applied sciences

MDPI

Review

Singular and Multimodal Techniques of 3D Object Detection: Constraints, Advancements and Research Direction

Tajbia Karim *, Zainal Rasyid Mahayuddin and Mohammad Kamrul Hasan

Center for Artificial Intelligence Technology, Universiti Kebangsaan Malaysia, Bangi 43600, Selangor, Malaysia; zainalr@ukm.edu.my (Z.R.M.); mkhasan@ukm.edu.my (M.K.H.)

* Correspondence: tajbia.karim@gmail.com (T.K.)

Abstract: Two-dimensional object detection techniques can detect multiscale objects in images. However, they lack depth information. Three-dimensional object detection provides the location of the object in the image along with depth information. To provide depth information, 3D object detection involves the application of depth-perceiving sensors such as LiDAR, stereo cameras, RGB-D, RADAR, etc. The existing review articles on 3D object detection techniques are found to be focusing on either a singular modality (e.g., only LiDAR point cloud-based) or a singular application field (e.g., autonomous vehicle navigation). However, to the best of our knowledge, there is no review paper that discusses the applicability of 3D object detection techniques in other fields such as agriculture, robot vision or human activity detection. This study analyzes both singular and multimodal techniques of 3D object detection techniques applied in different fields. A critical analysis comprising strengths and weaknesses of the 3D object detection techniques is presented. The aim of this study is to facilitate future researchers and practitioners to provide a holistic view of 3D object detection techniques. The critical analysis of the singular and multimodal techniques is expected to help the practitioners find the appropriate techniques based on their requirement.

Keywords: 3D object detection; singular technique; multimodal technique

Citation: Karim, T.; Mahayuddin, Z.R.; Hasan, M.K. Singular and Multimodal Techniques of 3D Object Detection: Constraints, Advancements and Research Direction. *Appl. Sci.* **2023**, *13*, 13267. https://doi.org/10.3390/app132413267

Academic Editors: Xianpeng Wang, Andres Alvarez-Meza and David Cárdenas-Peña

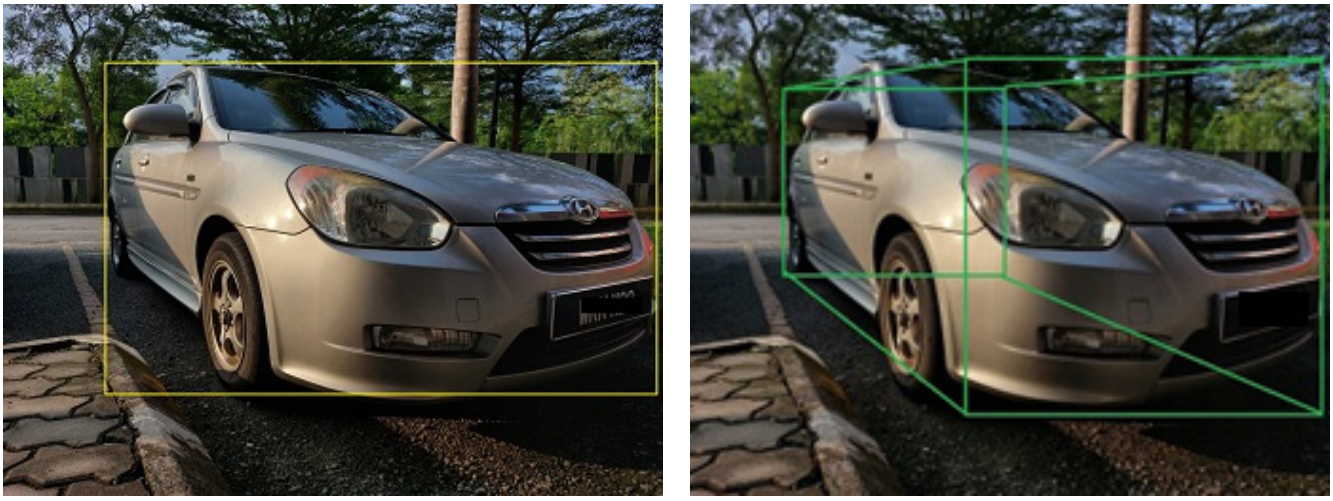
Received: 16 October 2023
Revised: 21 November 2023
Accepted: 30 November 2023
Published: 15 December 2023

1. Introduction

Object detection includes identifying both object class and location. Object detection is essential in versatile fields such as medical diagnosis, industrial production inspection, surveillance, etc. Researchers have gained high precision and real-time speed using deep learning-based 2D object detection methods [1–8]. Convolutional Neural Network (CNN) is the backbone of these algorithms. CNN can detect the pattern and shape of different objects. Although 2D object detection can precisely detect objects, it is deficient in depth and height information. Depth and height information is vital in obstacle avoidance, autonomous vehicle navigation, household robots, medical surgery, etc. However, unlike 2D object detection, 3D object detection is more complex in terms of model training, data availability, annotation and processing. Figure 1 illustrates the difference between 2D and 3D object detection of a car. Figure 1 shows that, in 2D object detection, the objects are detected in two-dimensional form (length and width in the image). On the other hand, in 3D object detection, a prediction of depth is also perceived along with length and width.

The most popularly used datasets for 3D object detection are KITTI [9], SUN RGB-D [10], ScanNet [11], nuScenes [12] and the Waymo Open Dataset [13]. Among the unmanned aerial vehicle (UAV) point of view datasets, Shahbazi et al., 2019, presented a 3D dataset generated using both a multi-view stereo camera and a LASER scanner. The site it covers is a gravel mine [14]. Vélez et al. published a high resolution RGB image dataset of a pistachio orchard in Spain captured from UAV [15]. This dataset is intended to use for 3D Photogrammetric Reconstruction Another large dataset containing human activity is produced by Li et al., 2021, in RGB-D format that can be used for 3D detection [16]. 3D

detection is being applied in medical sector for internal organ disease detection. In the medical sector, 3D CNN is being used on 3D images obtained by Computed Tomography (CT), Diffusion tensor imaging (DTI), magnetic resonance imaging (MRI), Functional magnetic resonance imaging and Ultrasound [17]. But medical imaging differs in terms of data representation and is difficult to compare with other domains.



(**a**) (**b**)

Figure 1. (**a**) Two-dimensional detection of a car; (**b**) 3D detection of a car.

Since computers with GPUs are more available now, it has enhanced the scope of 3D detection with deep learning. Several review papers are found on 3D object detection. For example, Fernandes et al., 2021, presented a survey paper on 3D object detection [18]. This paper solely focuses on LiDAR-generated point cloud-based 3D object detection. Moreover, only self-driving vehicle-related methods are discussed in this paper. Zamanakos et al., 2021, published another survey paper on current trends in 3D object detection [19]. This survey exclusively covers LiDAR-based methods and focuses on the self-driving field. Arnold et al., 2019, published a survey article on 3D object detection techniques obtained by sensors other than LiDAR. This article discusses some fusion methods as well [20]. This research emphasizes 3D detection techniques relevant to autonomous driving. After carefully studying the "3D object detection"-related review papers published in Scopus and Web of Science journals, the contents of the most relevant review papers are noted in Table 1. This visual presentation has paved the way to existing research gaps and new research scopes on 3D object detection techniques. Table 1 contains the relevant review papers.

Three-dimensional object detection techniques are recently gaining popularity due to the necessity of object shape and orientation estimation in real-world space. However, Table 1 clearly shows that, a comprehensive review paper on 3D object detection is still lacking which analyses both singular and multimodal techniques in recently popular application domains. This review article is written with a view to reducing this gap to some extent.

1.1. Paper Motivation

The previous survey papers focus on a single application sector (mostly autonomous vehicle navigation) or a singular modality of 3D object detection. Table 1 clearly depicts the scenario. However, 3D object detection is recently growing in other application domains besides autonomous driving. Precision agriculture, household robots, surveillance services, etc. can benefit from 3D object detection. As 3D object detection is rapidly gaining popularity with researchers, there have been some new discoveries as well which are not included in the previously published review papers.

Table 1. Relevant review articles on 3D object detection.

Ref.	Modality						Application Domain				Content
	Point Cloud		Camera		RADAR	Multi Modal	Autonomous Vehicles	Robot Vision	Precision Agriculture	Human Pose	
	LiDAR	RGB-D	Mono	Stereo							
[21]	H	M	H	L	L	L	H	H	L	L	This research focuses on point cloud and monocular camera-based 3D object detection. However, the scope is limited to autonomous vehicle navigation and robot vision.
[22]	H	M	H	H	M	H	H	L	L	L	This research elaborates different modality (camera, LiDAR, RADAR) performance, but the scope is limited to the autonomous vehicle sector.
[23]	H	H	H	M	L	M	H	H	L	L	This paper explores 3D object detection focusing on autonomous vehicles and indoor robot vision.
[24]	H	L	M	L	L	M	H	L	L	L	This article discusses 3D object detection techniques with point cloud methods, especially in the autonomous vehicle navigation sector. However, the other modalities such as stereo sensors are not discussed.
[25]	H	H	H	M	L	M	H	H	L	M	This research focuses on 3D object detection techniques based on point cloud and monocular camera.
[26]	L	H	H	M	L	M	L	H	L	L	This research discusses 3D object detection techniques for room shape assessment.
[27]	H	L	L	L	L	M	L	M	L	H	This article discusses 3D object detection techniques specifically with LiDAR sensors. However, the scope is limited to human detection.

Symbol: H—Detailed discussion. M—Fundamental information. L—no information or few information

We have studied and analyzed the trend of 3D object detection techniques used in diverse fields since 2017. Before 2017, the research related to 3D object detection is observed to be very scarce. Each technique, whether using LiDAR, monocular camera, stereo camera or RADAR, possesses individual strengths and some limitations. From our observation, there is a lack of a thorough study depicting opportunities and obstacles of singular and multimode 3D object detection techniques in different fields. The motivation of this paper is to present a holistic study to assist future practitioners and researchers to choose the suitable method for their application in 3D object detection research.

1.2. Paper Contribution

To the best of our knowledge, this is the first review paper covering diverse sectors of 3D object detection considering different modalities. The key contributions of this review paper are as follows:

- The fundamental concepts of 3D object detection are illustrated in this paper (Section 2);
- Three-dimensional object detection is booming in different sectors. This paper presents 3D object detection techniques applied in different sectors, benchmark datasets in various fields and compares sensor types (Section 3);
- The most frequently used evaluation metrics are discussed (Section 4);
- This paper compares the speed and performance of different methods in a popular benchmark dataset (Section 5);
- A SWOT (Strengths, Weaknesses, Opportunities and Threats) analysis is presented on the singular and multimodal techniques (Section 5);
- After thorough analysis, future directions and limitations of the existing methods are provided (Section 6).

2. Three-dimensional Object Detection Techniques

Three-dimensional object detection requires inferring height and depth information along with object location. To acquire this knowledge about objects, researchers have used several modalities of data from different kinds of device or sensors. Different benchmark datasets of 3D object detection techniques incorporate different modalities, and they exhibit different scopes and constraints. Table 2 briefly discusses the modalities, annotated object category and constraints of the benchmark datasets.

Table 2. Benchmark datasets of 3D object detection.

Dataset	Modalities					Annotated Object Category	Constraint
	LiDAR	RGB Camera	Stereo Camera	RADAR	RGB-D		
KITTI [9]	✓	✓	✓ (Greyscale)	✗		11 classes	Focuses on road scene. Not suitable for indoor robotic vision research.
nuScenes [12]	✓	✓	✗	✓		23 classes	Focuses on autonomous driving scenario.
Waymo [13]	✓	✓	✗	✓		23 classes	Exclusively built for autonomous driving.
SUN RGB-D [10]	✗	✓	✗	✗	✓	700 classes	Focuses only on indoor scene.
ScanNet [11]	✗	✓	✗	✗	✓	20 classes	Focused on indoor scenes.

In singular modality only one type of device is used, whereas multimodal methods combine more than one device. These modalities are reviewed in this section. Sections 2.1 and 2.2, respectively, describe singular modalities and multimodal methods practiced by researchers.

2.1. Singular Modalities

2.1.1. Point Clouds

Numerous 3D object detections are performed from a point cloud. A point cloud can be generated from Kinect devices (SUN RGB-D dataset), CAD models, mobile scanners or LiDAR [28]. Point cloud is the illumination value along with the 3D location of points reflected from the object surface. LiDAR sensors having LASERs can directly produce a point cloud. From the depth information (usually obtained by infrared or Time of Flight camera) of RGB-D datasets, the point cloud is inferred by processing. For 3D detection, point clouds are processed in two ways, i.e., i. Direct point cloud processing; ii. Projection of the point cloud in Bird's Eye View (BEV) or 2D plane.

Direct Point Cloud Processing

In direct point cloud processing, the point information (such as intensity, 3D coordinate values, color, etc.) of point clouds are directly processed by neural networks to extract features and predict 3D object boundaries. The network has to handle higher dimension data compared to the 2D plane projection-based method. A visual representation is shown in Figure 2. Object segmentation can be performed directly on a point cloud using a neural network [29,30]. In these research works, authors performed semantic segmentation using the features of points by applying max pooling for the overall skeleton of the object. They further improved this process by applying this technique in different layers and detected fine patterns as well. But this process is not demonstrated for instance level segmentation. Instance segmentation can be required when each object needs to be identified individually. Semantic segmentation can locate the same class of all objects with common category names, such as cars, people, bicycles, etc., whereas instance segmentation is capable of distinguishing one car from the others. This is often essential in object tracking or surveillance.

Shi et al. illustrated another point cloud-based network called PointRCNN where points are segmented as foreground and background in the first stage to generate 3D region proposals. Then, coordinates of the proposed regions are identified for performing object detection. However, this network showed poor performance on pedestrian detection compared to vehicles detection. The sparse nature of the point cloud from a small-sized object (here a pedestrian) may be a reason for the lower precision rate [31].

Qi et al. utilized an interesting geometric concept, Hough voting, to identify the core of the 3D object. From the point cloud of the surface of the object, the center can be located using Hough voting. They have successfully demonstrated 3D object detection by locating household items such as a sofa, chair, table, etc. in the SUN RGB-D and ScanNet datasets. But the applicability of this method in long-range detection such as on UAV images is uncertain based on the information of this research [32].

Some researchers have rasterized the information of point clouds, which are called voxels, and then the feature of every voxel is encoded [33–36]. An improved method of this network is introduced by Zhou, Y. and Tuzel, O. (2018). Instead of handcrafted feature encoding, they have incorporated a region proposal network for detection. They have detected different vehicles and pedestrians in the KITTI dataset, proving this method to be suitable in autonomous vehicles. Interestingly pedestrians detection required close distance LiDAR, compared to vehicles detection, It suggests that small objects such as pedestrians and cyclists generate too sparse point cloud to detect from long distance [37].

Projection of Point Cloud in Bird's Eye View (BEV) or 2D Plane

Another commonly practiced method for 3D object detection is projecting the 3D point cloud to a pseudo-2D or Birds Eye View plane. After 2D view projection, a neural network is applied to extract features and generate object prediction. Some researchers have applied CNNs after converting the point clouds into 2D planes in their research [38,39]. A similar conversion technique from 3D point cloud to 2D plane is found in [40,41]. After

the perspective view conversion, CNN-based feature recognition network is applied in this research.

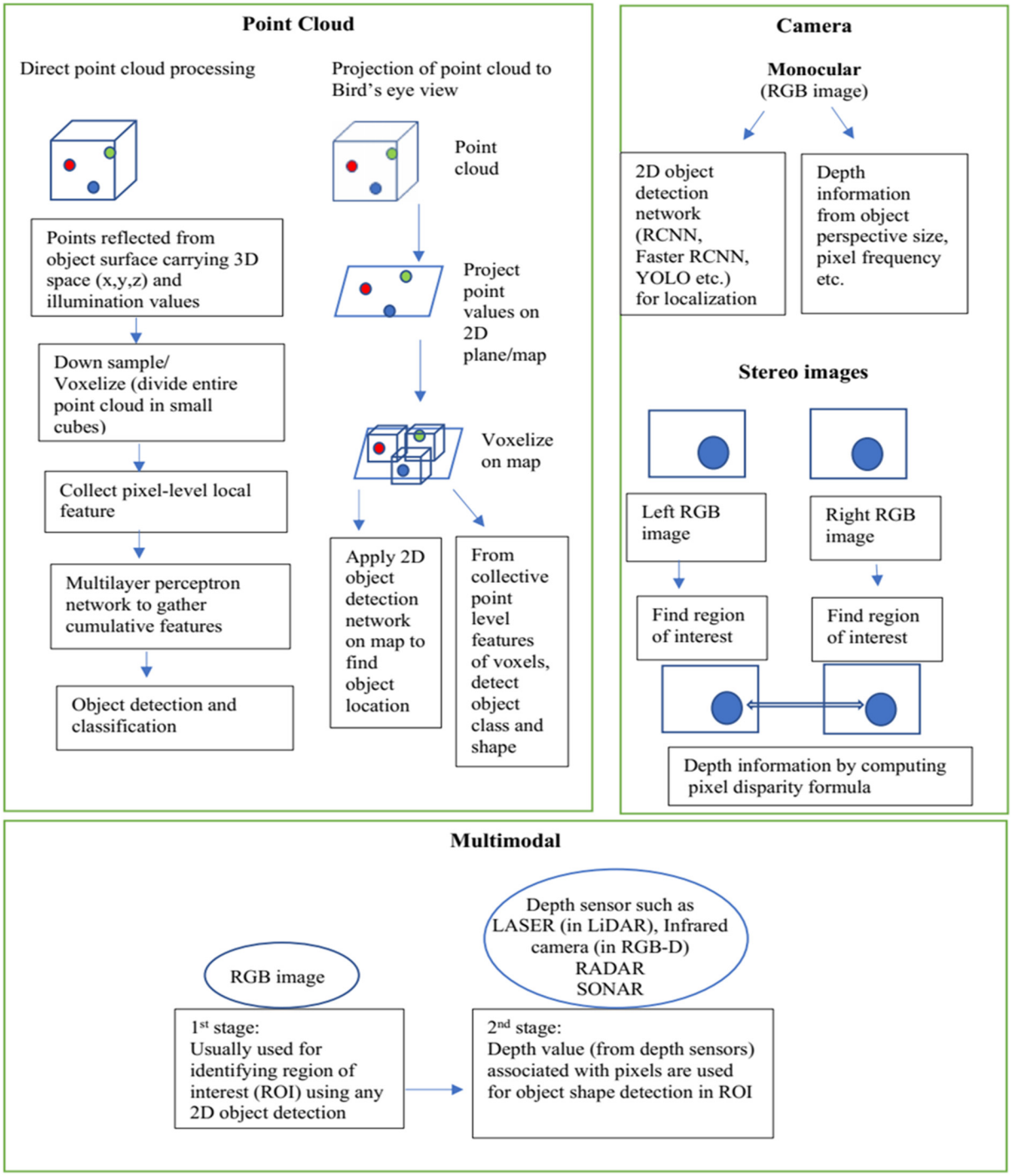


Figure 2. Simplified illustration of 3D object detection technique stages.

Fang, Jin et al. used a LiDAR point cloud in the form of VoxelNet or PointPillars and then applied 3D CNN in feature map. A heat map is generated where the local maxima is the object center. Identifying the center made the object tracking faster. The nuScenes and Waymo Open Datasets are used in their research to demonstrate the precision [42]. Simon, Martin et al. (2019) created a Birds Eye View projection from a point cloud and then applied Complex YOLO following the YOLO V2 technique to predict the object class. After that, Euler Region Proposal was used to detect the orientation of the object [43].

In research of Lang, Alex H. et al. (2019), the features of point clouds are stacked in the form of pillars and named PointPillars. A pseudo-2D image is then projected with the PointPillar features. They claimed that any 2D CNN-based network can be applied at this stage to detect the object. In this paper, they applied Single Shot Detector (SSD) [44]. Though research on projecting 3D points to 2D planes has successfully detected road scene objects such as vehicles or pedestrians, the data transformation to another plane may lead to information loss and cause additional computation loss.

2.1.2. Camera/Vision

Vision-based systems have been successfully applied for both object detection and tracking [45–48]. Mahayuddin et al. performed vision-based moving object detection using semantic convolutional features and achieved a greater detection rate than YOLO V3 and faster than RCNN [46]. Besides single objects, researchers even estimated dynamic crowds from UAV imagery using a vision-based system [49]. While this research focuses on 2D object detection, vision-based moving object detection is also possible in three dimensions [50,51]. Three-dimensional detection using cameras can be performed using two methods, i.e., i. monocular and ii. stereo images.

Monocular (RGB)

Monocular camera is a single camera that is used to generate 2D images. From single 2D images, 2D object detection task i.e., object classification and localization can be done. For 3D object detection, additional depth information is inferred from geometric cues, contrast or prior shape information of the known object.

Several researchers performed 3D object detection only from a single image [52–54]. Monocular image-based 3D reconstruction is proposed by Shapii et al., 2020, where multiple images are used for generating a 3D view of a human activity pose [55]. This is the cheapest method for 3D object detection. But the accuracy is lower than stereo and LiDAR-based detection. However, monocular images can be combined with other 3D detection techniques to achieve better precision. This is discussed in Section 2.2, entitled "Multimodal methods".

Stereo

Stereo cameras utilize two cameras, one capturing the left and another the right image. Its working principle follows that of human vision. By comparing the disparity between the same pixels of two camera images, it is possible to perceive the depth information of the object. Stereo images are proved to be usable for 3D object detection, with a slight compromise in precision compared to LiDAR [56–58]. Their research proves that stereo images can be used for 3D object detection precisely and economically.

Zhang et al. applied the CNN-based network TrackleNet for detecting and tracking pedestrians [50]. Stereo R-CNN [59], Pseudo-LiDAR from Visual Depth [56] and Triangulation Learning Network [60] are 3D deep learning networks developed and successfully applied for autonomous vehicle navigation. You et al. utilized stereo images guided by an economical few point laser for improving precision [57].

2.2. Multimodal Methods

Three-dimensional object detection can be performed by applying multiple sensors or devices. Based on the application and requirements, researchers have combined different sensors having different strengths. Recent research works on multimodal methods are briefly discussed in this section.

2.2.1. Two-dimensional Image and Point Cloud

Qi et al., 2018, used both RGB images and depth information for 3D object detection. In this research, CNN is used on an RGB image to generate region proposals, and then in the proposed regions depth information is merged to create a 3D frustum. After that, 3D

instance segmentation and "amodal 3D box estimation" is performed to view the object behind an obstacle. This method works at real-time speed and exhibits high recall, even for detecting small objects. But this method relies heavily on the region proposed by the 2D object detector at the beginning. Therefore, an appropriate algorithm for region proposal in the 2D plane is essential for the overall success of the entire process [61]. Similar methods are observed in [62–65].

LiDAR-only methods (i.e., point cloud-based) are improved by incorporating RGB images in Point-RCNN [31], VoxelNet [37,66] and PointPillars [44]. These methods are similar to PointPainting [67–70].

Another blended method between a point cloud and an RGB image fuses features of point louds and features of images [71–90]. From collective features, the region of interest is identified. In these regions, 3D object detection is performed afterwards. Although combining RGB and LiDAR point cloud information is complex, it exhibited better performance in outdoor scene 3D object detection than the standalone LiDAR voxel method [91].

2.2.2. RADAR and 2D Image

Radio Detecting and Ranging (RADAR) senses distance using radio signals. As RADAR cannot anticipate color information, it can only assume the shape or size rather than classify. A big advantage of RADAR is being less prone to harsh weather compared to LiDAR or camera images. But combining RADAR with images has successfully improved precision of 3D object detection [92–94]. In [92], the object center is detected from a 2D image. Then, targeting that object center, a RADAR point cloud is used for obtaining depth information in frustrum shape. It improved image-based detection in the nuScenes dataset. Nabati et al. (2020) used RADAR-generated point clouds for region proposal for objects. Then this region is imposed on 2D image to perform 3D detection [93].

2.2.3. Other 3D Object Detection Methods

Wang et al., 2020, proposed a multimodal system for leveraging the advantages of images, LiDAR and RADAR. In this method, the first image is used to obtain the regions of interest. Then, in these regions, point clouds from LiDAR are used to obtain depth information along with orientation of the object. RADAR is used for further refinement of 3D detection [95].

Apart from the above-mentioned methods, simultaneous localization and mapping (SLAM) is applied for constructing 3D views. Chen et al. (2022) applied SLAM for waste sorting [96]. They used both image and LiDAR data for the SLAM application. They created their own dataset. Zhang et al. (2019) used SLAM on Multiview stereo images for 3D detection of humans from a UAV [50]. Structure-from-motion (SFM) is another 3D detection and localization approach recently being applied from UAVs. Gené-Mola et al. (2020) applied SFM for 3D detection of fruits [97]. This approach involves capturing multiple 2D images and then constructing a 3D view by combining these images by matching the key features. As this method is time-consuming, instance segmentation is performed to reduce the regions of interest before generating the 3D view. Teng et al. (2023) compared both LiDAR-based SLAM and SFM. In their research, they found SLAM to be faster and more accurate than SFM [98]. For better visualization, an illustration of 3D object detection technique stages is presented in Figure 2 in a very simplified manner.

3. Branches of 3D Object Detection Techniques

Three-dimensional object detection is gaining popularity in various domains. Recent research works on 3D object detection in different application sectors are shown in Figure 3.

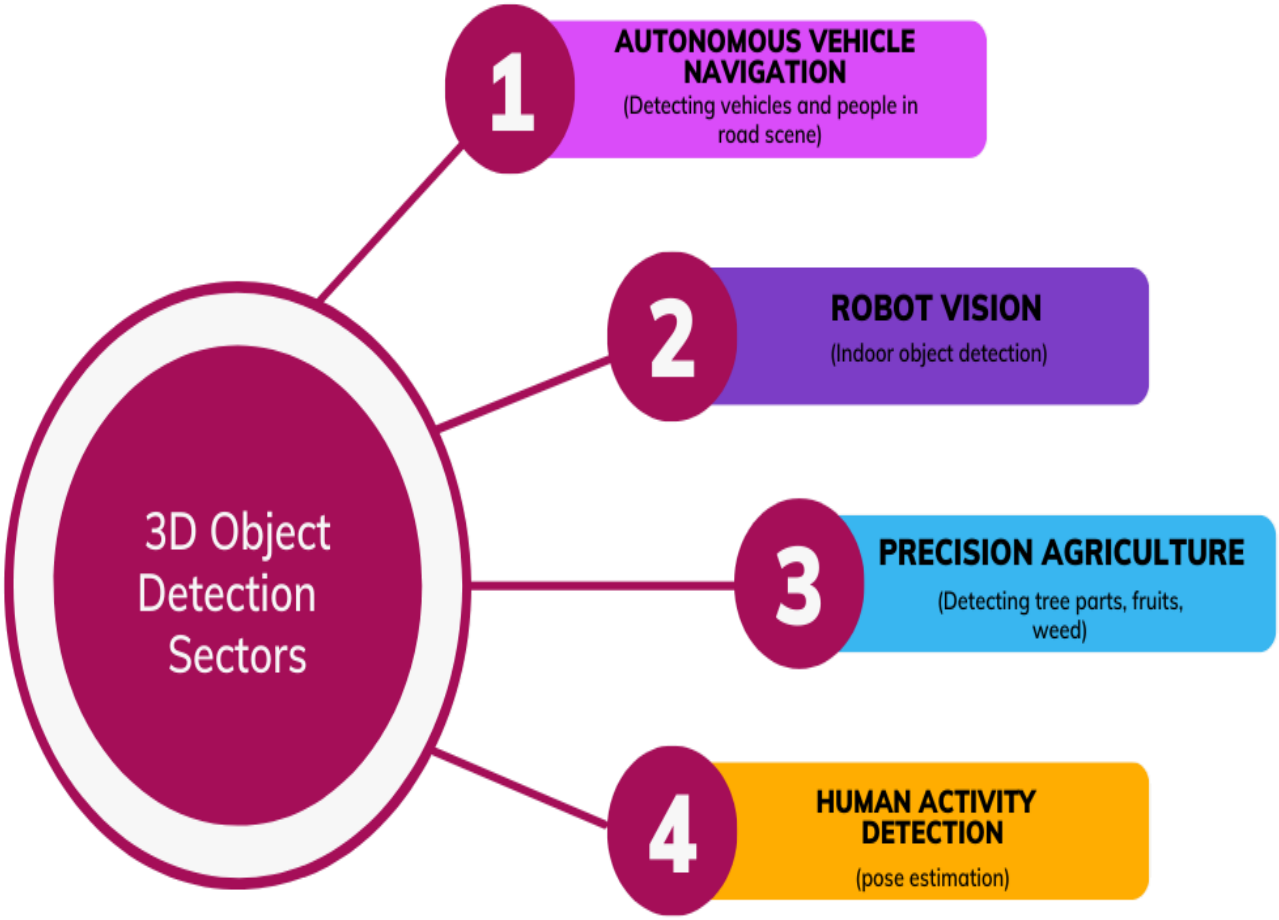


Figure 3. Applications of 3D object detection in different sectors.

The application of 3D object detection is expanding day by day. Figure 3 depicts the commonly practiced sectors of 3D object detection. Among all the four sectors, most research work is conducted in the autonomous vehicle navigation domain. The benchmark datasets are most versatile in this sector as well. Research works in these domains are tabularized in Tables 3 and 4.

Apart from the above-discussed applications, 3D object detection is also performed in some recent medical research, such as augmented assistance in surgery or diagnosis. The input data format and acquisition process of 3D object detection in the medical field is different than other sectors. Computed Tomography (CT), Diffusion tensor imaging (DTI), magnetic resonance imaging (MRI), Functional magnetic resonance imaging and Ultrasound are three-dimensional. Three-dimensional convolutional networks can be applied to these images for 3D detection of organs for better disease detection. Three-dimensional CNN is successfully applied for disease severity prediction and classification [99–101].

However, 3D datasets are very few in number in the medical field. Patient confidentiality may be a reason behind it. Image augmentation is sometimes applied to increase the dataset size [102,103]. In Table 3, we have summarized singular modality 3D object detection methods practiced since 2017.

Multimodal techniques involve multiple sensors for 3D object detection. In Table 4, recent multimodal techniques are mentioned. This table indicates that multimodal techniques usually involve LiDAR or RGB images accompanied by some other modality.

Tables 3 and 4 exhibit singular and multimodal techniques of 3D object detection in different fields. Observing the trends, the following insights are found.

- Autonomous Vehicle navigation: LiDAR is a very popular modality in this field. It can be used both in singular and multimodal methods. LiDAR exhibits long-range

LASER scanning, making it capable of designing a standalone end-to-end 3D object detection system. RGB-D sensors have low range (usually lower than 10 m). Due to this constraint, autonomous vehicle navigation-related research works are not found to use this modality.

Among the vision-based technologies, a monocular RGB camera was used as a singular modality by few researchers, but it could not exhibit as much accuracy as the LiDAR sensor. Moreover, detectability range was lower than LiDAR. Some researchers have mentioned stereo cameras as a modality with high potential in autonomous vehicle navigation. Even the automotive company Tesla is focusing on stereoscopic vision rather than LiDAR, considering it more nature-inspired, economical, and similar to human vision. However, stereoscopic camera ranges are far less than that of LiDAR.

RADAR is not found to be used as a singular modality system for 3D object detection since, it cannot perceive color information of the objects. But in some research works, RADAR was implemented along with other modalities.

- Robot Vision: RGB-D sensor is the most popular sensor, serving both in singular and multimodal techniques of robot vision. Most of these research works are conducted in indoor environments. For this reason, the long-range detection requirement is absent in this field. This makes RGB-D a wonderful option for perceiving both color and depth information of objects in indoor environments. RGB-D cameras are constructed with an RGB camera (for color perception) and infrared sensors (for depth perception).
- Precision Agriculture: In agriculture, LiDAR is found to be used for long-range 3D detection. Specially, precision agriculture involving UAVs from high altitude is benefitted by LiDAR. LiDAR has been applied as singular modality or multimodal technique with other sensors such as RGB cameras or narrow beam SONAR (Sound Navigation and Ranging). However, monocular cameras are being used as a singular modality in the Multiview 3D detection technique. In this method, 2D images captured from different angles around the object contribute to 3D detection. In the case of multimodal detection, RGB cameras can be used with RGB-D sensors. Monocular cameras, having lower range than LiDAR can serve for 3D object detection from a closer range.
- Human activity/pose detection: Monocular RGB cameras are widely used for human pose detection using the Multiview 3D object detection technique. For long-range detection, LiDAR has been used by researchers. To enhance the detectability of LiDAR-based detection, some other modalities such as inertial measurement unit (IMU) are also used in existing research works. However, in indoor robot vision, human activity detection and precision agriculture, RADAR is not usually preferred for 3D object detection technique. The reason may be the low spatial resolution of RADAR (compared to LiDAR or camera) which makes detection of thin objects or close-proximity objects difficult and ambiguous.

Multimodal techniques have evolved for utilizing the benefits of multiple sensors. For example, structural information of objects is well perceived using LiDAR for long range, whereas cameras are good at perceiving fine texture information. However, multimodal techniques come with the additional cost and complexity of synchronizing different formats of data from different sensors.

Table 3. Singular modality techniques of 3D object detection methods from 2017 to 2022.

Domain	Ref.	Sensor Data Type						Dataset					
		LiDAR	RGB-D	Monocular (RGB) image	Stereo Image Pair	RADAR	other	KITTI [9]	nuScenes [12]	Waymo [13]	SUN RGB-D [10]	ScanNet [11]	Others
Autonomous Vehicle navigation (Detecting vehicles and people in road scene)	[37]	✓						✓					
	[31]	✓						✓					
	[43]	✓						✓					
	[44]	✓						✓					
	[56]				✓			✓					
	[59]				✓			✓					
	[60]				✓			✓					
	[53]			✓				✓					
	[42]	✓							✓	✓			
	[54]			✓					✓				
	[104]							✓					
	[105]	✓						✓	✓				
Indoor objects (Robotic Vision)	[104]		✓								✓	✓	
	[105]		✓								✓	✓	
	[29]		✓										[106,107]
	[30]		✓									✓	[107,108]
	[32]		✓								✓	✓	
	[109]		✓										[110]
	[111]		✓										Self-made
	[112]		✓										TUM RGB-D
Precision Agriculture	[113]						Multi spectral						Self-made
	[97]	✓											Self-made
	[114]			✓ (Multi View)									Self-made
Human Pose/ Activity Detection	[115]			✓ (Multi View)									[116–118]
	[119]			✓ (Multi View)									[116–118]

Table 4. Multimodal modality techniques of 3D object detection methods from 2017 to 2022.

Domain	Ref.	Sensor Data Type						Dataset					
		LiDAR	RGB-D	Monocular (RGB) image	Stereo Image	RADAR	Other	KITTI [9]	nuScenes [12]	Waymo [13]	SUN RGB-D [10]	ScanNet [11]	Others
Autonomous Vehicle navigation	[61]	✓		✓				✓					
	[120]	✓		✓				✓					
	[90]	✓		✓				✓					[121]
	[67]	✓		✓				✓	✓				
	[92]			✓		✓			✓				
	[94]			✓		✓			✓				[122]
	[95]	✓		✓		✓			✓				
Indoor objects (Robotic Vision)	[61]		✓	✓							✓		
Precision Agriculture	[123]		✓	✓									Self-made
	[124]		✓	✓									Self-made
	[125]	✓		✓			narrow beam SONAR						Self-made
Human Pose/ Activity Detection	[126]	✓					inertial measurement unit (IMU)						[127]

4. Evaluation Metrics

A commonly used metric for measuring the performance of object detection is Average Precision [103]. Precision, average precision and mean average precision are widely utilized metrics for evaluating performance of 2D object detection. For 3D object detection, different benchmark competitions use average precision (AP) as well. AP is used for performance comparison in [37,61,128,129]. Average Orientation Estimation is also used as 3D object detection evaluation metric which shows the angle of the object [44,129]. A slightly different version of AP is Mean Average Precision, used by [32,42,44,129].

Wang et al. used Average Precision at different IoU (Intersection over Union) values, i.e., 0.5 or 0.7, for evaluation [56]. IoU is used to measure the amount of matched area between ground truth and the detected object. If there is a perfect match, i.e., the object is perfectly predicted, then the IoU value will be 1.

If TP indicates True Positive (correctly detected object), FP means False Positive (mistakenly detected as object), TN means True Negative (correct mentioning of absence of object) and FN means False Negative (could not predict the presence of object), then, from [130,131], the simplified equations of precision and recall are as follows:

$$\text{Precision} = \text{TP}/(\text{TP} + \text{FP}) \tag{1}$$

$$\text{Recall} = \text{TP}/(\text{TP} + \text{FN}) \tag{2}$$

The parameters of precision and recall are explained in Figure 4. The figure illustrates the meaning of TP, FP, TN and FN.

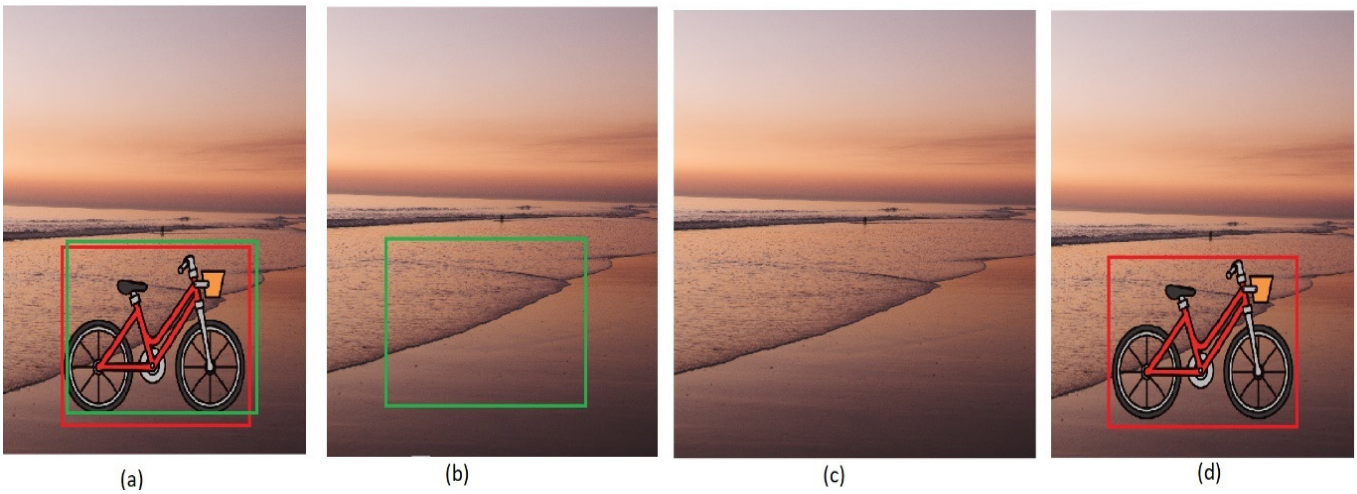


Figure 4. (**a**) True Positive; (**b**) False Positive; (**c**) True Negative; (**d**) False Negative (the red rectangle indicates ground truth, and the green rectangle stands for detection).

In Figure 4a, when the object (in this case the bicycle) is present in the image and it is detected, this scenario is called true positive. In Figure 4b, there was no bicycle in the image, but it was falsely detected. This scenario is called false positive. In Figure 4c, there was no bicycle and there was no detection of bicycles either. It is classed as a true negative. In Figure 4d, there was a bicycle, but it was not detected by the model. This is called false negative.

If the precision and recall curve is *p(r)*, then the formula to compute average precision is:

$$\mathrm{AP} = \int_0^1 p(r) \tag{3}$$

Mean Average Precision (mAP) is computed by computing the average value of all the AP corresponding to an entire dataset. If a dataset has 10 categories, then after finding the AP for each category the mean is computed for these 20 categories to obtain mAP.

5. Observation and Analysis

Since 2017, most of the 3D object detection methods have been implemented for autonomous vehicle navigation. Numerous competitions have been organized on different open benchmarks, especially on road scenes. Three-dimensional object detection is gradually becoming popular in other domains as well, with the availability of GPU-based computers allowing high computation capabilities.

A comparison of the common devices for 3D detection is presented in Table 5. Information in this table is collected from the research work of Chen et al. (2021) [132].

Table 5. Performance comparison of 3D object detection sensors and devices.

Performance Criterion	RADAR	LiDAR	RGB Camera
Object detection	Good	Good	Fair
Object classification	Poor	Fair	Good
Distance inference	Good	Good	Fair
Detecting edge	Poor	Good	Good
Visibility range	Good	Fair	Fair
Adverse weather performance	Good	Fair	Poor
Performance in low light condition	Good	Good	Fair

The performance of 3D object detection methods depending on the type of sensor can be seen from Table 6. Table 6 contains the top average precision methods exhibited by different modalities in the KITTI vision benchmark leaderboard.

Table 6. Comparison of networks in KITTI Vision benchmark Suite.

Modality	Method	Reference	Category	Easy	Moderate	Hard	Run Time
LiDAR+ RGB	VirConv-S	[133]	Car	92.48%	87.20%	82.45%	0.09 s
LiDAR+ RGB	LoGoNet	[134]	Car	91.80%	85.06%	80.74%	0.1 s
LiDAR+ RGB	LoGoNet	[134]	Pedestrian	54.04%	47.43%	44.56%	0.1 s
LiDAR	CasA++	[135]	Car	90.68%	84.04%	79.69%	0.1 s
LiDAR	CasA++	[135]	Pedestrian	56.33%	49.29%	46.70%	0.1 s
Stereo	DSGN++	[58]	Car	83.21%	67.37%	59.91%	0.2 s
Stereo	DSGN++	[58]	Pedestrian	43.05%	32.74%	29.54%	0.2 s
Stereo	DMF	[136]	Car	77.55%	67.33%	62.44%	0.2 s
Stereo	DMF	[136]	Pedestrian	37.21%	29.77%	27.62%	0.2 s
Monocular (RGB)	CIE + DM3D	[137]	Car	35.96%	25.02%	21.47%	0.1 s
Monocular (RGB)	QD-3DT [LSTM on RGB]	[138]	Car	12.81%	9.33%	7.86%	0.03 s
Monocular (RGB)	QD-3DT [LSTM on RGB]	[138]	Pedestrian	5.53%	3.37%	3.02%	0.03 s

Table 6 shows that LiDAR-based methods exhibited maximum average precision. Multimodal techniques are a recent addition; in particular, LiDAR with camera techniques is showing good potential in terms of precision value. Thin surface objects (such as pedestrians) detection achieved average precision than wide surface cars. The monocular techniques captured only RGB image frames and then the images were trained in sequential network (e.g., LSTM). For dealing with only RGB, the speed was found to be fast, but the average precision was very much less than stereo or LiDAR.

Analyzing the practiced modalities of 3D object detections, following SWOT (Strengths, Weaknesses, Opportunities and Threats) analysis, Tables 7–9 are developed.

Table 7. SWOT analysis on LiDAR (point cloud)-based 3D object detection.

Strengths • LiDAR produces a direct point cloud, conveying the three-dimensional information (position and depth) and corresponding illumination value from the object surface. • Up to now, LiDAR-based 3D object detection has been proved to have the highest precision.	**Weaknesses** • LiDAR-generated point clouds are sparse. For thin surface objects, the sparsity is greater, which makes the detection process less precise. • Thin objects can be detected only at a closer distance from the LiDAR position. Hence, average precision in detecting pedestrians or cyclists is always lower than vehicles.
Opportunities • It can be applied for real-time applications. • Waymo and some other vehicle companies are using LiDAR-based 3D object detection in self-driving cars.	**Threats** • LiDARs are not commonly available. Researchers rely on open datasets if they cannot use LiDARs of their own. • High price of LiDARs is not economic for numerous appli-cations. • Majority of open datasets of LiDAR point clouds are fo-cused on road scenes, making it difficult for other sectors to grow.

Table 8. SWOT analysis on camera/vision-based 3D object detection.

Strengths • It is the most inexpensive method, making it readily available to researchers even for developing new da-tasets for deep learning. • Camera-based methods can classify object types with less processing compared to LiDAR or RGB-D. • Unlike LiDAR, thin object detection does not suffer from lower precision.	**Weaknesses** • Fog or dusty weather can easily hinder object detection. • Depth information in single 2D images is less precise than any other method.

Table 8. *Cont.*

Opportunities	Threats
• While integrating with other modalities, the vi-sion-based method is proved to be very precise.	• Using multi-view 2D images for 3D detection (e.g., in the Structure-from-motion technique) is time-consuming, making it difficult for real-time applications.

Table 9. SWOT analysis on Multimodal 3D object detection.

Strengths	Weaknesses
• In combined methods, one device or sensor can over-come the individual weakness of another. Hence, it provides better precision than single modality.	• Integrating different modality data is challenging. • It often requires more time to integrate the stages.
Opportunities	**Threats**
• With greater availability of sensors and datasets, combined methods are gaining popularity among re-searchers and practitioners.	• Failure of one modality can affect the overall success of the entire process. • Unavailability of datasets in different sectors can ham-per combined modality research.

6. Advancement, Challenges and Future Directions

Analyzing the strengths and weaknesses of existing research works, some valuable insights are learnt. These are listed as follows:

- Point cloud-based 3D object detection can be performed in both indoor and outdoor environments. However, LiDAR can generate point clouds at longer range in variable weather conditions, whereas RGB-D or Kinect-based point clouds face limitations in terms of range and weather conditions. For this reason, point cloud-based 3D object detection is performed with the help of LiDAR sensors in autonomous vehicle navigation research. However, RGB-D sensors are less expensive, and the generated point clouds are successfully implemented in close-range research works on precision agriculture or indoor robotic vision;
- Three-dimensional object detection technology is significantly supported by deep learning. Deep learning networks are comprised of multilayer neural networks which can learn the patterns of data. In significant 3D object detection networks, such as PointNet, PointNet++, VoxelNet, CenterNet, etc., deep learning is used to learn object information from points or group of points. Also, in the two-stage networks where initially RGB images are used for region proposals, deep learning is applied for predicting the regions of objects. Future research work may include deep learning for leveraging more opportunities such as transfer learning in 3D object detection-related research;
- Development of end-to-end 3D object detection networks is becoming popular for their ease of application. End-to-end networks directly need to collect raw sensor data and provide 3D bounding-box output predictions. To develop such networks, it is essential to choose the necessary type of sensor (LiDAR, camera or RADAR), pre-process the data, design a neural network to learn the features of the data and train, validate and evaluate the model. Both hardware and software knowledge are required for the developers of end-to-end networks;
- Data collection and annotation for 3D object detection is more complex compared to 2D object detection. Three-dimensional object detection data collection involves fusing the data from different types of sensors, such as LiDAR, monocular or stereo cameras, RADAR, etc. This process requires calibration of different devices and synchronization of the data. The data annotation for 3D object detection needs to describe not only the object location, but also its dimensions in space, position and orientation. The data description involves parameters such as object length, width, height, yaw, pitch,

roll, occlusion amount, etc. More expertise in terms of 3D geometry is necessary to annotate data in the case of 3D object detection;

- The limitation of point cloud-based object detection, especially in outdoor environments, is sparsity. Hence, thin object detection is an open research question in this field. Comparing the average precision values of cars and pedestrians in Table 6, it is clearly visible that detection of the thin surface of pedestrians achieved far less precision. Hence, how to increase precision for thin object detection with point cloud methods is open research question;
- Data scarcity is one of the main constraints in 3D object detection-related research. Due to the support and sponsorship of automobile companies in their pursuit of building self-driving cars, some enriched benchmark datasets such as KITTI, Waymo, nuScenes, etc. are widely available. A few indoor benchmark datasets are also available for robot vision research, such as SUN RGB-D and ScanNet. But there is scarcity of open benchmark datasets for other fields. Specially, 3D object detection is becoming popular in precision agriculture, but the conducted research works are found to be using self-collected datasets. However, these datasets are not publicly available. This is a constraint on conducting 3D object detection research in agriculture.

To reduce the issue of data scarcity, the 2D object detection sector is getting benefit from the new technology Generative Adversarial Networks (GAN). GAN is a type of augmentation technique that can generate new synthetic data. In the near future, 3D object detection techniques can be benefitted by GAN as well. To enhance 3D datasets, more 3D CAD model datasets (e.g., ShapeNet) can emerge in the near future.

7. Conclusions

Three-dimensional object detection is rapidly gaining popularity among researchers and practitioners. As the real world is 3D, hence depth and height information are vital too. Previously the low computational power of computers, unavailability of sensors and lower number of datasets hindered research on 3D object detection. With time, these obstacles are being reduced, and 3D object detection is being performed in numerous sectors. This review paper is aimed to assist researchers and practitioners to be informed about the recent trends in the 3D object detection sector. This study has carefully analyzed the advantages and disadvantages of different modalities so that future researchers can benefit while selecting suitable methods and sensors for their application.

Author Contributions: Conceptualization, Z.R.M. and T.K.; methodology, T.K.; Analysis, Z.R.M. and M.K.H.; investigation, Z.R.M.; writing—original draft preparation, T.K.; writing—review and editing, T.K.; visualization, M.K.H.; supervision, Z.R.M.; project administration, Z.R.M.; funding acquisition, Z.R.M. All authors have read and agreed to the published version of the manuscript.

Funding: The authors would like to thank Universiti Kebangsaan Malaysia for providing financial support under the "Geran Universiti Penyelidikan" research grant, GUP-2020-064.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Redmon, J.; Divvala, S.; Girshick, R.; Farhadi, A. You Only Look Once: Unified, Real-Time Object Detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016. Available online: http://pjreddie.com/yolo/ (accessed on 19 January 2023).
2. Redmon, J.; Farhadi, A. YOLO9000: Better, Faster, Stronger. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017. Available online: http://pjreddie.com/yolo9000/ (accessed on 19 January 2023).
3. Redmon, J.; Farhadi, A. YOLOv3: An Incremental Improvement. *in Computer vision and pattern recognition. arXiv* **2018**, arXiv:1804.02767. [CrossRef]
4. Bochkovskiy, A.; Wang, C.-Y.; Liao, H. YOLOv4: Optimal Speed and Accuracy of Object Detection. *arXiv* **2020**, arXiv:2004.10934. [CrossRef]

5. Thuan, D. Evolution of Yolo Algorithm and Yolov5: The State-of-the-Art Object Detention Algorithm. 2021. Available online: http://www.theseus.fi/handle/10024/452552 (accessed on 19 January 2023).
6. Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S.; Fu, C.-Y.; Berg, A.C. SSD: Single shot multibox detector. In Proceedings of the Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, 11–14 October 2016; Proceedings, Part I; Springer: Berlin/Heidelberg, Germany, 2016; pp. 21–37.
7. Girshick, R. Fast R-CNN. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 7–13 December 2015; pp. 1440–1448. Available online: https://github.com/rbgirshick/ (accessed on 19 January 2023).
8. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. In Proceedings of the Advances in Neural Information Processing Systems 28 (NIPS 2015), Montreal, QC, Canada, 7–12 December 2015; Available online: https://github.com/ (accessed on 19 January 2023).
9. Geiger, A.; Lenz, P.; Urtasun, R. Are we ready for autonomous driving? The KITTI vision benchmark suite. In Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Providence, RI, USA, 16–21 June 2012; pp. 3354–3361. [CrossRef]
10. Song, S.; Lichtenberg, S.P.; Xiao, J. SUN RGB-D: A RGB-D Scene Understanding Benchmark Suite. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015. Available online: https://rgbd.cs.princeton.edu/ (accessed on 11 January 2023).
11. Dai, A.; Chang, A.X.; Savva, M.; Halber, M.; Funkhouser, T.; Nießner, M. ScanNet | Richly-annotated 3D Reconstructions of Indoor Scenes. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017. Available online: http://www.scan-net.org/ (accessed on 11 January 2023).
12. Caesar, H.; Bankiti, V.; Lang, A.H.; Vora, S.; Liong, V.E.; Xu, Q.; Krishnan, A.; Pan, Y.; Baldan, G.; Beijbo, O. nuScenes: A Multimodal Dataset for Autonomous Driving. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 13–19 June 2020; pp. 11621–11631.
13. Sun, P.; Kretzschmar, H.; Dotiwalla, X.; Chouard, A.; Patnaik, V.; Tsui, P.; Guo, J.; Zhou, Y.; Chai, Y.; Caine, B. Scalability in perception for autonomous driving: Waymo open dataset. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020. Available online: https://waymo.com/open/ (accessed on 12 January 2023).
14. Shahbazi, M.; Ménard, P.; Sohn, G.; Théau, J. Unmanned aerial image dataset: Ready for 3D reconstruction. *Data Brief* **2019**, *25*, 103962. [CrossRef] [PubMed]
15. SVélez, S.; Vacas, R.; Martín, H.; Ruano-Rosa, D.; Álvarez, S. High-Resolution UAV RGB Imagery Dataset for Precision Agriculture and 3D Photogrammetric Reconstruction Captured over a Pistachio Orchard (*Pistacia vera* L.) in Spain. *Data* **2022**, *7*, 157. [CrossRef]
16. Li, T.; Liu, J.; Zhang, W.; Ni, Y.; Wang, W.; Li, Z. UAV-Human: A Large Benchmark for Human Behavior Understanding With Unmanned Aerial Vehicles. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 16266–16275.
17. Singh, S.P.S.; Wang, L.; Gupta, S.; Goli, H.; Padmanabhan, P.; Gulyás, B. 3d deep learning on medical images: A review. *Sensors* **2023**, *20*, 5097. [CrossRef] [PubMed]
18. Fernandes, D.; Silva, A.; Névoa, R.; Simões, C.; Gonzalez, D.; Guevara, M.; Novais, P.; Monteiro, J.; Melo-Pinto, P. Point-cloud based 3D object detection and classification methods for self-driving applications: A survey and taxonomy. *Inf. Fusion* **2020**, *68*, 161–191. [CrossRef]
19. Zamanakos, G.; Tsochatzidis, L.; Amanatiadis, A.; Pratikakis, I. A comprehensive survey of LIDAR-based 3D object detection methods with deep learning for autonomous driving. *Comput. Graph.* **2021**, *99*, 153–181. [CrossRef]
20. Arnold, E.; Al-Jarrah, O.Y.; Dianati, M.; Fallah, S.; Oxtoby, D.; Mouzakitis, A. A Survey on 3D Object Detection Methods for Autonomous Driving Applications. *IEEE Trans. Intell. Transp. Syst.* **2019**, *20*, 3782–3795. [CrossRef]
21. Liang, W.; Xu, P.; Guo, L.; Bai, H.; Zhou, Y.; Chen, F. A survey of 3D object detection. *Multimedia Tools Appl.* **2021**, *80*, 29617–29641. [CrossRef]
22. Mao, J.; Shi, S.; Wang, X.; Li, H. 3D Object Detection for Autonomous Driving: A Comprehensive Survey. *Int. J. Comput. Vis.* **2023**, *131*, 1–55. [CrossRef]
23. MDrobnitzky, M.; Friederich, J.; Egger, B.; Zschech, P. Survey and Systematization of 3D Object Detection Models and Methods. *Vis. Comput.* **2023**, 1–47. [CrossRef]
24. Wu, Y.; Wang, Y.; Zhang, S.; Ogai, H. Deep 3D Object Detection Networks Using LiDAR Data: A Review. *IEEE Sens. J.* **2021**, *21*, 1152–1171. [CrossRef]
25. Hoque, S.; Arafat, Y.; Xu, S.; Maiti, A.; Wei, Y. A Comprehensive Review on 3D Object Detection and 6D Pose Estimation with Deep Learning. *IEEE Access* **2021**, *9*, 143746–143770. [CrossRef]
26. Mohan, N.; Kumar, M. Room layout estimation in indoor environment: A review. *Multimedia Tools Appl.* **2022**, *81*, 1921–1951. [CrossRef]
27. Hasan, M.; Hanawa, J.; Goto, R.; Suzuki, R.; Fukuda, H.; Kuno, Y.; Kobayashi, Y. LiDAR-based detection, tracking, and property estimation: A contemporary review. *Neurocomputing* **2022**, *506*, 393–405. [CrossRef]
28. Tong, G.; Li, Y.; Chen, D.; Sun, Q.; Cao, W.; Xiang, G. CSPC-Dataset: New LiDAR Point Cloud Dataset and Benchmark for Large-Scale Scene Semantic Segmentation. *IEEE Access* **2020**, *8*, 87695–87718. [CrossRef]

29. Qi, C.R.; Su, H.; Mo, K.; Guibas, L.J. PointNet: Deep Learning on Point Sets for 3D Classification and Segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017.
30. Li, C.R.Q.; Hao, Y.; Leonidas, S.; Guibas, J. PointNet++: Deep Hierarchical Feature Learning on Point Sets in a Metric Space. *arXiv* **2017**, arXiv:1706.02413.
31. Shi, S.; Wang, X.; Li, H. PointRCNN: 3D Object Proposal Generation and Detection from Point Cloud. 2018. Available online: http://arxiv.org/abs/1812.04244 (accessed on 29 November 2023).
32. Qi, C.R.; Litany, O.; He, K.; Guibas, L. Deep Hough Voting for 3D Object Detection in Point Clouds. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 2 November 2019. Available online: http://arxiv.org/abs/1904.09664 (accessed on 29 November 2023).
33. Huang, X.; Wang, P.; Cheng, X.; Zhou, D.; Geng, Q.; Yang, R. The ApolloScape Open Dataset for Autonomous Driving and Its Application. *IEEE Trans. Pattern Anal. Mach. Intell.* **2020**, *42*, 2702–2719. [CrossRef] [PubMed]
34. Casas, S.; Gulino, C.; Liao, R.; Urtasun, R. SpAGNN: Spatially-Aware Graph Neural Networks for Relational Behavior Forecasting from Sensor Data. In Proceedings of the 2020 IEEE International Conference on Robotics and Automation (ICRA), Paris, France, 31 May–31 August 2020; pp. 9491–9497. [CrossRef]
35. Halder, S.; Lalonde, J.-F.; De Charette, R. Physics-Based Rendering for Improving Robustness to Rain. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 10203–10212. Available online: https://team.inria.fr/rits/computer-vision/weather-augment/ (accessed on 21 January 2023).
36. He, K.; Zhang, X.; Ren, S.; Sun, J. Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Santiago, Chile, 7–13 December 2015; pp. 1026–1034.
37. Zhou, Y.; Tuzel, O. VoxelNet: End-to-End Learning for Point Cloud Based 3D Object Detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018. Available online: http://arxiv.org/abs/1711.06396 (accessed on 29 November 2023).
38. Su, H.; Maji, S.; Kalogerakis, E.; Learned-Miller, E. Multi-View Convolutional Neural Networks for 3D Shape Recognition. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 7–13 December 2015; pp. 945–953. Available online: http://vis-www.cs.umass.edu/mvcnn (accessed on 21 January 2023).
39. Qi, C.R.; Su, H.; Nießner, M.; Dai, A.; Yan, M.; Guibas, L.J. Volumetric and Multi-View CNNs for Object Classification on 3D Data. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 26 June–1 July 2016; pp. 5648–5656.
40. Premebida, C.; Carreira, J.; Batista, J.; Nunes, U. Pedestrian detection combining RGB and dense LIDAR data. In Proceedings of the IEEE International Conference on Intelligent Robots and Systems, Chicago, IL, USA, 14–18 September 2014; pp. 4112–4117. [CrossRef]
41. Gonzalez, A.; Villalonga, G.; Xu, J.; Vazquez, D.; Amores, J.; Lopez, A.M. Multiview random forest of local experts combining RGB and LIDAR data for pedestrian detection. In Proceedings of the IEEE Intelligent Vehicles Symposium, Seoul, Republic of Korea, 28 June–1 July 2015; pp. 356–361. [CrossRef]
42. Yin, T.; Zhou, X.; Krähenbühl, P. Center-based 3D Object Detection and Tracking. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020. Available online: http://arxiv.org/abs/2006.11275 (accessed on 29 November 2023).
43. Simon, M.; Milz, S.; Amende, K.; Gross, H.-M. Complex-YOLO: Real-Time 3D Object Detection on Point Clouds. 2018. Available online: http://arxiv.org/abs/1803.06199 (accessed on 29 November 2023).
44. Lang, A.H.; Vora, S.; Caesar, H.; Zhou, L.; Yang, J.; Beijbom, O. PointPillars: Fast Encoders for Object Detection from Point Clouds. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, Long Beach, CA, USA, 15–20 June 2019. Available online: http://arxiv.org/abs/1812.05784 (accessed on 29 November 2023).
45. Mahayuddin, Z.R.; Saif, A.F.M.S. Edge Feature based Moving Object Detection Using Aerial Images: A Comparative Study. In Proceedings of the 6th International Conference on Computing, Engineering, and Design, ICCED 2020, Sukabumi, Indonesia, 15–16 October 2020.
46. Mahayuddin, Z.R.; Saif, A.F.M.S. Moving Object Detection Using Semantic Convolutional Features. *J. Inf. Syst. Technol. Manag.* **2022**, *7*, 24–41. [CrossRef]
47. Saif, A.F.M.S.; Mahayuddin, Z.R.; Arshad, H. Vision-Based Efficient Collision Avoidance Model Using Distance Measurement. In *Soft Computing Approach for Mathematical Modeling of Engineering Problems*; CRC Press: Boca Raton, FL, USA, 2021; pp. 191–202. [CrossRef]
48. Mahayuddin, Z.R.; Saif, A.S. View of A Comparative Study of Three Corner Feature Based Moving Object Detection Using Aerial Images. *Malays. J. Comput. Sci.* 2019, 25–33. Available online: http://adum.um.edu.my/index.php/MJCS/article/view/21461/10985 (accessed on 13 February 2023).
49. Saif, A.F.M.S.; Mahayuddin, Z.R. Crowd Density Estimation from Autonomous Drones Using Deep Learning: Challenges and Applications. *J. Eng. Sci. Res.* **2021**, *5*, 1–6. [CrossRef]
50. Zhang, H.; Wang, G.; Lei, Z.; Hwang, J.-N. Eye in the Sky. In Proceedings of the 27th ACM International Conference on Multimedia, New York, NY, USA, 21–25 October 2019; pp. 899–907. [CrossRef]

51. Saif, S.; Zainal, F.; Mahayuddin, R. Vision based 3D Object Detection using Deep Learning: Methods with Challenges and Applications towards Future Directions. *Int. J. Adv. Comput. Sci. Appl.* **2022**, *13*, 203–214. [CrossRef]
52. Brazil, G.; Liu, X. M3D-RPN: Monocular 3D Region Proposal Network for Object Detection. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019. Available online: http://arxiv.org/abs/1907.06038 (accessed on 29 November 2023).
53. Liu, Z.; Wu, Z.; Tóth, R. SMOKE: Single-Stage Monocular 3D Object Detection via Keypoint Estimation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, Seattle, WA, USA, 13–19 June 2020. Available online: http://arxiv.org/abs/2002.10111 (accessed on 29 November 2023).
54. Wang, T.; Zhu, X.; Pang, J.; Lin, D. Fcos3d: Fully convolutional one-stage monocular 3d object detection. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021. Available online: https://paperswithcode.com/paper/fcos3d-fully-convolutional-one-stage (accessed on 11 January 2023).
55. Shapii, A.; Pichak, S.; Mahayuddin, Z.R. 3D Reconstruction Technique from 2d Sequential Human Body Images in Sports: A Review. *Technol. Rep. Kansai Univ. 2020, 62, 4973–4988.* Available online: https://www.researchgate.net/publication/345392953 (accessed on 13 February 2023).
56. Wang, Y.; Chao, W.-L.; Garg, D.; Hariharan, B.; Campbell, M.; Weinberger, K.Q. Pseudo-LiDAR from Visual Depth Estimation: Bridging the Gap in 3D Object Detection for Autonomous Driving. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019. Available online: http://arxiv.org/abs/1812.07179 (accessed on 29 November 2023).
57. You, Y.; Wang, Y.; Chao, W.L.; Garg, D.; Pleiss, G.; Hariharan, B.; Campbell, M.; Weinberger, K.Q. Pseudo-LiDAR++: Accurate Depth for 3D Object Detection in Autonomous Driving. 2019. Available online: http://arxiv.org/abs/1906.06310 (accessed on 29 November 2023).
58. Chen, Y.; Huang, S.; Liu, S.; Yu, B.; Jia, J. DSGN++: Exploiting Visual-Spatial Relation for Stereo-based 3D Detectors. In Proceedings of the IEEE Transactions on Pattern Analysis and Machine Intelligence. 2022. Available online: http://arxiv.org/abs/2204.03039 (accessed on 29 November 2023).
59. Li, P.; Chen, X.; Shen, S. Stereo R-CNN based 3D Object Detection for Autonomous Driving. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019. Available online: http://arxiv.org/abs/1902.09738 (accessed on 29 November 2023).
60. Qin, Z.; Wang, J.; Lu, Y. Triangulation Learning Network: From Monocular to Stereo 3D Object Detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019. Available online: http://arxiv.org/abs/1906.01193 (accessed on 29 November 2023).
61. Qi, C.R.; Liu, W.; Wu, C.; Su, H.; Guibas, L.J. Frustum PointNets for 3D Object Detection from RGB-D Data. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018. Available online: http://arxiv.org/abs/1711.08488 (accessed on 29 November 2023).
62. Wang, Z.; Jia, K. Frustum ConvNet: Sliding Frustums to Aggregate Local Point-Wise Features for Amodal. In Proceedings of the IEEE International Conference on Intelligent Robots and Systems, Macau, China, 3–8 November 2019; pp. 1742–1749. [CrossRef]
63. Shin, K.; Kwon, Y.P.; Tomizuka, M. RoarNet: A Robust 3D object detection based on region approximation refinement. In Proceedings of the IEEE Intelligent Vehicles Symposium, Paris, France, 9–12 June 2019; pp. 2510–2515. [CrossRef]
64. Paigwar, A.; Sierra-Gonzalez, D.; Erkent, Ö.; Laugier, C. Frustum-PointPillars: A Multi-Stage Approach for 3D Object Detection Using RGB Camera and LiDAR. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021.
65. Du, X.; Ang, M.H.; Karaman, S.; Rus, D. A General Pipeline for 3D Detection of Vehicles. In Proceedings of the IEEE International Conference on Robotics and Automation (ICRA), Brisbane, QLD, Australia, 21–25 May 2018; pp. 3194–3200. [CrossRef]
66. Yan, Y.; Mao, Y.; Li, B. SECOND: Sparsely Embedded Convolutional Detection. *Sensors* **2018**, *18*, 3337. [CrossRef]
67. Vora, S.; Lang, A.H.; Helou, B.; Beijbom, O. PointPainting: Sequential Fusion for 3D Object Detection. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, Seattle, WA, USA, 13–19 June 2020; pp. 4604–4612.
68. Xu, S.; Zhou, D.; Fang, J.; Yin, J.; Bin, Z.; Zhang, L. FusionPainting: Multimodal Fusion with Adaptive Attention for 3D Object Detection. In Proceedings of the IEEE Conference on Intelligent Transportation Systems, Proceedings, ITSC, Indianapolis, IN, USA, 19–22 September 2021; pp. 3047–3054. [CrossRef]
69. Simon, M.; Amende, K.; Kraus, A.; Honer, J.; Samann, T.; Kaulbersch, H.; Milz, S.; Gross, H.M. Complexer-YOLO: Real-Time 3D Object Detection and Tracking on Semantic Point Clouds. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, Long Beach, CA, USA, 16–17 June 2019.
70. Meyer, G.P.; Charland, J.; Hegde, D.; Laddha, A.; Vallespi-Gonzalez, C. Sensor Fusion for Joint 3D Object Detection and Semantic Segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, Long Beach, CA, USA, 16–17 June 2019.
71. Wang, S.; Suo, S.; Ma, W.-C.; Pokrovsky, A.; Urtasun, R. Deep Parametric Continuous Convolutional Neural Networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 2589–2597.
72. Liang, M.; Yang, B.; Wang, S.; Urtasun, R. Deep Continuous Fusion for Multi-Sensor 3D Object Detection. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 641–656.

73. Sindagi, A.V.; Zhou, Y.; Tuzel, O. MVX-net: Multimodal VoxelNet for 3D object detection. In Proceedings of the IEEE International Conference on Robotics and Automation, Montreal, Canada, 20–24 May 2019; pp. 7276–7282. [CrossRef]
74. Li, Y.; Yu, A.W.; Meng, T.; Caine, B.; Ngiam, J.; Peng, D.; Shen, J.; Lu, Y.; Zhou, D.; Le, Q.V.; et al. DeepFusion: Lidar-Camera Deep Fusion for Multi-Modal 3D Object Detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 17182–17191. Available online: https://github.com/NVIDIA/semantic-segmentation (accessed on 13 January 2023).
75. Zhang, Y.; Chen, J.; Huang, D. CAT-Det: Contrastively Augmented Transformer for Multi-Modal 3D Object Detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 908–917.
76. Yoo, J.H.; Kim, Y.; Kim, J.; Choi, J.W. 3D-CVF: Generating Joint Camera and LiDAR Features Using Cross-view Spatial Feature Fusion for 3D Object Detection. *Lect. Notes Comput. Sci.* **2020**, *12372*, 720–736.
77. Chen, X.; Zhang, T.; Wang, Y.; Wang, Y.; Zhao, H. FUTR3D: A Unified Sensor Fusion Framework for 3D Detection. *arXiv* **2022**, arXiv:2203.10642. [CrossRef]
78. Liu, Z.; Tang, H.; Amini, A.; Yang, X.; Mao, H.; Rus, D.L.; Han, S. BEVFusion: Multi-Task Multi-Sensor Fusion with Unified Bird's-Eye View Representation. *arXiv* **2022**, arXiv:2205.13542. [CrossRef]
79. Chen, Z.; Li, Z.; Zhang, S.; Fang, L.; Jiang, Q.; Zhao, F.; Zhou, B.; Zhao, H. AutoAlign: Pixel-Instance Feature Aggregation for Multi-Modal 3D Object Detection. *Int. Jt. Conf. Artif. Intell.* **2022**, 827–833. [CrossRef]
80. Bai, X.; Hu, Z.; Zhu, X.; Huang, Q.; Chen, Y.; Fu, H.; Tai, C.L. TransFusion: Robust LiDAR-Camera Fusion for 3D Object Detection With Transformers. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 1090–1099.
81. Dou, J.; Xue, J.; Fang, J. SEG-VoxelNet for 3D vehicle detection from RGB and LiDAR data. In Proceedings of the 2019 International Conference on Robotics and Automation, Montreal, Canada, 20–24 May 2019; pp. 4362–4368. [CrossRef]
82. Chen, Y.; Li, H.; Gao, R.; Zhao, D. Boost 3-D Object Detection via Point Clouds Segmentation and Fused 3-D GIoU-L Loss. *IEEE Trans. Neural Netw. Learn. Syst.* **2022**, *33*, 762–773. [CrossRef]
83. Wang, C.; Ma, C.; Zhu, M.; Yang, X.; Key, M. PointAugmenting: Cross-Modal Augmentation for 3D Object Detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 11794–11803.
84. Xu, D.; Anguelov, D.; Jain, A. PointFusion: Deep Sensor Fusion for 3D Bounding Box Estimation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 244–253.
85. Huang, T.; Liu, Z.; Chen, X.; Bai, X. EPNet: Enhancing Point Features with Image Semantics for 3D Object Detection. In Proceedings of the Computer Vision–ECCV 2020: 16th European Conference, Springer Science and Business Media Deutschland GmbH, Glasgow, UK, 23–28 August 2020; pp. 35–52.
86. Xie, L.; Xiang, C.; Yu, Z.; Xu, G.; Yang, Z.; Cai, D.; He, X. PI-RCNN: An Efficient Multi-Sensor 3D Object Detector with Point-Based Attentive Cont-Conv Fusion Module. *Proc. AAAI Conf. Artif. Intell.* **2020**, *34*, 12460–12467. [CrossRef]
87. Wang, Z.; Zhao, Z.; Jin, Z.; Che, Z.; Tang, J.; Shen, C.; Peng, Y. Multi-Stage Fusion for Multi-Class 3D Lidar Detection. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Nashville, TN, USA, 20–25 June 2021; pp. 3120–3128.
88. Zhu, M.; Ma, C.; Ji, P.; Yang, X. Cross-Modality 3D Object Detection. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, Waikola, HI, USA, 3–8 January 2021; pp. 3772–3781.
89. Li, Y.; Qi, X.; Chen, Y.; Wang, L.; Li, Z.; Sun, J.; Jia, J. Voxel Field Fusion for 3D Object Detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 18–24 June 2022; pp. 1120–1129. Available online: https://github.com/dvlab-research/VFF (accessed on 30 May 2023).
90. Liang, M.; Yang, B.; Chen, Y.; Hu, R.; Urtasun, R. Multi-Task Multi-Sensor Fusion for 3D Object Detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 7345–7353.
91. An, P.; Liang, J.; Yu, K.; Fang, B.; Ma, J. Deep structural information fusion for 3D object detection on LiDAR–camera system. *Comput. Vis. Image Underst.* **2022**, *214*, 103295. [CrossRef]
92. Nabati, R.; Qi, H. CenterFusion: Center-based Radar and Camera Fusion for 3D Object Detection. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, Nashville, TN, USA, 20–25 June 2021. Available online: https://github.com/mrnabati/CenterFusion (accessed on 26 January 2023).
93. Nabati, R.; Qi, H. Radar-Camera Sensor Fusion for Joint Object Detection and Distance Estimation in Autonomous Vehicles. 2020. Available online: http://arxiv.org/abs/2009.08428 (accessed on 29 November 2023).
94. Nobis, F.; Geisslinger, M.; Weber, M.; Betz, J.; Lienkamp, M. A Deep Learning-based Radar and Camera Sensor Fusion Architecture for Object Detection. In Proceedings of the 2019 Symposium on Sensor Data Fusion: Trends, Solutions, Applications, SDF 2019, Bonn, Germany, 15–17 October 2019. [CrossRef]
95. Wang, L.; Chen, T.; Anklam, C.; Goldluecke, B. High Dimensional Frustum PointNet for 3D Object Detection from Camera, LiDAR, and Radar. In Proceedings of the IEEE Intelligent Vehicles Symposium, Las Vegas, NV, USA, 19 October–13 November 2020; pp. 1621–1628. [CrossRef]
96. Chen, X.; Huang, H.; Liu, Y.; Li, J.; Liu, M. Robot for automatic waste sorting on construction sites. *Autom. Constr.* **2022**, *141*, 104387. [CrossRef]

97. Gené-Mola, J.; Sanz-Cortiella, R.; Rosell-Polo, J.R.; Morros, J.-R.; Ruiz-Hidalgo, J.; Vilaplana, V.; Gregorio, E. Fruit detection and 3D location using instance segmentation neural networks and structure-from-motion photogrammetry. *Comput. Electron. Agric.* **2020**, *169*, 105165. [CrossRef]
98. Teng, P.; Zhang, Y.; Yamane, T.; Kogoshi, M.; Yoshida, T.; Ota, T.; Nakagawa, J. Accuracy Evaluation and Branch Detection Method of 3D Modeling Using Backpack 3D Lidar SLAM and UAV-SfM for Peach Trees during the Pruning Period in Winter. *Remote Sens.* **2023**, *15*, 408. [CrossRef]
99. Parmar, H.S.; Nutter, B.; Long, R.; Antani, S.; Mitra, S. Deep learning of volumetric 3D CNN for fMRI in Alzheimer's disease classification. In *Medical Imaging 2020: Biomedical Applications in Molecular, Structural, and Functional Imaging*; SPIE: Bellingham, WA, USA, 2020; Volume 11317, pp. 66–71. [CrossRef]
100. Wegmayr, V.; Aitharaju, S.; Buhmann, J. Classification of brain MRI with big data and deep 3D convolutional neural networks. In *Medical Imaging 2018: Computer-Aided Diagnosis*; SPIE: Bellingham, WA, USA, 2018; Volume 10575, pp. 406–412. [CrossRef]
101. Nie, D.; Zhang, H.; Adeli, E.; Liu, L.; Shen, D. 3D Deep Learning for Multi-modal Imaging-Guided Survival Time Prediction of Brain Tumor Patients. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2016: 19th International Conference, Athens, Greece, 17–21 October 2016*; Proceedings, Part II; Springer: Cham, Switzerland, 2016; pp. 212–220. [CrossRef]
102. Tang, Z.; Chen, K.; Pan, M.; Wang, M.; Song, Z. An Augmentation Strategy for Medical Image Processing Based on Statistical Shape Model and 3D Thin Plate Spline for Deep Learning. *IEEE Access* **2019**, *7*, 133111–133121. [CrossRef]
103. Han, C.; Kitamura, Y.; Kudo, A.; Ichinose, A.; Rundo, L.; Furukawa, Y.; Umemoto, K.; Li, Y.; Nakayama, H. Synthesizing Diverse Lung Nodules Wherever Massively: 3D Multi-Conditional GAN-Based CT Image Augmentation for Object Detection. In Proceedings of the 2019 International Conference on 3D Vision, 3DV 2019, Québec, Canada, 16–19 September 2019; pp. 729–737. [CrossRef]
104. Feng, M.; Gilani, S.Z.; Wang, Y.; Zhang, L.; Mian, A. Relation Graph Network for 3D Object Detection in Point Clouds. *IEEE Trans. Image Process.* **2021**, *30*, 92–107. [CrossRef]
105. Pan, X.; Xia, Z.; Song, S.; Li, L.E.; Huang, G. 3D Object Detection with Pointformer. In Proceedings of theIEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021.
106. Armeni, I. 3D Semantic Parsing of Large-Scale Indoor Spaces (a) Raw Point Cloud (b) Space Parsing and Alignment in Canonical 3D Space (c) Building Element Detection Enclosed Spaces. Available online: http://buildingparser.stanford.edu/ (accessed on 9 September 2023).
107. Princeton ModelNet. Available online: https://modelnet.cs.princeton.edu/ (accessed on 11 January 2023).
108. SHREC15. Non-Rigid 3D Shape Retrieval. Available online: https://www.icst.pku.edu.cn/zlian/representa/3d15/dataset/index.htm (accessed on 13 February 2023).
109. Wang, L.; Li, R.; Sun, J.; Liu, X.; Zhao, L.; Seah, H.S.; Quah, C.K.; Tandianus, B. Multi-View Fusion-Based 3D Object Detection for Robot Indoor Scene Perception. *Sensors* **2019**, *19*, 4092. [CrossRef]
110. Hua, B.-S.; Pham, Q.-H.; Nguyen, D.T.; Tran, M.-K.; Yu, L.-F.; Yeung, S.-K. SceneNN: A scene meshes dataset with aNNotations. In Proceedings of the 2016 4th International Conference on 3D Vision, 3DV, Stanford, CA, USA, 25–28 October 2016; pp. 92–101. [CrossRef]
111. Tao, C.; Gao, Z.; Yan, J.; Li, C.; Cui, G. Indoor 3D Semantic Robot VSLAM based on mask regional convolutional neural network. *IEEE Access* **2020**, *8*, 52906–52916. [CrossRef]
112. Guan, H.; Qian, C.; Wu, T.; Hu, X.; Duan, F.; Ye, X. A Dynamic Scene Vision SLAM Method Incorporating Object Detection and Object Characterization. *Sustainability* **2023**, *15*, 3048. [CrossRef]
113. Comba, L.; Biglia, A.; Aimonino, D.R.; Gay, P. Unsupervised detection of vineyards by 3D point-cloud UAV photogrammetry for precision agriculture. *Comput. Electron. Agric.* **2018**, *155*, 84–95. [CrossRef]
114. Ge, L.; Zou, K.; Zhou, H.; Yu, X.; Tan, Y.; Zhang, C.; Li, W. Three dimensional apple tree organs classification and yield estimation algorithm based on multi-features fusion and support vector machine. *Inf. Process. Agric.* **2022**, *9*, 431–442. [CrossRef]
115. Tu, H.; Wang, C.; Zeng, W. VoxelPose: Towards Multi-camera 3D Human Pose Estimation in Wild Environment. In Proceedings of the European Conference on Computer Vision, Springer Science and Business Media Deutschland GmbH, Glasgow, UK, 23–28 August 2020; pp. 197–212.
116. Belagiannis, V.; Amin, S.; Andriluka, M.; Schiele, B.; Navab, N.; Ilic, S. 3D Pictorial Structures for Multiple Human Pose Estimation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Columbus, OH, USA, 23–28 June 2014; pp. 1669–1676.
117. Joo, H.; Soo, H.; Sheikh, P.Y. MAP Visibility Estimation for Large-Scale Dynamic 3D Reconstruction. In Proceedings of the Computer Vision and Pattern Recognition Conference, Columbus, OH, USA, 23–28 June 2014. Available online: http://www.cs.cmu.edu/ (accessed on 29 November 2023).
118. Joo, H.; Liu, H.; Tan, L.; Gui, L.; Nabbe, B.; Matthews, I.; Kanade, T.; Nobuhara, S.; Sheikh, Y. Panoptic Studio: A Massively Multiview System for Social Motion Capture. In Proceedings of the IEEE International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019. Available online: http://www.cs.cmu.edu/ (accessed on 29 November 2023).
119. Liu, H.; Wu, J.; He, R. Center point to pose: Multiple views 3D human pose estimation for multi-person. *PLoS ONE* **2022**, *17*, e0274450. [CrossRef]

120. Ku, J.; Mozifian, M.; Lee, J.; Harakeh, A.; Waslander, S. Joint 3D Proposal Generation and Object Detection from View Aggregation. In Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Madrid, Spain, 1–5 October 2018. Available online: http://arxiv.org/abs/1712.02294 (accessed on 29 November 2023).
121. Yang, B.; Luo, W.; Urtasun, R. PIXOR: Real-time 3D Object Detection from Point Clouds. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018.
122. Computer Vision Group—Datasets—RGB-D SLAM Dataset and Benchmark. Available online: https://cvg.cit.tum.de/data/datasets/rgbd-dataset (accessed on 10 September 2023).
123. Kang, H.; Chen, C. Fruit detection, segmentation and 3D visualisation of environments in apple orchards. *Comput. Electron. Agric.* **2020**, *171*, 105302. [CrossRef]
124. Wu, G.; Li, B.; Zhu, Q.; Huang, M.; Guo, Y. Using color and 3D geometry features to segment fruit point cloud and improve fruit recognition accuracy. *Comput. Electron. Agric.* **2020**, *174*, 105475. [CrossRef]
125. Pretto, A.; Aravecchia, S.; Burgard, W.; Chebrolu, N.; Dornhege, C.; Falck, T.; Fleckenstein, F.V.; Fontenla, A.; Imperoli, M.; Khanna, R.; et al. Building an Aerial-Ground Robotics System for Precision Farming: An Adaptable Solution. *IEEE Robot. Autom. Mag.* **2021**, *28*, 29–49. [CrossRef]
126. Patil, A.K.; Balasubramanyam, A.; Ryu, J.Y.; N, P.K.B.; Chakravarthi, B.; Chai, Y.H. Fusion of multiple lidars and inertial sensors for the real-time pose tracking of human motion. *Sensors* **2020**, *20*, 5342. [CrossRef] [PubMed]
127. Trumble, M.; Gilbert, A.; Malleson, C.; Hilton, A.; Collomosse, J. Total Capture: 3D Human Pose Estimation Fusing Video and Inertial Sensors. In Proceedings of the 28th British Machine Vision Conference, London, UK, 21–24 November 2017; pp. 1–13. Available online: https://openresearch.surrey.ac.uk/esploro/outputs/conferencePresentation/Total-Capture-3D-Human-Pose-Estimation-Fusing-Video-and-Inertial-Sensors/99512708202346 (accessed on 1 June 2023).
128. Chen, Y.; Liu, S.; Shen, X.; Jia, J. DSGN: Deep Stereo Geometry Network for 3D Object Detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–18 June 2020.
129. Mousavian, A.; Anguelov, D.; Flynn, J.; Košecká, J. 3D Bounding Box Estimation Using Deep Learning and Geometry. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017.
130. Maxwell, A.E.; Warner, T.A.; Guillén, L.A. Accuracy assessment in convolutional neural network-based deep learning remote sensing studies—Part 1: Literature review. *Remote Sens.* **2021**, *13*, 2450. [CrossRef]
131. Hung, W.-C.; Kretzschmar, H.; Casser, V.; Hwang, J.-J.; Anguelov, D. LET-3D-AP: Longitudinal Error Tolerant 3D Average Precision for Camera-Only 3D Detection. 2022. Available online: http://arxiv.org/abs/2206.07705 (accessed on 29 November 2023).
132. Chen, X.; Jin, Z.; Zhang, Q.; Wang, P. Research on Comparison of LiDAR and Camera in Autonomous Driving. *J. Phys. Conf. Ser.* **2021**, *2093*, 012032. [CrossRef]
133. Wu, H.; Wen, C.; Shi, S.; Li, X.; Wang, C. Virtual Sparse Convolution for Multimodal 3D Object Detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and, and Pattern Recognition, Vancouver, BC, Canada, 17–24 June 2023. Available online: http://openaccess.thecvf.com/content/CVPR2023/html/Wu_Virtual_Sparse_Convolution_for_Multimodal_3D_Object_Detection_CVPR_2023_paper.html (accessed on 15 September 2023).
134. Li, X.; Ma, T.; Hou, Y.; Shi, B.; Yang, Y.; Liu, Y.; Wu, X.; Chen, Q.; Li, Y.; Qiao, Y.; et al. LoGoNet: Towards Accurate 3D Object Detection with Local-to-Global Cross-Modal Fusion. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 17–24 June 2023. Available online: https://github.com/sankin97/LoGoNet (accessed on 29 November 2023).
135. Wu, H.; Deng, J.; Wen, C.; Li, X.; Wang, C.; Li, J. CasA: A cascade attention network for 3-D object detection from LiDAR point clouds. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 1–11. Available online: https://ieeexplore.ieee.org/abstract/document/9870747/ (accessed on 17 September 2023).
136. Chen, J.; Wang, Q.; Peng, W.; Xu, H.; Li, X.; Xu, W. Disparity-Based Multiscale Fusion Network for Transportation Detection. *IEEE Trans. Intell. Transp. Syst.* **2022**, *23*, 18855–18863. [CrossRef]
137. Ye, Q.; Jiang, L.; Zhen, W.; Du, Y.; Chuxing, D. Consistency of Implicit and Explicit Features Matters for Monocular 3D Object Detection. *arXiv* **2022**, arXiv:2207.07933.
138. Hu, H.-N.; Yang, Y.-H.; Fischer, T.; Darrell, T.; Yu, F.; Sun, M. Monocular Quasi-Dense 3D Object Tracking. *IEEE Trans. Pattern Anal. Mach. Intell.* **2023**, *45*, 1992–2008. [CrossRef]