89.81	79.28	74.59	-	-	-	-	-	-	-
EPNet++ (Liu et al., 2022a)	Fusion-Deep	2022	-	**92.51**	83.17	82.27	**74.27**	-	-	76.57	76.10	68.29	67.86
MSMDFusion (Jiao et al., 2022)	Fusion-Deep	2022	2	-	-	-	-	**71.50**	**74.00**	-	-	-	-
LoGoNet (Li et al., 2023)	Fusion-Deep	2023	-	**91.80**	85.06	80.74	**74.40**	-	-	**88.33**	**87.87**	**82.17**	**81.72**
SDVRF (Ren & Yin, 2023)	Fusion-Deep	2023	-	**93.15**	86.89	84.63	**75.12**	-	-	-	-	-	-
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	65.10	63.10	68.70	-	-	-	-
MapFusion (Fang et al., 2021)	Fusion-Other	2021	-	-	-	-	-	60.61	67.97	-	-	-	-
SFD (Wu et al., 2022)	Fusion-Other	2022	-	**91.73**	84.76	77.92	**76.58**	-	-	-	-	-	-
BiProDet (Zhang et al., 2023)	Cross-Modality Interaction	2023	-	89.13	82.97	80.05	**70.13**	-	-	78.36	77.91	69.45	69.04

Table 2: Comparative analysis and experimental results of 3D object detection.

detectors. Recently, more papers have been utilizing point cloud enhancement methods to solve occlusion and long-range detection problems concurrently. For adverse weather conditions, effective solutions include the use of specialized datasets (K-Radar and Ithaca365) and multi-modality (especially using the Camera-radar or radar-LiDAR approach).

4. 3D Object Detection Experimental Evaluation

Through extensive screening and analysis, we have obtained the experimental measurement tables for the categories of camera, point cloud, and multi-modality, as shown in Table 2. The sections of the table in bold font are the areas that require focused attention.

Table 2 showcases representative methods selected from over 200 published conferences and journals. It provides the original inference time (ms) as stated in the papers, and reports the AP (%) for 3D car detection on the KITTI test benchmark, mAP (%), and NDS on the nuScenes test set, Level 1 (L1) mAP, mean Average Precision Harmonic (mAPH) and Level 2 (L2) mAP, mAPH on the waymo validation set. The methods are categorized based on sensor types and methods and are arranged chronologically by their year of publication. It's worth noting that more recent and efficient methods are primarily concentrated after 2020. In terms of datasets, most methods still utilize the KITTI, nuScenes,waymo Open, and argoverse datasets. Among them, KITTI is more widely used, while nuScenes and waymo are more efficiently used with LiDAR-based and multi-modal methods. In terms of speed performance, camera-based methods are the fastest, with most Frames Per Second (FPS) above 200ms, and some LiDAR-point and LiDAR-voxel methods also perform well. In terms of KITTI mAP, some LiDAR-Voxel, LiDAR-Range and data augmentation methods, most

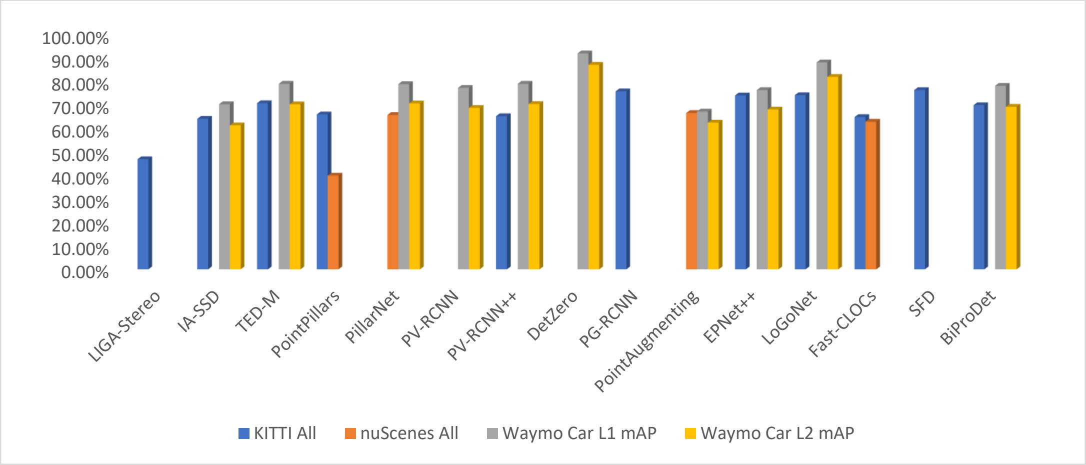

Figure 5: An analysis of mAP values. The figure presents a comparison of select methods and models from Table 2, focusing on four key metrics: KITTI mAP, nuScenes mAP, and Waymo Car L1 and L2 mAP. These methods, which have demonstrated relative efficiency, serve as benchmark models for 3D object detection and tracking.

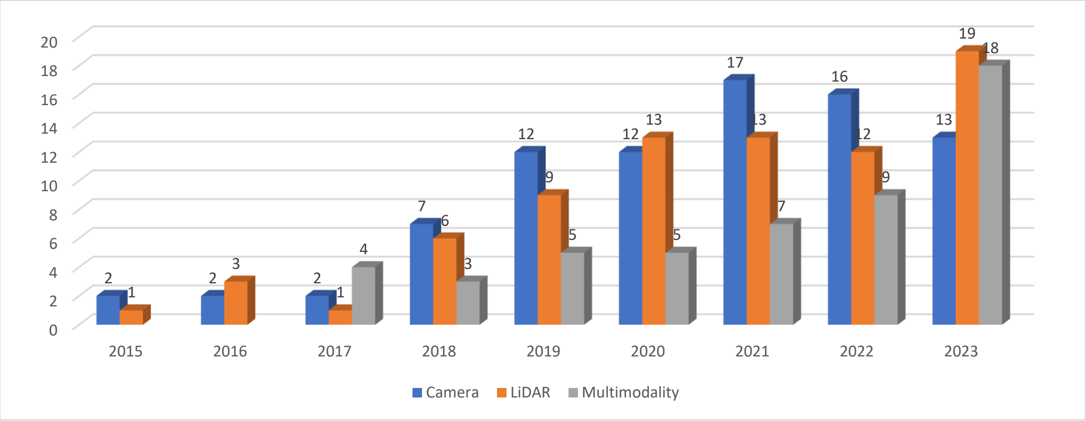

Figure 6: Annual categorization of 3D object detection. The figure presents an annual categorization of 3D object detection research from 2015 to 2023. It tallies the most influential publications from various journals or conferences during this period. These publications are classified based on the type of sensors they utilize: Camera-based, LiDAR-based, or multi-modal.

deep fusion and cross-modality interaction can reach above 70%. In terms of nuScenes mAP, some LiDAR-Voxel methods and many camera-LiDAR fusion methods can exceed 60%. As for the L1 and L2 mAP of waymo cars, LiDAR-voxel, LiDAR-pillars, spatial-temporal and sequence, and camera-LiDAR Fusion methods perform better.

5. Conclusion and Future Directions

From the preceding description and analysis, our understanding of 3D object detection has become more lucid. Next, we will encapsulate this knowledge succinctly and highlight potential avenues for future research.

5.1 Conclusion

Through a comparative analysis of various methods, we can draw several clear conclusions:

Camera-based methods are widely employed for 3D object detection under standard weather conditions, owing to their affordability, rapid processing speed, and low model complexity. However, under complex weather conditions, these sensors may not deliver optimal performance and they lack depth information. Presently, the primary focus of research in camera-based detection is on pseudo-LiDAR and innovative deep learning methods rooted in monocular detection. Additionally, there has been a significant increase in research exploring multi-view and multi-camera approaches, with a special emphasis on BEV-based and BEVFormer.

Point cloud-based methods have the advantages of semantic richness and high accuracy, but the sparseness of point clouds can lead to some issues. LiDAR-voxel methods have the advantage of speed, but they require specific hardware and are prone to data loss. BEV-based methods are beneficial for path planning and collision avoidance, offering high accuracy and safety, making them an important basis for 3D MOT. However, they are relatively costly and require significant computational power. Data augmentation is a crucial method in 3D target detection, primarily addressing the issue of point cloud sparsity. It can improve detection efficiency and reduce labeling costs, and is widely used in LiDAR, radar, pseudo-point cloud, and multi-modal directions. However, excessive data augmentation may degrade the model's performance and even lead to incorrect detection. Spatial-temporal and sequence methods have recently published several papers and are one of the recent hot topics. These methods are more suitable as a technical basis for 3D MOT, focusing on generating the integrity of object trajectories. 4D radar is relatively low-cost and is not affected by adverse weather conditions. However, it is less suitable for long-distance sensing, and the related datasets and research results are scarce.

Multi-modal approaches encompass multi-modal fusion and cross-modality interaction. In terms of Camera-LiDAR fusion, early fusion benefits from a simple structure and fast processing speed, but it demands high hardware and computational power. Late fusion, while having a simpler structure, is effective in improving perception accuracy, though it has fewer research results. Currently, most of the research direction of Camera-LiDAR fusion classification is focused on Deep fusion. The challenge with Deep fusion lies in solving the feature alignment problem of multiple modalities, but it can address issues such as difficult localization, small targets, and point cloud sparsity. Camera-Radar fusion is a relatively new research direction, offering advantages such as low cost and suitability for special weather detection, but it is not ideal for long-distance detection. LiDAR-Radar fusion research shows promise, with the advantage of using homologous point cloud data, which is convenient for feature fusion. However, this method has a relatively high hardware cost and also faces issues such as point cloud sparsity. Novel fusion methods such as dual point cloud fusion, point cloud and pseudo-point cloud fusion are emerging research directions in the field of multi-modality. Cross-modal interaction represents a relatively new area of research. It primarily employs methods such as knowledge distillation and data augmentation for the extraction and fusion of information. This approach allows for more robust and comprehensive understanding of data by leveraging the strengths of different modalities.

Currently, research on pedestrian detection, lane detection, roadside cameras, etc., is relatively sparse. Integrating these into the whole 3D target detection system is a direction worth exploring in depth. Additionally, there is a lack of specialized datasets with better labeling information, such as for LiDAR-radar fusion, Camera-radar fusion, and Cross-modality interaction.

5.2 Future Directions

The field of 3D object detection is indeed diverse, encompassing a wide range of classification and modeling methods. This can often be overwhelming for researchers, especially novices. This review aims to demystify this complexity by providing a comprehensive overview of the classifications, methodologies, and experiments involved in 3D object detection. It offers a holistic understanding of all facets of this field. A well-rounded review, such as this, also necessitates a strong foresight into the field.

Based on our thorough analysis and experiments, it's evident that the following represent the most recent and future significant areas of research in 3D object detection:

- **Camera-based** Recent advancements in 3D object detection from single images have leveraged monocular depth estimation to produce 3D point clouds, transforming cameras into pseudo-LiDAR sensors. This approach is gaining significance in camera-based applications and is increasingly being applied to Camera-LiDAR and Camera-radar multi-modality. There has been a surge in work related to multi-cameras, multi-view, and stereo cameras. Stereo-based 3D detection, which detects 3D object bounding boxes from stereo images using intermediate depth maps or implicit 3D geometry representations, provides a cost-effective solution for 3D perception. Multi-view and multi-cameras 3D object detection, due to its low cost and high efficiency, has shown promising application prospects. Furthermore, BEVFormer is also a significant area of research. It is anticipated that in the future, the combination of BEV and Transformer will likely supersede the previous model of 2D combined with CNN, gradually becoming the mainstream approach for autonomous driving perception. This implies that everything from the hardware chip, sensor camera, software algorithms, model deployment, to data collection calibration, will need to adapt and change accordingly.

- **Point Cloud-based** Key areas of LiDAR research include data augmentation, 4D radar, and the spatial-temporal and serialization of point clouds. Data enhancement undoubtedly plays a significant role in LiDAR, radar, and multi-modal research, addressing issues such as sparsity and miss detection. 4D radar, recognized for its resilience and cost-effectiveness under adverse weather conditions, plays a pivotal role in autonomous driving. Recently, spatial-temporal and serialization methods for point clouds have demonstrated significant research results in object detection, particularly in 3D MOT.

- **Multi-modal-based** Deep fusion techniques are proving highly effective in multi-modal applications and are currently the most widely used. Several important methods, such as Fine-Grained, distillation-based, data enhancement, and good detection network, have emerged.

- **Cross-modal Interaction** Cross-modality interaction has recently emerged as a hot topic in research. The alignment problem in cross-modality is a critical aspect of multi-modal fusion and cross-modality interaction. The most common approach is to employ a comprehensive knowledge distillation framework across different modalities, tasks, and stages. As we look to the future, we can expect an increasing number of applications to focus on the interaction between LiDAR (radar) and stereo-cameras, multi-cameras, and multi-views.

- **Cooperative Perception** Since 2023, Cooperative Perception on 3D MOT has garnered significant attention due to its ability to enhance scene comprehension by integrating data from various agents, such as vehicles and infrastructure. This field can be broadly divided into two main categories: **Vehicle-to-Everything (V2X)** perception methods, which include techniques like AR2VP (Tan et al., 2023), FFNet (Yu et al., 2023), and DI-V2X (Xiang et al., 2023), among others. **Vehicle-to-Vehicle (V2V)** perception technologies, which encompass methods such as OFDM (Sheng et al., 2023) and SiCP (Qu et al., 2023). Each of these categories contributes uniquely to the overall understanding of the scene, thereby enhancing the effectiveness of 3D MOT.

- **Other Classifications** There is increasing attention towards pedestrian detection, lane detection, roadside camera detection, and detection in special weather conditions. This will be one of the more promising research directions as the relevant datasets, labeling information, experimental results, and other supports are still scarce.

- **Detecting Safety** Identifying unsafe conditions is a crucial component of autonomous driving, as it directly influences the trustworthiness of environmental perception. Any reduction in this trustworthiness can lead to serious consequences for the operation and safety of self-driving vehicles. Future investigations in this field may consider a range of possibilities. These encompass multi-modal attacks, adversarial patching, spoofing attacks, and the issue of false negatives. These topics present substantial challenges and opportunities for improving the safety and dependability of autonomous driving applications.

Acknowledgments

This work is supported by National Science and Technology Major Project (2021ZD0111000/ 2021ZD0111004), the Science and Technology Commission of Shanghai Municipality Grant (No. 21511100101, 22511105901,22DZ2229004), the Open Research Fund of Key Laboratory of Advanced Theory and Application in Statistics and Data Science (East China Normal University), Ministry of Education. Xin Lin and Xin Li are the corresponding author. Xin Lin is also a member of Key Laboratory of Advanced Theory and Application in Statistics and Data Science (East China Normal University), Ministry of Education.

References

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*, pp. 1090–1099.

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*, pp. 3517–3523. IEEE.

Brazil, G., & Liu, X. (2019). M3d-rpn: Monocular 3d region proposal network for object detection. In *ICCV*, pp. 9287–9296.

Brazil, G., Pons-Moll, G., Liu, X., & Schiele, B. (2020). Kinematic 3d object detection in monocular video. In *ECCV*, pp. 135–152. Springer.

Brenner, M., Reyes, N. H., Susnjak, T., & Barczak, A. L. (2023). Rgb-d and thermal sensor fusion: a systematic literature review. *IEEE Access*.

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*, pp. 11621–11631.

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*, pp. 2040–2049.

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*, pp. 16000–16009.

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*, pp. 8748–8757.

Chen, D., Li, J., Guizilini, V., Ambrus, R., & Gaidon, A. (2023). Viewpoint equivariance for multi-view 3d object detection. *arXiv preprint arXiv:2303.14548*.

Chen, X., Ma, H., Wan, J., Li, B., & Xia, T. (2017). Multi-view 3d object detection network for autonomous driving. In *CVPR*, pp. 1907–1915.

Chen, X., Shi, S., Zhu, B., Cheung, K. C., Xu, H., & Li, H. (2022a). Mppnet: Multi-frame feature intertwining with proxy points for 3d temporal object detection. In *ECCV*, pp. 680–697. Springer.

Chen, Y.-N., Dai, H., & Ding, Y. (2022b). Pseudo-stereo for monocular 3d object detection in autonomous driving. In *CVPR*, pp. 887–897.

Chen, Y., Liu, S., Shen, X., & Jia, J. (2019). Fast point r-cnn. In *ICCV*, pp. 9775–9784.

Chen, Y., Liu, S., Shen, X., & Jia, J. (2020). Dsgn: Deep stereo geometry network for 3d object detection. In *CVPR*, pp. 12536–12545.

Chen, Y., Yu, Z., Chen, Y., Lan, S., Anandkumar, A., Jia, J., & Alvarez, J. M. (2023a). Focalformer3d: Focusing on hard instance for 3d object detection. In *ICCV*, pp. 8394–8405.

Chen, Y., Liu, J., Zhang, X., Qi, X., & Jia, J. (2023b). Voxelnext: Fully sparse voxelnet for 3d object detection and tracking. *arXiv preprint arXiv:2303.11301*.

Chen, Z., Luo, Y., Wang, Z., Baktashmotlagh, M., & Huang, Z. (2023c). Revisiting domain-adaptive 3d object detection by reliable, diverse and class-balanced pseudo-labeling. In *ICCV*, pp. 3714–3726.

Cheng, J.-H., Kuan, S.-Y., Latapie, H., Liu, G., & Hwang, J.-N. (2023a). Centerradarnet: Joint 3d object detection and tracking framework using 4d fmcw radar. *arXiv preprint arXiv:2311.01423*.

Cheng, Z., Choi, H., Liang, J., Feng, S., Tao, G., Liu, D., Zuzak, M., & Zhang, X. (2023b). Fusion is not enough: Single-modal attacks to compromise fusion models in autonomous driving. *arXiv preprint arXiv:2304.14614*.

Cheng, Z., Liang, J., Choi, H., Tao, G., Cao, Z., Liu, D., & Zhang, X. (2022). Physical attack on monocular depth estimation with optimal adversarial patches. In *ECCV*, pp. 514–532.

Cho, M., Cao, Y., Zhou, Z., & Mao, Z. M. (2023). Adopt: Lidar spoofing attack detection based on point-level temporal consistency. *arXiv preprint arXiv:2310.14504*.

Choi, J., Song, Y., & Kwak, N. (2021). Part-aware data augmentation for 3d object detection in point cloud. In *IROS*, pp. 3391–3397. IEEE.

Chu, X., Deng, J., Li, Y., Yuan, Z., Zhang, Y., Ji, J., & Zhang, Y. (2021). Neighbor-vote: Improving monocular 3d object detection through neighbor distance voting. In *Proceedings of the 29th ACM International Conference on Multimedia*, pp. 5239–5247.

Cui, C., Ma, Y., Lu, J., & Wang, Z. (2023). Radar enlighten the dark: Enhancing low-visibility perception for automated vehicles with camera-radar fusion. *arXiv preprint arXiv:2305.17318*.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., & Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. In *CVPR*, pp. 248–255. Ieee.

Deng, J., Shi, S., Li, P., Zhou, W., Zhang, Y., & Li, H. (2021). Voxel r-cnn: Towards high performance voxel-based 3d object detection. In *AAAI*, Vol. 35, pp. 1201–1209.

Diaz-Ruiz, C. A., Xia, Y., You, Y., Nino, J., Chen, J., Monica, J., Chen, X., Luo, K., Wang, Y., Emond, M., et al. (2022). Ithaca365: Dataset and driving perception under repeated and challenging weather conditions. In *CVPR*, pp. 21383–21392.

Duan, K., Bai, S., Xie, L., Qi, H., Huang, Q., & Tian, Q. (2019). Centernet: Keypoint triplets for object detection. In *ICCV*, pp. 6569–6578.

Efrat, N., Bluvstein, M., Oron, S., Levi, D., Garnett, N., & Shlomo, B. E. (2020). 3d-lanenet+: Anchor free lane detection using a semi-local representation. *arXiv preprint arXiv:2011.01535*.

Elharrouss, O., Hassine, K., Zayyan, A., Chatri, Z., Al-Maadeed, S., Abualsaud, K., et al. (2023). 3d objects and scenes classification, recognition, segmentation, and reconstruction using 3d point cloud data: A review. *arXiv preprint arXiv:2306.05978*.

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*, pp. 1355–1361. IEEE.

Fan, L., Pang, Z., Zhang, T., Wang, Y.-X., Zhao, H., Wang, F., Wang, N., & Zhang, Z. (2022a). Embracing single stride 3d object detector with sparse transformer. In *CVPR*, pp. 8458–8468.

Fan, L., Wang, F., Wang, N., & ZHANG, Z.-X. (2022b). Fully sparse 3d object detection. *Advances in Neural Information Processing Systems*, *35*, 351–363.

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*, pp. 2918–2927.

Fang, J., Zhou, D., Song, X., & Zhang, L. (2021). Mapfusion: A general framework for 3d object detection with hdmaps. In *IROS*, pp. 3406–3413. IEEE.

Fu, Y., Tian, D., Duan, X., Zhou, J., Lang, P., Lin, C., & You, X. (2020). A camera–radar fusion method based on edge computing. In *EDGE*, pp. 9–14. IEEE.

Garg, D., Wang, Y., Hariharan, B., Campbell, M., Weinberger, K. Q., & Chao, W.-L. (2020). Wasserstein distances for stereo disparity estimation. *Advances in Neural Information Processing Systems*, *33*, 22517–22529.

Garnett, N., Cohen, R., Pe'er, T., Lahav, R., & Levi, D. (2019). 3d-lanenet: end-to-end 3d multiple lane detection. In *ICCV*, pp. 2921–2930.

Ge, R., Ding, Z., Hu, Y., Wang, Y., Chen, S., Huang, L., & Li, Y. (2020a). Afdet: Anchor free one stage 3d object detection. *arXiv preprint arXiv:2006.12671*.

Ge, R., Ding, Z., Hu, Y., Wang, Y., Chen, S., Huang, L., & Li, Y. (2020b). Afdet: Anchor free one stage 3d object detection. *arXiv preprint arXiv:2006.12671*.

Geiger, A., Lenz, P., Stiller, C., & Urtasun, R. (2013). Vision meets robotics: The kitti dataset. *The International Journal of Robotics Research*, *32*(11), 1231–1237.

Gu, J., Xiang, Z., Zhao, P., Bai, T., Wang, L., Zhao, X., & Zhang, Z. (2022). Cvfnet: Real-time 3d object detection by learning cross view features. In *IROS*, pp. 568–574. IEEE.

Guo, X., Shi, S., Wang, X., & Li, H. (2021). Liga-stereo: Learning lidar geometry aware representations for stereo-based 3d detector. In *ICCV*, pp. 3153–3163.

Gupta, S., Kanjani, J., Li, M., Ferroni, F., Hays, J., Ramanan, D., & Kong, S. (2023). Far3det: Towards far-field 3d detection. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 692–701.

Hahner, M., Dai, D., Liniger, A., & Van Gool, L. (2020). Quantifying data augmentation for lidar based 3d object detection. *arXiv preprint arXiv:2004.01643*.

Hahner, M., Sakaridis, C., Bijelic, M., Heide, F., Yu, F., Dai, D., & Van Gool, L. (2022). Lidar snowfall simulation for robust 3d object detection. In *CVPR*, pp. 16364–16374.

Hallyburton, R. S., Liu, Y., Cao, Y., Mao, Z. M., & Pajic, M. (2022). Security analysis of {Camera-LiDAR} fusion against {Black-Box} attacks on autonomous vehicles. In *31st USENIX Security Symposium (USENIX Security 22)*, pp. 1903–1920.

Han, W., & Shen, J. (2023). Decoupling the curve modeling and pavement regression for lane detection. *arXiv preprint arXiv:2309.10533*.

He, C., Zeng, H., Huang, J., Hua, X.-S., & Zhang, L. (2020). Structure aware single-stage 3d object detection from point cloud. In *CVPR*, pp. 11873–11882.

He, T., & Soatto, S. (2019). Mono3d++: Monocular 3d vehicle detection with two-scale 3d hypotheses and task priors. In *AAAI*, Vol. 33, pp. 8409–8416.

He, X., Yang, F., Lin, J., Fu, H., Yuan, J., Yang, K., & Li, Z. (2023). Ssd-monodtr: Supervised scale-constrained deformable transformer for monocular 3d object detection. *arXiv preprint arXiv:2305.07270*.

Hoiem, D., Divvala, S. K., & Hays, J. H. (2009). Pascal voc 2008 challenge. *World Literature Today*, *24*.

Hong, Y., Dai, H., & Ding, Y. (2022). Cross-modality knowledge distillation network for monocular 3d object detection. In *ECCV*, pp. 87–104. Springer.

Hu, J. S., Kuai, T., & Waslander, S. L. (2022). Point density-aware voxels for lidar 3d object detection. In *CVPR*, pp. 8469–8478.

Huang, K., Shi, B., Li, X., Li, X., Huang, S., & Li, Y. (2022a). Multi-modal sensor fusion for auto driving perception: A survey. *arXiv preprint arXiv:2202.02703*.

Huang, K.-C., Wu, T.-H., Su, H.-T., & Hsu, W. H. (2022b). Monodtr: Monocular 3d object detection with depth-aware transformer. In *CVPR*, pp. 4012–4021.

Huang, L., Li, Z., Sima, C., Wang, W., Wang, J., Qiao, Y., & Li, H. (2023). Leveraging vision-centric multi-modal expertise for 3d object detection. *arXiv preprint arXiv:2310.15670*.

Huang, T., Liu, Z., Chen, X., & Bai, X. (2020). Epnet: Enhancing point features with image semantics for 3d object detection. In *ECCV*, pp. 35–52. Springer.

Huang, X., Wang, P., Cheng, X., Zhou, D., Geng, Q., & Yang, R. (2019). The apolloscape open dataset for autonomous driving and its application. *IEEE transactions on pattern analysis and machine intelligence*, *42*(10), 2702–2719.

Hwang, J.-J., Kretzschmar, H., Manela, J., Rafferty, S., Armstrong-Crews, N., Chen, T., & Anguelov, D. (2022). Cramnet: Camera-radar fusion with ray-constrained cross-attention for robust 3d object detection. In *ECCV*, pp. 388–405. Springer.

Jiang, X., Li, S., Liu, Y., Wang, S., Jia, F., Wang, T., Han, L., & Zhang, X. (2023). Far3d: Expanding the horizon for surround-view 3d object detection. *arXiv preprint arXiv:2308.09616*.

Jiang, Y., Zhang, L., Miao, Z., Zhu, X., Gao, J., Hu, W., & Jiang, Y.-G. (2022). Polarformer: Multi-camera 3d object detection with polar transformers. *arXiv preprint arXiv:2206.15398*.

Jiang, Y., Zhang, L., Miao, Z., Zhu, X., Gao, J., Hu, W., & Jiang, Y.-G. (2023). Polarformer: Multi-camera 3d object detection with polar transformer. In *AAAI*, Vol. 37, pp. 1042–1050.

Jiao, Y., Jie, Z., Chen, S., Chen, J., Wei, X., Ma, L., & Jiang, Y.-G. (2022). Msmdfusion: Fusing lidar and camera at multiple scales with multi-depth seeds for 3d object detection. *arXiv preprint arXiv:2209.03102*.

Jiao, Y., Jie, Z., Chen, S., Cheng, L., Chen, J., Ma, L., & Jiang, Y.-G. (2023). Instance-aware multi-camera 3d object detection with structural priors mining and self-boosting learning. *arXiv preprint arXiv:2312.08004*.

Ju, B., Zou, Z., Ye, X., Jiang, M., Tan, X., Ding, E., & Wang, J. (2022). Paint and distill: Boosting 3d object detection with semantic passing network. In *Proceedings of the 30th ACM International Conference on Multimedia*, pp. 5639–5648.

Khoche, A., Sánchez, L. P., Batool, N., Mansouri, S. S., & Jensfelt, P. (2023). Fully sparse long range 3d object detection using range experts and multimodal virtual points. *arXiv preprint arXiv:2310.04800*.

Kim, C., Kim, U.-H., & Kim, J.-H. (2022). Self-supervised 3d object detection from monocular pseudo-lidar. In *2022 IEEE International Conference on Multisensor Fusion and Integration for Intelligent Systems (MFI)*, pp. 1–6. IEEE.

Kim, Y., Kim, S., Shin, J., Choi, J. W., & Kum, D. (2023). Crn: Camera radar net for accurate, robust, efficient 3d perception. *arXiv preprint arXiv:2304.00670*.

Klingner, M., Borse, S., Kumar, V. R., Rezaei, B., Narayanan, V., Yogamani, S., & Porikli, F. (2023). X^3KD: Knowledge distillation across modalities, tasks and stages for multi-camera 3d object detection. *arXiv preprint arXiv:2303.02203*.

Koh, J., Lee, J., Lee, Y., Kim, J., & Choi, J. W. (2022). Mgtanet: Encoding sequential lidar points using long short-term motion-guided temporal attention for 3d object detection. *arXiv preprint arXiv:2212.00442*.

Königshof, H., Salscheider, N. O., & Stiller, C. (2019). Realtime 3d object detection for automated driving using stereo vision and semantic information. In *ITSC*, pp. 1405–1410. IEEE.

Koo, I., Lee, I., Kim, S.-H., Kim, H.-S., Jeon, W.-j., & Kim, C. (2023). Pg-rcnn: Semantic surface point generation for 3d object detection. In *ICCV*, pp. 18142–18151.

Ku, J., Mozifian, M., Lee, J., Harakeh, A., & Waslander, S. L. (2018). Joint 3d proposal generation and object detection from view aggregation. In *IROS*, pp. 1–8. IEEE.

Kundu, A., Li, Y., & Rehg, J. M. (2018). 3d-rcnn: Instance-level 3d object reconstruction via render-and-compare. In *CVPR*, pp. 3559–3568.

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*, pp. 12697–12705.

Le, D. T., Shi, H., Rezatofighi, H., & Cai, J. (2022). Accurate and real-time 3d pedestrian detection using an efficient attentive pillar network. *IEEE Robotics and Automation Letters*, *8*(2), 1159–1166.

Li, B. (2017). 3d fully convolutional network for vehicle detection in point cloud. In *IROS*, pp. 1513–1518. IEEE.

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. (2019). Gs3d: An efficient 3d object detection framework for autonomous driving. In *CVPR*, pp. 1019–1028.

Li, C., Ku, J., & Waslander, S. L. (2020). Confidence guided stereo 3d object detection with split depth estimation. In *IROS*, pp. 5776–5783. IEEE.

Li, J., Dai, H., Shao, L., & Ding, Y. (2021). Anchor-free 3d single stage detector with mask-guided attention for point cloud. In *Proceedings of the 29th ACM International Conference on Multimedia*, pp. 553–562.

Li, J., Luo, C., & Yang, X. (2023). Pillarnext: Rethinking network designs for 3d object detection in lidar point clouds. *arXiv preprint arXiv:2305.04925*.

Li, P., Chen, X., & Shen, S. (2019). Stereo r-cnn based 3d object detection for autonomous driving. In *CVPR*, pp. 7644–7652.

Li, X., Ma, T., Hou, Y., Shi, B., Yang, Y., Liu, Y., Wu, X., Chen, Q., Li, Y., Qiao, Y., et al. (2023). Logonet: Towards accurate 3d object detection with local-to-global cross-modal fusion. *arXiv preprint arXiv:2303.03595*.

Li, X., Shi, B., Hou, Y., Wu, X., Ma, T., Li, Y., & He, L. (2022). Homogeneous multi-modal feature fusion and interaction for 3d object detection. In *ECCV*, pp. 691–707. Springer.

Li, Y., Xu, S., Lin, M., Yin, J., Zhang, B., & Cao, X. (2023a). Representation disparity-aware distillation for 3d object detection. In *ICCV*, pp. 6715–6724.

Li, Y., Qi, C. R., Zhou, Y., Liu, C., & Anguelov, D. (2023b). Modar: Using motion forecasting for 3d object detection in point cloud sequences. In *CVPR*, pp. 9329–9339.

Li, Y., Yu, A. W., Meng, T., Caine, B., Ngiam, J., Peng, D., Shen, J., Lu, Y., Zhou, D., Le, Q. V., et al. (2022). Deepfusion: Lidar-camera deep fusion for multi-modal 3d object detection. In *CVPR*, pp. 17182–17191.

Li, Y., Ge, Z., Yu, G., Yang, J., Wang, Z., Shi, Y., Sun, J., & Li, Z. (2023). Bevdepth: Acquisition of reliable depth for multi-view 3d object detection. In *AAAI*, Vol. 37, pp. 1477–1485.

Li, Z., Wang, F., & Wang, N. (2021). Lidar r-cnn: An efficient and universal 3d object detector. In *CVPR*, pp. 7546–7555.

Li, Z., Wang, W., Li, H., Xie, E., Sima, C., Lu, T., Qiao, Y., & Dai, J. (2022). Bevformer: Learning bird's-eye-view representation from multi-camera images via spatiotemporal transformers. In *ECCV*, pp. 1–18. Springer.

Li, Z., Yu, Z., Wang, W., Anandkumar, A., Lu, T., & Alvarez, J. M. (2023a). Fb-bev: Bev representation from forward-backward view transformations. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 6919–6928.

Li, Z., Zhang, C., Ma, W.-C., Zhou, Y., Huang, L., Wang, H., Lim, S., & Zhao, H. (2023b). Voxelformer: Bird's-eye-view feature generation based on dual-view attention for multi-view 3d object detection. *arXiv preprint arXiv:2304.01054*.

Li, Z., Guo, J., Cao, T., Bingbing, L., & Yang, W. (2023c). Gpa-3d: Geometry-aware prototype alignment for unsupervised domain adaptive 3d object detection from point clouds. In *ICCV*, pp. 6394–6403.

Lian, Q., Li, P., & Chen, X. (2022). Monojsg: Joint semantic and geometric cost volume for monocular 3d object detection. In *CVPR*, pp. 1070–1079.

Liang, M., Yang, B., Chen, Y., Hu, R., & Urtasun, R. (2019). Multi-task multi-sensor fusion for 3d object detection. In *CVPR*, pp. 7345–7353.

Liang, M., Yang, B., Wang, S., & Urtasun, R. (2018). Deep continuous fusion for multi-sensor 3d object detection. In *ECCV*, pp. 641–656.

Liang, Z., Zhang, Z., Zhang, M., Zhao, X., & Pu, S. (2021). Rangeioudet: Range image based real-time 3d object detector optimized by intersection over union. In *CVPR*, pp. 7140–7149.

Lis, K., & Kryjak, T. (2023). Pointpillars backbone type selection for fast and accurate lidar object detection. In *ICCVG 2022*, pp. 99–119.

Liu, C., Gao, C., Liu, F., Li, P., Meng, D., & Gao, X. (2023a). Hierarchical supervision and shuffle data augmentation for 3d semi-supervised object detection. *arXiv preprint arXiv:2304.01464*.

Liu, H., Teng, Y., Lu, T., Wang, H., & Wang, L. (2023b). Sparsebev: High-performance sparse 3d object detection from multi-camera videos. In *ICCV*, pp. 18580–18590.

Liu, Y.-C., Ma, C.-Y., & Kira, Z. (2022a). Unbiased teacher v2: Semi-supervised object detection for anchor-free and anchor-based detectors. In *CVPR*, pp. 9819–9828.

Liu, Y., Wang, T., Zhang, X., & Sun, J. (2022b). Petr: Position embedding transformation for multi-view 3d object detection. In *ECCV*, pp. 531–548. Springer.

Liu, Y., Wang, L., & Liu, M. (2021). Yolostereo3d: A step back to 2d for efficient stereo 3d detection. In *ICRA*, pp. 13018–13024. IEEE.

Liu, Z., Wu, Z., & Tóth, R. (2020). Smoke: Single-stage monocular 3d object detection via keypoint estimation. In *CVPR Workshops*, pp. 996–997.

Liu, Z., Huang, T., Li, B., Chen, X., Wang, X., & Bai, X. (2022a). Epnet++: Cascade bi-directional fusion for multi-modal 3d object detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

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*.

Liu, Z., Tang, H., Lin, Y., & Han, S. (2019). Point-voxel cnn for efficient 3d deep learning. *Advances in Neural Information Processing Systems*, *32*.

Lu, D., Xie, Q., Wei, M., Gao, K., Xu, L., & Li, J. (2022a). Transformers in 3d point clouds: A survey. *arXiv preprint arXiv:2205.07417*.

Lu, J., Zhou, Z., Zhu, X., Xu, H., & Zhang, L. (2022b). Learning ego 3d representation as ray tracing. In *ECCV*, pp. 129–144. Springer.

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*, pp. 3111–3121.

Luo, S., Dai, H., Shao, L., & Ding, Y. (2021). M3dssd: Monocular 3d single stage object detector. In *CVPR*, pp. 6145–6154.

Ma, T., Yang, X., Zhou, H., Li, X., Shi, B., Liu, J., Yang, Y., Liu, Z., He, L., Qiao, Y., et al. (2023). Detzero: Rethinking offboard 3d object detection with long-term sequential point clouds. *arXiv preprint arXiv:2306.06023*.

Ma, X., Liu, S., Xia, Z., Zhang, H., Zeng, X., & Ouyang, W. (2020). Rethinking pseudo-lidar representation. In *ECCV*, pp. 311–327. Springer.

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, X., Wang, Z., Li, H., Zhang, P., Ouyang, W., & Fan, X. (2019). Accurate monocular 3d object detection via color-embedded 3d reconstruction for autonomous driving. In *ICCV*, pp. 6851–6860.

Manhardt, F., Kehl, W., & Gaidon, A. (2019). Roi-10d: Monocular lifting of 2d detection to 6d pose and metric shape. In *CVPR*, pp. 2069–2078.

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*, pp. 2723–2732.

Mao, J., Xue, Y., Niu, M., Bai, H., Feng, J., Liang, X., Xu, H., & Xu, C. (2021b). Voxel transformer for 3d object detection. In *ICCV*, pp. 3164–3173.

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 Robotics and Automation Letters*, *6*(2), 526–533.

Meyer, G. P., Laddha, A., Kee, E., Vallespi-Gonzalez, C., & Wellington, C. K. (2019). Lasernet: An efficient probabilistic 3d object detector for autonomous driving. In *CVPR*, pp. 12677–12686.

Mohapatra, S., Yogamani, S., Gotzig, H., Milz, S., & Mader, P. (2021). Bevdetnet: bird's eye view lidar point cloud based real-time 3d object detection for autonomous driving. In *2021 IEEE International Intelligent Transportation Systems Conference (ITSC)*, pp. 2809–2815. IEEE.

Nabati, R., & Qi, H. (2021). Centerfusion: Center-based radar and camera fusion for 3d object detection. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 1527–1536.

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*, pp. 61–65. IEEE.

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*.

Paek, D.-H., Kong, S.-H., & Wijaya, K. T. (2022a). K-radar: 4d radar object detection for autonomous driving in various weather conditions. In *NeurIPS*.

Paek, D.-H., Kong, S.-H., & Wijaya, K. T. (2022b). K-radar: 4d radar object detection for autonomous driving in various weather conditions. *Advances in Neural Information Processing Systems*, *35*, 3819–3829.

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*, pp. 2926–2933.

Palmer, P., Krueger, M., Altendorfer, R., Adam, G., & Bertram, T. (2023). Reviewing 3d object detectors in the context of high-resolution 3+ 1d radar. *arXiv preprint arXiv:2308.05478*.

Pan, X., Xia, Z., Song, S., Li, L. E., & Huang, G. (2021). 3d object detection with pointformer. In *CVPR*, pp. 7463–7472.

Pan, Z., Ding, F., Zhong, H., & Lu, C. X. (2023a). Moving object detection and tracking with 4d radar point cloud. *arXiv preprint arXiv:2309.09737*.

Pan, Z., Ding, F., Zhong, H., & Lu, C. X. (2023b). Moving object detection and tracking with 4d radar point cloud. *arXiv preprint arXiv:2309.09737*.

Pang, S., Morris, D., & Radha, H. (2020). Clocs: Camera-lidar object candidates fusion for 3d object detection. In *IROS*, pp. 10386–10393. IEEE.

Pang, S., Morris, D., & Radha, H. (2022). Fast-clocs: Fast camera-lidar object candidates fusion for 3d object detection. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 187–196.

Park, D., Ambrus, R., Guizilini, V., Li, J., & Gaidon, A. (2021). Is pseudo-lidar needed for monocular 3d object detection?. In *ICCV*, pp. 3142–3152.

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*, pp. 9552–9557. IEEE.

Peng, L., Wu, X., Yang, Z., Liu, H., & Cai, D. (2022). Did-m3d: Decoupling instance depth for monocular 3d object detection. In *ECCV*, pp. 71–88. Springer.

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*, pp. 13015–13024.

Peng, X., Zhu, X., & Ma, Y. (2023). Cl3d: Unsupervised domain adaptation for cross-lidar 3d detection. In *AAAI*, Vol. 37, pp. 2047–2055.

Peng, X., Zhu, X., Wang, T., & Ma, Y. (2022). Side: Center-based stereo 3d detector with structure-aware instance depth estimation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 119–128.

Piroli, A., Dallabetta, V., Kopp, J., Walessa, M., Meissner, D., & Dietmayer, K. (2023). Towards robust 3d object detection in rainy conditions. *arXiv preprint arXiv:2310.00944*.

Pon, A. D., Ku, J., Li, C., & Waslander, S. L. (2020). Object-centric stereo matching for 3d object detection. In *ICRA*, pp. 8383–8389. IEEE.

Qi, C. R., Litany, O., He, K., & Guibas, L. J. (2019). Deep hough voting for 3d object detection in point clouds. In *ICCV*, pp. 9277–9286.

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*, pp. 918–927.

Qi, C. R., Su, H., Mo, K., & Guibas, L. J. (2017a). Pointnet: Deep learning on point sets for 3d classification and segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 652–660.

Qi, C. R., Yi, L., Su, H., & Guibas, L. J. (2017b). Pointnet++: Deep hierarchical feature learning on point sets in a metric space. *Advances in neural information processing systems*, *30*.

Qian, R., Lai, X., & Li, X. (2022). 3d object detection for autonomous driving: a survey. *Pattern Recognition*, *130*, 108796.

Qin, Y., Wang, C., Kang, Z., Ma, N., Li, Z., & Zhang, R. (2023a). Supfusion: Supervised lidar-camera fusion for 3d object detection. In *ICCV*, pp. 22014–22024.

Qin, Z., Chen, J., Chen, C., Chen, X., & Li, X. (2023b). Unifusion: Unified multi-view fusion transformer for spatial-temporal representation in bird's-eye-view. In *ICCV*, pp. 8690–8699.

Qu, D., Chen, Q., Bai, T., Qin, A., Lu, H., Fan, H., Fu, S., & Yang, Q. (2023). Sicp: Simultaneous individual and cooperative perception for 3d object detection in connected and automated vehicles. *arXiv preprint arXiv:2312.04822*.

Răduţoiu, A., Schulze, J.-P., Sperl, P., & Böttinger, K. (2023). Physical adversarial examples for multi-camera systems. *arXiv preprint arXiv:2311.08539*.

Reading, C., Harakeh, A., Chae, J., & Waslander, S. L. (2021). Categorical depth distribution network for monocular 3d object detection. In *CVPR*, pp. 8555–8564.

Ren, B., & Yin, J. (2023). Sdvrf: Sparse-to-dense voxel region fusion for multi-modal 3d object detection. *arXiv preprint arXiv:2304.08304*.

Roh, W., Chang, G., Moon, S., Nam, G., Kim, C., Kim, Y., Kim, J., & Kim, S. (2022a). Ora3d: Overlap region aware multi-view 3d object detection. *arXiv preprint arXiv:2207.00865*.

Roh, W., Chang, G., Moon, S., Nam, G., Kim, C., Kim, Y., Kim, S., & Kim, J. (2022b). Ora3d: Overlap region aware multi-view 3d object detection. *arXiv preprint arXiv:2207.00865*.

Rukhovich, D., Vorontsova, A., & Konushin, A. (2022). Imvoxelnet: Image to voxels projection for monocular and multi-view general-purpose 3d object detection. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 2397–2406.

Šebek, P., Pokornỳ, Š., Vacek, P., & Svoboda, T. (2022). Real3d-aug: Point cloud augmentation by placing real objects with occlusion handling for 3d detection and segmentation. *arXiv preprint arXiv:2206.07634*.

Shan, Y., Xia, Y., Chen, Y., & Cremers, D. (2023). Scp: Scene completion pre-training for 3d object detection. *arXiv preprint arXiv:2309.06199*.

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*, pp. 2743–2752.

Sheng, Y., Ye, H., Liang, L., Jin, S., & Li, G. Y. (2023). Semantic communication for cooperative perception based on importance map. *arXiv preprint arXiv:2311.06498*.

Shi, G., Li, R., & Ma, C. (2022). Pillarnet: High-performance pillar-based 3d object detection. *arXiv preprint arXiv:2205.07403*.

Shi, H., Pang, C., Zhang, J., Yang, K., Wu, Y., Ni, H., Lin, Y., Stiefelhagen, R., & Wang, K. (2023). Cobev: Elevating roadside 3d object detection with depth and height complementarity. *arXiv preprint arXiv:2310.02815*.

Shi, S., Wang, X., Li, H. P., et al. (2019). 3d object proposal generation and detection from point cloud. In *CVPR, Long Beach, CA, USA*, pp. 16–20.

Shi, S., Guo, C., Jiang, L., Wang, Z., Shi, J., Wang, X., & Li, H. (2020). Pv-rcnn: Point-voxel feature set abstraction for 3d object detection. In *CVPR*, pp. 10529–10538.

Shi, S., Jiang, L., Deng, J., Wang, Z., Guo, C., Shi, J., Wang, X., & Li, H. (2023). Pv-rcnn++: Point-voxel feature set abstraction with local vector representation for 3d object detection. *International Journal of Computer Vision*, *131*(2), 531–551.

Shi, W., & Rajkumar, R. (2020). Point-gnn: Graph neural network for 3d object detection in a point cloud. In *CVPR*, pp. 1711–1719.

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 *CVPR Workshops*, pp. 0–0.

Sun, P., Kretzschmar, H., Dotiwalla, X., Chouard, A., Patnaik, V., Tsui, P., Guo, J., Zhou, Y., Chai, Y., Caine, B., et al. (2020). Scalability in perception for autonomous driving: Waymo open dataset. In *CVPR*, pp. 2446–2454.

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*, pp. 5725–5734.

Tan, J., Lyu, F., Li, L., Hu, F., Feng, T., Xu, F., & Yao, R. (2023). Dynamic v2x autonomous perception from road-to-vehicle vision. *arXiv preprint arXiv:2310.19113*.

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*, pp. 685–702. Springer.

Tian, Z., Shen, C., Chen, H., & He, T. (2019). Fcos: Fully convolutional one-stage object detection. In *ICCV*, pp. 9627–9636.

Tseng, C.-Y., Chen, Y.-R., Lee, H.-Y., Wu, T.-H., Chen, W.-C., & Hsu, W. H. (2023). Crossdtr: Cross-view and depth-guided transformers for 3d object detection. In *ICRA*, pp. 4850–4857. IEEE.

Veit, A., Matera, T., Neumann, L., Matas, J., & Belongie, S. (2016). Coco-text: Dataset and benchmark for text detection and recognition in natural images. *arXiv preprint arXiv:1601.07140*.

Vora, S., Lang, A. H., Helou, B., & Beijbom, O. (2020). Pointpainting: Sequential fusion for 3d object detection. In *CVPR*, pp. 4604–4612.

Wang, C., Ma, C., Zhu, M., & Yang, X. (2021a). Pointaugmenting: Cross-modal augmentation for 3d object detection. In *CVPR*, pp. 11794–11803.

Wang, C., Ma, C., Zhu, M., & Yang, X. (2021b). Pointaugmenting: Cross-modal augmentation for 3d object detection. In *CVPR*, pp. 11794–11803.

Wang, D. Z., & Posner, I. (2015). Voting for voting in online point cloud object detection.. In *Robotics: science and systems*, Vol. 1, pp. 10–15. Rome, Italy.

Wang, H., Tang, H., Shi, S., Li, A., Li, Z., Schiele, B., & Wang, L. (2023). Unitr: A unified and efficient multi-modal transformer for bird's-eye-view representation. In *ICCV*, pp. 6792–6802.

Wang, J., Ma, Y., Huang, S., Hui, T., Wang, F., Qian, C., & Zhang, T. (2022). A keypoint-based global association network for lane detection. In *CVPR*, pp. 1392–1401.

Wang, L., Du, L., Ye, X., Fu, Y., Guo, G., Xue, X., Feng, J., & Zhang, L. (2021a). Depth-conditioned dynamic message propagation for monocular 3d object detection. In *CVPR*, pp. 454–463.

Wang, L., Zhang, L., Zhu, Y., Zhang, Z., He, T., Li, M., & Xue, X. (2021b). Progressive coordinate transforms for monocular 3d object detection. *Advances in Neural Information Processing Systems*, *34*, 13364–13377.

Wang, S., Liu, Y., Wang, T., Li, Y., & Zhang, X. (2023a). Exploring object-centric temporal modeling for efficient multi-view 3d object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 3621–3631.

Wang, S., Zhao, X., Xu, H.-M., Chen, Z., Yu, D., Chang, J., Yang, Z., & Zhao, F. (2023b). Towards domain generalization for multi-view 3d object detection in bird-eye-view. *arXiv preprint arXiv:2303.01686*.

Wang, T., Pang, J., & Lin, D. (2022a). Monocular 3d object detection with depth from motion. In *ECCV*, pp. 386–403. Springer.

Wang, T., Xinge, Z., Pang, J., & Lin, D. (2022b). Probabilistic and geometric depth: Detecting objects in perspective. In *Conference on Robot Learning*, pp. 1475–1485. PMLR.

Wang, T., Zhu, X., Pang, J., & Lin, D. (2021). Fcos3d: Fully convolutional one-stage monocular 3d object detection. In *ICCV*, pp. 913–922.

Wang, X., & Kitani, K. M. (2023). Cost-aware evaluation and model scaling for lidar-based 3d object detection. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 9260–9267. IEEE.

Wang, Y., Chao, W.-L., Garg, D., Hariharan, B., Campbell, M., & Weinberger, K. Q. (2019). Pseudo-lidar from visual depth estimation: Bridging the gap in 3d object detection for autonomous driving. In *CVPR*, pp. 8445–8453.

Wang, Y., Yin, J., Li, W., Frossard, P., Yang, R., & Shen, J. (2022). Ssda3d: Semi-supervised domain adaptation for 3d object detection from point cloud. *arXiv preprint arXiv:2212.02845*.

Wang, Y., Deng, J., Li, Y., Hu, J., Liu, C., Zhang, Y., Ji, J., Ouyang, W., & Zhang, Y. (2023a). Bi-lrfusion: Bi-directional lidar-radar fusion for 3d dynamic object detection. In *CVPR*, pp. 13394–13403.

Wang, Y., Mao, Q., Zhu, H., Deng, J., Zhang, Y., Ji, J., Li, H., & Zhang, Y. (2023b). Multi-modal 3d object detection in autonomous driving: a survey. *International Journal of Computer Vision*, 1–31.

Wang, Y., Fathi, A., Kundu, A., Ross, D. A., Pantofaru, C., Funkhouser, T., & Solomon, J. (2020). Pillar-based object detection for autonomous driving. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXII 16*, pp. 18–34.

Wang, Y., Guizilini, V. C., Zhang, T., Wang, Y., Zhao, H., & Solomon, J. (2022). Detr3d: 3d object detection from multi-view images via 3d-to-2d queries. In *Conference on Robot Learning*, pp. 180–191. PMLR.

Wang, Z., Li, D., Luo, C., Xie, C., & Yang, X. (2023a). Distillbev: Boosting multi-camera 3d object detection with cross-modal knowledge distillation. In *ICCV*, pp. 8637–8646.

Wang, Z., Li, Y., Chen, X., Zhao, H., & Wang, S. (2023b). Uni3detr: Unified 3d detection transformer. *Advances in Neural Information Processing Systems*.

Wanga, S., & Zheng, J. (2023). Monoskd: General distillation framework for monocular 3d object detection via spearman correlation coefficient. *arXiv preprint arXiv:2310.11316*.

Wu, H., Wen, C., Li, W., Li, X., Yang, R., & Wang, C. (2022). Transformation-equivariant 3d object detection for autonomous driving. *arXiv preprint arXiv:2211.11962*.

Wu, H., Wen, C., Shi, S., Li, X., & Wang, C. (2023). Virtual sparse convolution for multimodal 3d object detection. In *CVPR*, pp. 21653–21662.

Wu, X., Peng, L., Yang, H., Xie, L., Huang, C., Deng, C., Liu, H., & Cai, D. (2022). Sparse fuse dense: Towards high quality 3d detection with depth completion. In *CVPR*, pp. 5418–5427.

Wu, Z., Chen, G., Gan, Y., Wang, L., & Pu, J. (2023a). Mvfusion: Multi-view 3d object detection with semantic-aligned radar and camera fusion. *arXiv preprint arXiv:2302.10511*.

Wu, Z., Gan, Y., Wang, L., Chen, G., & Pu, J. (2023b). Monopgc: Monocular 3d object detection with pixel geometry contexts. *arXiv preprint arXiv:2302.10549*.

Wu, Z., Wu, Y., Pu, J., Li, X., & Wang, X. (2022). Attention-based depth distillation with 3d-aware positional encoding for monocular 3d object detection. *arXiv preprint arXiv:2211.16779*.

Xiang, L., Yin, J., Li, W., Xu, C.-Z., Yang, R., & Shen, J. (2023). Di-v2x: Learning domain-invariant representation for vehicle-infrastructure collaborative 3d object detection. *arXiv preprint arXiv:2312.15742*.

Xiang, Y., Choi, W., Lin, Y., & Savarese, S. (2015). Data-driven 3d voxel patterns for object category recognition. In *CVPR*, pp. 1903–1911.

Xiang, Y., Choi, W., Lin, Y., & Savarese, S. (2017). Subcategory-aware convolutional neural networks for object proposals and detection. In *2017 IEEE winter conference on applications of computer vision (WACV)*, pp. 924–933. IEEE.

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*, pp. 3095–3101. IEEE.

Xie, L., Xiang, C., Yu, Z., Xu, G., Yang, Z., Cai, D., & He, X. (2020). Pi-rcnn: An efficient multi-sensor 3d object detector with point-based attentive cont-conv fusion module. In *AAAI*, Vol. 34, pp. 12460–12467.

Xie, W., Hu, T., Ling, N., Xing, G., Liu, S., & Guan, N. (2023). Timely fusion of surround radar/lidar for object detection in autonomous driving systems. *Design, Automation and Test in Europe*.

Xiong, K., Gong, S., Ye, X., Tan, X., Wan, J., Ding, E., Wang, J., & Bai, X. (2023). Cape: Camera view position embedding for multi-view 3d object detection. *arXiv preprint arXiv:2303.10209*.

Xu, D., Anguelov, D., & Jain, A. (2018). Pointfusion: Deep sensor fusion for 3d bounding box estimation. In *CVPR*, pp. 244–253.

Xu, H., Liu, F., Zhou, Q., Hao, J., Cao, Z., Feng, Z., & Ma, L. (2021). Semi-supervised 3d object detection via adaptive pseudo-labeling. In *ICIP*, pp. 3183–3187. IEEE.

Xu, J., Peng, L., Cheng, H., Li, H., Qian, W., Li, K., Wang, W., & Cai, D. (2023a). Mononerd: Nerf-like representations for monocular 3d object detection. In *ICCV*, pp. 6814–6824.

Xu, J., Peng, L., Cheng, H., Xia, L., Zhou, Q., Deng, D., Qian, W., Wang, W., & Cai, D. (2023b). Regulating intermediate 3d features for vision-centric autonomous driving. *arXiv preprint arXiv:2312.11837*.

Xu, Q., Zhong, Y., & Neumann, U. (2022a). Behind the curtain: Learning occluded shapes for 3d object detection. In *AAAI*, Vol. 36, pp. 2893–2901.

Xu, Q., Zhong, Y., & Neumann, U. (2022b). Behind the curtain: Learning occluded shapes for 3d object detection. In *AAAI*, Vol. 36, pp. 2893–2901.

Xu, R., Wang, T., Zhang, W., Chen, R., Cao, J., Pang, J., & Lin, D. (2023). Mv-jar: Masked voxel jigsaw and reconstruction for lidar-based self-supervised pre-training. *arXiv preprint arXiv:2303.13510*.

Xu, S., Zhou, D., Fang, J., Yin, J., Bin, Z., & Zhang, L. (2021). Fusionpainting: Multimodal fusion with adaptive attention for 3d object detection. In *ITSC*, pp. 3047–3054. IEEE.

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*, Vol. 34, pp. 12557–12564.

Yan, C., & Salman, E. (2017). Mono3d: Open source cell library for monolithic 3-d integrated circuits. *IEEE Transactions on Circuits and Systems I: Regular Papers*, *65*(3), 1075–1085.

Yan, F., Nie, M., Cai, X., Han, J., Xu, H., Yang, Z., Ye, C., Fu, Y., Mi, M. B., & Zhang, L. (2022). Once-3dlanes: Building monocular 3d lane detection. In *CVPR*, pp. 17143–17152.

Yang, B., Liang, M., & Urtasun, R. (2018a). Hdnet: Exploiting hd maps for 3d object detection. In *Conference on Robot Learning*, pp. 146–155. PMLR.

Yang, B., Luo, W., & Urtasun, R. (2018b). Pixor: Real-time 3d object detection from point clouds. In *CVPR*, pp. 7652–7660.

Yang, B., & Ji, X. (2023). Exploring adversarial robustness of lidar-camera fusion model in autonomous driving. *arXiv preprint arXiv:2312.01468*.

Yang, C., Chen, Y., Tian, H., Tao, C., Zhu, X., Zhang, Z., Huang, G., Li, H., Qiao, Y., Lu, L., et al. (2023a). Bevformer v2: Adapting modern image backbones to bird's-eye-view recognition via perspective supervision. In *CVPR*, pp. 17830–17839.

Yang, H., Wang, W., Chen, M., Lin, B., He, T., Chen, H., He, X., & Ouyang, W. (2023b). Pvt-ssd: Single-stage 3d object detector with point-voxel transformer. *arXiv preprint arXiv:2305.06621*.

Yang, J., Shi, S., Wang, Z., Li, H., & Qi, X. (2021). St3d: Self-training for unsupervised domain adaptation on 3d object detection. In *CVPR*, pp. 10368–10378.

Yang, K., Yang, D., Zhang, J., Li, M., Liu, Y., Liu, J., Wang, H., Sun, P., & Song, L. (2023a). Spatio-temporal domain awareness for multi-agent collaborative perception. In *ICCV*, pp. 23383–23392.

Yang, L., Tang, T., Li, J., Chen, P., Yuan, K., Wang, L., Huang, Y., Zhang, X., & Yu, K. (2023b). Bevheight++: Toward robust visual centric 3d object detection. *arXiv preprint arXiv:2309.16179*.

Yang, L., Yu, J., Zhang, X., Li, J., Wang, L., Huang, Y., Zhang, C., Wang, H., & Li, Y. (2023c). Monogae: Roadside monocular 3d object detection with ground-aware embeddings. *arXiv preprint arXiv:2310.00400*.

Yang, L., Yu, K., Tang, T., Li, J., Yuan, K., Wang, L., Zhang, X., & Chen, P. (2023d). Bevheight: A robust framework for vision-based roadside 3d object detection. In *CVPR*, pp. 21611–21620.

Yang, Y., Liu, J., Huang, T., Han, Q.-L., Ma, G., & Zhu, B. (2022). Ralibev: Radar and lidar bev fusion learning for anchor box free object detection system. *arXiv preprint arXiv:2211.06108*.

Yang, Z., Sun, Y., Liu, S., & Jia, J. (2020). 3dssd: Point-based 3d single stage object detector. In *CVPR*, pp. 11040–11048.

Yang, Z., Sun, Y., Liu, S., Shen, X., & Jia, J. (2018). 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*, pp. 1951–1960.

Yao, S., Guan, R., Huang, X., Li, Z., Sha, X., Yue, Y., Lim, E. G., Seo, H., Man, K. L., Zhu, X., et al. (2023). Radar-camera fusion for object detection and semantic segmentation in autonomous driving: A comprehensive review. *arXiv preprint arXiv:2304.10410*.

Ye, M., Xu, S., & Cao, T. (2020). Hvnet: Hybrid voxel network for lidar based 3d object detection. In *CVPR*, pp. 1631–1640.

Yin, J., Zhou, D., Zhang, L., Fang, J., Xu, C.-Z., Shen, J., & Wang, W. (2022). Proposalcontrast: Unsupervised pre-training for lidar-based 3d object detection. In *ECCV*, pp. 17–33. Springer.

Yin, T., Zhou, X., & Krahenbuhl, P. (2021a). Center-based 3d object detection and tracking. In *CVPR*, pp. 11784–11793.

Yin, T., Zhou, X., & Krähenbühl, P. (2021b). Multimodal virtual point 3d detection. *Advances in Neural Information Processing Systems*, *34*, 16494–16507.

Yin, Z., Sun, H., Liu, N., Zhou, H., & Shen, J. (2023). Fgfusion: Fine-grained lidar-camera fusion for 3d object detection. *arXiv preprint arXiv:2309.11804*.

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*, pp. 720–736. Springer.

You, Y., Wang, Y., Chao, W.-L., Garg, D., Pleiss, G., Hariharan, B., Campbell, M., & Weinberger, K. Q. (2019). Pseudo-lidar++: Accurate depth for 3d object detection in autonomous driving. *arXiv preprint arXiv:1906.06310*.

Yu, H., Tang, Y., Xie, E., Mao, J., Luo, P., & Nie, Z. (2023). Flow-based feature fusion for vehicle-infrastructure cooperative 3d object detection. *arXiv preprint arXiv:2311.01682*.

Yuan, J., Zhang, B., Yan, X., Chen, T., Shi, B., Li, Y., & Qiao, Y. (2023). Bi3d: Bi-domain active learning for cross-domain 3d object detection. *arXiv preprint arXiv:2303.05886*.

Zhang, W., Liu, D., Ma, C., & Cai, W. (2023a). Odm3d: Alleviating foreground sparsity for enhanced semi-supervised monocular 3d object detection..

Zhang, X., Wang, L., Chen, J., Fang, C., Yang, L., Song, Z., Yang, G., Wang, Y., Zhang, X., & Li, J. (2023b). Dual radar: A multi-modal dataset with dual 4d radar for autononous driving. *International Conference on Neural Information Processing*.

Zhang, Y., Chen, J., & Huang, D. (2022). Cat-det: Contrastively augmented transformer for multi-modal 3d object detection. In *CVPR*, pp. 908–917.

Zhang, Y., Huang, D., & Wang, Y. (2021). Pc-rgnn: Point cloud completion and graph neural network for 3d object detection. In *AAAI*, Vol. 35, pp. 3430–3437.

Zhang, Y., Dong, Z., Yang, H., Lu, M., Tseng, C.-C., Du, Y., Keutzer, K., Du, L., & Zhang, S. (2023). Qd-bev: Quantization-aware view-guided distillation for multi-view 3d object detection. In *ICCV*, pp. 3825–3835.

Zhang, Y., Hu, Q., Xu, G., Ma, Y., Wan, J., & Guo, Y. (2022). Not all points are equal: Learning highly efficient point-based detectors for 3d lidar point clouds. In *CVPR*, pp. 18953–18962.

Zhang, Y., Zhang, Q., Hou, J., Yuan, Y., & Xing, G. (2023). Bidirectional propagation for cross-modal 3d object detection. *arXiv preprint arXiv:2301.09077*.

Zheng, A., Zhang, Y., Zhang, X., Qi, X., & Sun, J. (2022a). Progressive end-to-end object detection in crowded scenes. In *CVPR*, pp. 857–866.

Zheng, L., Ma, Z., Zhu, X., Tan, B., Li, S., Long, K., Sun, W., Chen, S., Zhang, L., Wan, M., et al. (2022b). Tj4dradset: A 4d radar dataset for autonomous driving. In *ITSC*, pp. 493–498. IEEE.

Zheng, T., Huang, Y., Liu, Y., Tang, W., Yang, Z., Cai, D., & He, X. (2022c). Clrnet: Cross layer refinement network for lane detection. In *CVPR*, pp. 898–907.

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*, Vol. 35, pp. 3555–3562.

Zheng, W., Tang, W., Jiang, L., & Fu, C.-W. (2021b). Se-ssd: Self-ensembling single-stage object detector from point cloud. In *CVPR*, pp. 14494–14503.

Zhou, Q., Cao, J., Leng, H., Yin, Y., Kun, Y., & Zimmermann, R. (2024). Sogdet: Semantic-occupancy guided multi-view 3d object detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 38, pp. 7668–7676.

Zhou, X., Hou, J., Yao, T., Liang, D., Liu, Z., Zou, Z., Ye, X., Cheng, J., & Bai, X. (2023). Diffusion-based 3d object detection with random boxes. *arXiv preprint arXiv:2309.02049*.

Zhou, Y., Sun, P., Zhang, Y., Anguelov, D., Gao, J., Ouyang, T., Guo, J., Ngiam, J., & Vasudevan, V. (2020). End-to-end multi-view fusion for 3d object detection in lidar point clouds. In *Conference on Robot Learning*, pp. 923–932. PMLR.

Zhou, Y., & Tuzel, O. (2018). Voxelnet: End-to-end learning for point cloud based 3d object detection. In *CVPR*, pp. 4490–4499.

Zhou, Y., He, Y., Zhu, H., Wang, C., Li, H., & Jiang, Q. (2021). Monoef: Extrinsic parameter free monocular 3d object detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *44*(12), 10114–10128.

Zhou, Y., Liu, Q., Zhu, H., Li, Y., Chang, S., & Guo, M. (2022a). Mogde: Boosting mobile monocular 3d object detection with ground depth estimation. *Advances in Neural Information Processing Systems*, *35*, 2033–2045.

Zhou, Y., Liu, Q., Zhu, H., Li, Y., Chang, S., & Guo, M. (2022b). Mogde: Boosting mobile monocular 3d object detection with ground depth estimation. *Advances in Neural Information Processing Systems*, *35*, 2033–2045.

Zhou, Y., Zhu, H., Liu, Q., Chang, S., & Guo, M. (2023a). Monoatt: Online monocular 3d object detection with adaptive token transformer. *arXiv preprint arXiv:2303.13018*.

Zhou, Z., Lu, J., Zeng, Y., Xu, H., & Zhang, L. (2023b). Suit: Learning significance-guided information for 3d temporal detection. *arXiv preprint arXiv:2307.01807*.

Zhu, M., Ge, L., Wang, P., & Peng, H. (2023a). Monoedge: Monocular 3d object detection using local perspectives. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 643–652.

Zhu, Z., Meng, Q., Wang, X., Wang, K., Yan, L., & Yang, J. (2023b). Curricular object manipulation in lidar-based object detection. *arXiv preprint arXiv:2304.04248*.

Zimmer, W., Creß, C., Nguyen, H. T., & Knoll, A. C. (2023). A9 intersection dataset: All you need for urban 3d camera-lidar roadside perception. *arXiv preprint arXiv:2306.09266*.