World Electric Vehicle Journal

MDPI

Review

A Review of 3D Object Detection for Autonomous Driving of Electric Vehicles

Deyun Dai, Zonghai Chen, Peng Bao and Jikai Wang *

Department of Automation, University of Science and Technology of China (USTC), Hefei 230027, China; daideyun@mail.ustc.edu.cn (D.D.); chenzh@ustc.edu.cn (Z.C.); baopeng@mail.ustc.edu.cn (P.B.)
* Correspondence: wangjk@ustc.edu.cn; Tel.: +86-0551-63601514

Abstract: In recent years, electric vehicles have achieved rapid development. Intelligence is one of the important trends to promote the development of electric vehicles. As a result, autonomous driving system is becoming one of the core systems of electric vehicles. Considering that environmental perception is the basis of intelligent planning and safe decision-making for intelligent vehicles, this paper presents a survey of the existing perceptual methods in vehicles, especially 3D object detection, which guarantees the reliability and safety of vehicles. In this review, we first introduce the role of perceptual module in autonomous driving system and a relationship with other modules. Then, we classify and analyze the corresponding perception methods based on the different sensors. Finally, we compare the performance of the surveyed works on public datasets and discuss the possible future research interests.

Keywords: electric vehicles; autonomous driving; deep learning; 3D object detection

Citation: Dai, D.; Chen, Z.; Bao, P.; Wang, J. A Review of 3D Object Detection for Autonomous Driving of Electric Vehicles. *World Electr. Veh. J.* **2021**, *12*, 139. https://doi.org/10.3390/wevj12030139

Academic Editor: Joeri Van Mierlo

Received: 20 July 2021
Accepted: 25 August 2021
Published: 30 August 2021

1. Introduction

In recent years, electric vehicles (EVs) are gaining increasingly more favor and attention. Environmental protection and fiscal return are the advantages of EVs. On one hand, the driving process of EVs does not pollute and destroy the environment. On the other hand, the comprehensive cost of EVs is lower than that of traditional vehicles under a same mileage. With the strengthening of the supporting infrastructure of EVs, the technology of electric vehicles has been developed and improved. In the process, safety, comfort, energy conservation and environmental protection are the direction and eternal theme of vehicles development. Electric, intelligent, and renewable are the effective measures and approaches to achieve the aim. The development of intelligent electric vehicles can improve the safety, comfort, and economy of vehicles. Furthermore, the ability of autonomous driving is helpful to solve urban traffic congestion and beneficial to combining with the intelligent transportation environment of future urban.

An autonomous driving system consists of perception, planning, decision, and control, which is illustrated in Figure 1. The perception subsystem is the basis for other subsystems. It takes data captured from different sensors as input to obtain vehicle's position and location, also including the size and direction of surrounding objects. Autonomous driving vehicles [1–3] are often equipped with a variety of sensors, including LiDARs, cameras, millimeter-wave radars, GPS, and so on, which are illustrated in Figure 2.

A perception subsystem needs to be accurate and robust to ensure safe driving. It is composed of several important modules, such as object detection, tracking, Simultaneous Localization and Mapping (SLAM), etc. Object detection is a fundamental ability and aims to detect all interested objects to achieve their location and categories from captured data, such as images or point clouds. Images are captured by cameras and can provide rich texture information. Cameras are cheap but cannot achieve accurate depth information, and they are sensitive to changes in illumination and weather, such as low luminosity at night-time and extreme brightness disparity when entering or leaving tunnels, rainy, or

snowy weather. Point clouds are captured by LiDARs and can provide accurate 3D spatial information. They are robust to weather and extreme lighting conditions and demonstrate sparsity and ununiformity in spatial distribution. In addition, LiDARs are expensive sensors. Therefore, considering the complementary characteristics between point clouds and images, cameras and LiDARs are used as indispensable sensors to ensure intelligent vehicles' driving safety.

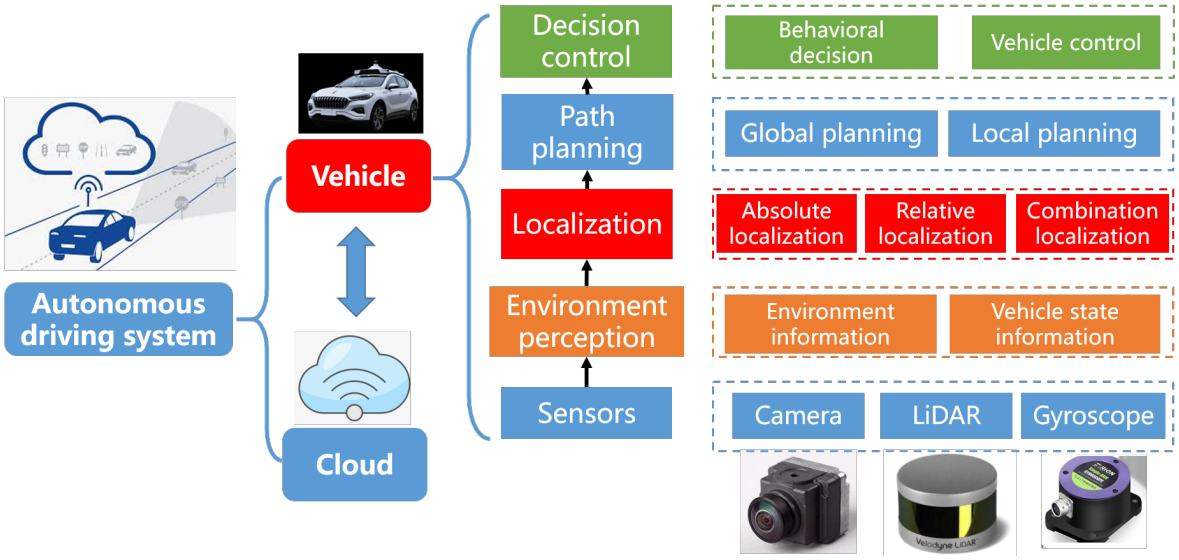

Figure 1. Illustration of autonomous driving system.

Figure 2. The autonomous car with multiple sensors including LiDAR, camera, GPS, and so on.

Notably, failure to detect objects might lead to safety-related incidents. It may result in traffic accidents, threatening human lives for failed detection of a leading vehicle [4]. To avoid collision with surrounding vehicles and pedestrians, object detection is an essential technique to analyze perceived images and point clouds, which needs to identify and localize objects. The general framework is illustrated in Figure 3. With the development of deep learning, 2D object detection is an extensive research topic in the field of computer

vision. CNN-based 2D object detections [5–8] have an excellent performance in some public datasets [9–11]. However, 2D object detection only provide 2D bounding boxes and can not provide depth information of objects that is crucial for safe driving. Compared with 2D object detection, 3D object detection provides more spatial information, such as location, direction, and object size, which makes it become more significant in autonomous driving. 3D detection needs to estimate more parameters for 3D-oriented boxes of objects, such as central 3D coordinates, length, width, height, and deflection angle of a bounding box. In addition, 3D object detection still faces arduous problems, including the complex interaction between objects, occlusion, changes in perspective and scale, and limited information provided by 3D data.

In this paper, we present a review of 3D object detection methods to summarize the development and challenges of 3D object detection. We analyze the potential advantages and limitations of these methods. The existing 3D object detection methods are divided into image-based methods, point cloud-based methods, and multimodal fusion-based methods. A general framework of the existing object detection methods is shown in Figure 3. The categories and their limitations are briefly described in Table 1.

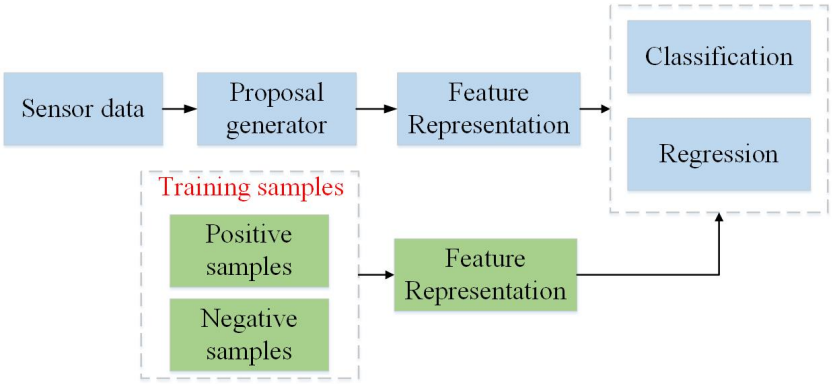

Figure 3. A general framework of object detection.

Table 1. Comparison of 3D object detection methods with different modes.

Mode		Methodology	Limitations
Image		Apply images to predict bounding boxes of 3D objects. 2D bounding boxes are predicted and then are extrapolated to 3D by reprojection constraints or regression model.	Depth information is deficient and the accuracy of detection results is low.
point cloud	Projection	Project a point cloud into a 2D plane and utilize 2D detection frameworks to regress 3D bounding boxes on projected images.	There is information loss in the process of projection.
	Volumetric	Conduct voxelization to achieve 3D voxels and generate representation by using convolutional operations in Voxels to predict 3D bounding boxes of objects.	Expensive 3D convolutional operations increase inference time. The computation is heavy.
	PointNet	Apply raw point cloud to predict 3D bounding boxes of objects directly.	Large scale of point cloud increases running time. It is difficult to generate region proposals.
Multi-sensor Fusion		Fuse image and point cloud to generate prediction on 3D bounding boxes. It is robust and complement each other.	Fusion methods are computationally expensive and are not mature enough.

2. Image-Based 3D Object Detection Methods

RGB-D images can provide depth information, which are used in some works. For example, Chen et al. [12] apply the poses of 3D bounding boxes to establish the energy function, and they use structured SVM for training to minimize the energy function. In DSS [13], multi-scale 3D RPN network is used to recommend objects on stereo images, which can detect objects of different sizes. Deng et al. [14] use the 2.5D method for object detection. They establish a model to detect 2D objects, and then convert 2D targets to 3D space to realize 3D object detection. Due to the large computation of RGB-D images, monocular images are used for 3D object detection.

In early days, Chen et al. propose Mono3d [15], which uses monocular images to generate 3D candidates, and then uses semantics, context information, hand-designed shape features, and location priors, which are illustrated in Figure 4, to score each candidates through energy model. Based on these candidates, Fast RCNN is used to further refine the 3D bounding boxes by location regression. The network improves the detection performance, but it is dependent on the object classes and needs a large number of candidates to achieve high recall, which leads to computational cost increase. To overcome this limitation, Pham and Jeon propose DeepStereoOP architecture [16] that is a class-independent algorithm, which exploits not noly RGB images but also depth inforamtion.

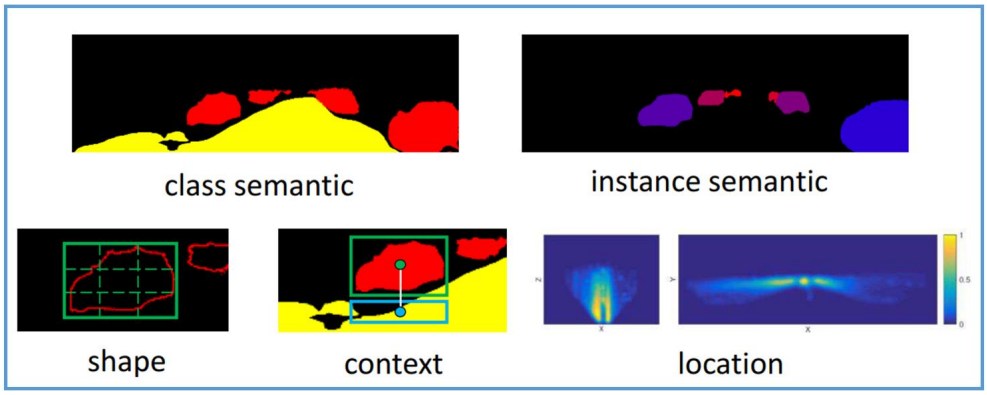

Figure 4. Information used in Mono3d.

Occlusion is a common phenomenon and also a great challenge in the driving environment. To alleviate this problem, Xiang et al. propose 3DVP [17] that introduces the 3D voxel patterns and uses RGB value, 3D shape and occlusion mask for appearance models. The 3D voxel patterns are illustrated in Figure 5. 3D detection is realized by minimizing the reprojection error between 3D frame projected on the image plane and 2D detection, which depends on the performance of regional recommendation network (RPN). Compared with the traditional region proposal methods, the region proposal network (RPN) can improve detection performance, but cannot deal with the problems of object scale changes, occlusion, and truncation. Therefore, SubCNN [18] uses subcategory information to generate region proposals and object candidates, where subcategories are objects with similar characteristics or attributes, such as 2D appearance, 3D pose or shape. A multi-scale image pyramid model is applied as the backbone network to improve the detection capability of small object. Although SubCNN improves the robustness to occlusion and truncation, the detection performance depnds on object categories.

Hu et al. [19] propose a multi-task framework to associate detections of objects in motion over time and estimate 3D bounding boxes information from a sequential images. They leverage 3D box depth-ordering matching for robust instance association and use 3D trajectory prediction for identification of occluded vehicles. Considering benefits from multi-task learning, Center3D [20] is proposed to efficiently estimate 3D location of objects and depth using only monocular images. It is an extension of CenterNet [21].

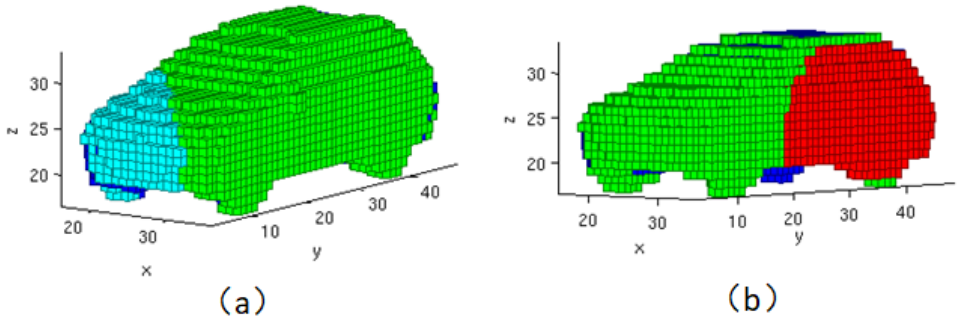

Figure 5. Illustration of 3D voxel pattern. Green of 3D voxel patterns represents visible in (**a**,**b**). Cyan of 3D voxel patterns represents truncated in (**a**). Red of 3D voxel patterns represents occluded in (**b**).

In recent years, 3D object detection from a 2D perspective has attracted the attention of many researchers. Lahoud and Ghanem [22] propose a 2D driven 3D object detection method to reduce the search space of 3D object. They apply manual features to train multi-layer perceptron network to predict 3D boxes. Later, they extend the work [23] and propose a multimodal region proposal network to generate region proposals, which uses an extended 2D boxes to generate 3D boxes. MonoDIS [24] leverages a novel disentangling transformation for 2D and 3D detection losses and a self-supervised confidence score for 3D bounding boxes.

Considering that depth information is helpful for 3D detection, pseudo-LiDAR is proposed based on stereo or monocular [25,26]. The depth map is first predicted and back-projecting is followed to generate a 3D point cloud in the LiDAR coordinate system. Wang et al. propose to convert image-based depth maps to pseudo-LiDAR representation, shown in Figure 6. Pseudo-LiDAR++ [27] is a 3D detection architecture based on pseudo-LiDAR. A depth-propagation algorithm is proposed based on initial depth estimates to diffuse these few exact measurements across the entire depth map. The architecture is independent of expensive LiDAR and performs almost on par with 64-beam LiDAR system but the depth map prediction is time-consuming. Moreover, there is a long tail problem when representing pseudo-LiDAR due to the inaccurate depth estimation around the boundaries of objects.

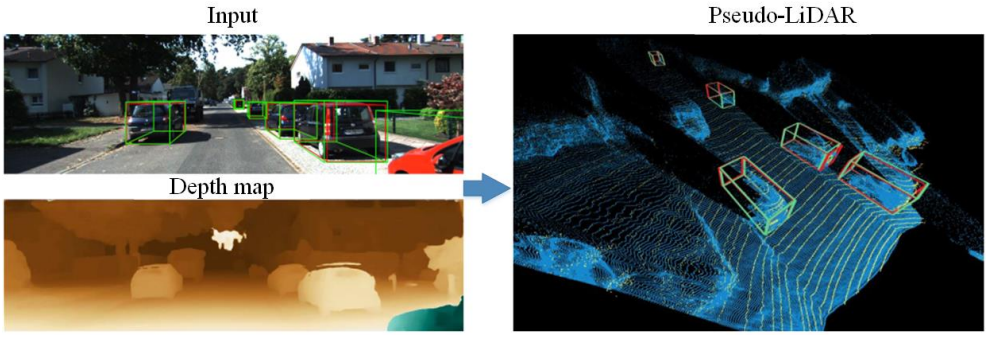

Figure 6. Pseudo-LiDAR point cloud from visual depth estimation.

Images can provide rich color and texture information. However, accuracte depth information cannot be obtained from images and depth information estimation with images has high computational cost and inaccury, which is necessary for accurate object size and location estimation, especially in the environments with occlusion and weak illumination.

3. Point Cloud-Based 3D Object Detection Methods

LiDAR sensors use laser beams to measure the distances of obstacles in the environment. The sensor outputs a set of 3D points. Compared with image-based methods, point clouds provide reliable depth information, which can be used to locate the object accurately. Unlike the structural information contained in the image, the LiDAR point cloud has the characteristics of disorder, sparsity, and limited information. Most Lidar-based object detection methods apply a two-stage strategy to detect object, which is illustrated Figure 7. In order to effectively utilize the point cloud, point cloud-based methods include 2D, 3D, and segmentation processing solutions.

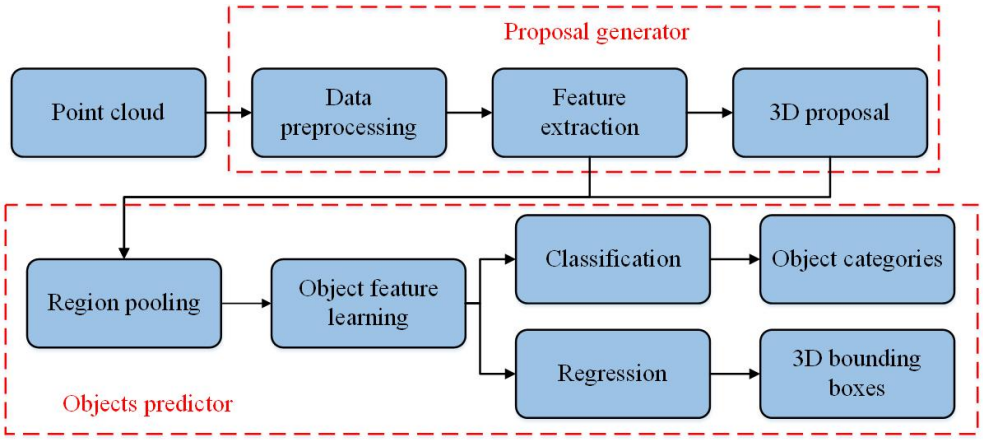

Figure 7. The flow chart of two-stage detection.

2D processing means that the point cloud is transformed into 2D planes. This kind of method does not directly process 3D point cloud data, but first projects the point cloud to some specific perspectives, such as the front view and bird's eye view (BEV), the projected images are shown in Figure 8. The pixels of the projected images are filled separately with density, average intensity, and height of each grid point as the RGB value of each pixel, which are inputed into the off-the-shelf 2D convolution network. The 3D boxes are predicted using the convolution features [28,29]. LMNet [30] projects the point cloud into the front view, and use FCN structure with dilated convolution [31] for single-stage object detection. The method can achieve a real-time detection performance, but low detection accuracy.

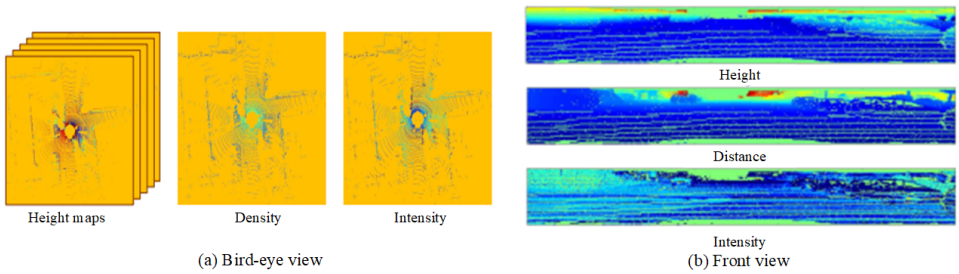

Figure 8. The projected images with different views of a point cloud.

Generally, BEV is adopted in AVs because the overlap between objects is little in BEV. The point cloud is projected into BEV with three channels (the channels are height, intensity, and density, respectively) in BirdNet [32]. Then, Faster R-CNN [33] is used to detect 2D directional bounding boxes of objects. Finally, directional 3D bounding boxes are obtained offline by combining 2D boxes and ground truth estimation. The method has a low efficiency. Based on the framework, BirdNet+ [34] utilizes ad hoc regression branches to eliminate the need for a postprocessing stage. RT3D [35] also uses a 2D object detection method to achieve 3D detection, in which point cloud is projected into BEV (the channels

are the maximum, average, and minimum height, respectively) and then R-FCN [36] is used to detect objects. The method improves the efficiency but has a low accuracy due to the loss of height information. Inspired by YoLo [37,38], Complex-YOLO [39] projects the point cloud into the BEV and then uses a single-stage strategy to estimate 3D bounding boxes of objects, which significantly improves the detection efficiency. Yange et al. [40] propose a single-stage detector which performs 2D convolution on the BEV. To avoid expensive computation of 3D CNNs, PointPillars [29] transform the point into vertical columns (pillars) and 2D operations is used to detect objects instead of 3D operations, which improves the computation efficiency.

Objects keep real physical dimensions and are naturally separable in BEV, but sparsity and variable point density of point cloud make great difficulties in detecting distant or small objects. To handle this problem, an end-to-end multiview fusion method [41] is proposed to synergize the BEV and the perspective view, which can effectively use the complementary information from both. In the method, dynamic voxelization is utilized to replace hard voxelization (HV), which eliminates the need to pad voxels to a predefined size and decreases the extra space and compute overhead of HV.

Range view (RV) is also a popular view in autonomous driving [42,43], which is shown in Figure 9. RangerCNN [44] and RangeNet++ [45] use 2D CNNs to achieve accurate 3D object detection based on range image representation, but they are subjected to the problem of scale variation. RangeIoUDet [46] learns point-wise features from the range image, which optimizes the point-wise feature and the 3D boxes by the point-based IoU and box-based IoU supervision. Although these methods are efficient, there is inevitable loss of spatial information. To improve the detection performance, MVFuseNet [47] fuses RV and BEV for spatiotemporal feature learning from a temporal sequence of LiDAR data to jointly perform both object detection and motion forecasting.

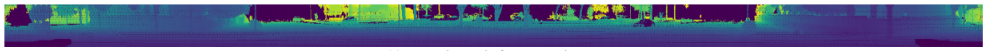

Figure 9. Range image.

3D processing directly uses the raw point cloud as the network input to extract the suitable point cloud features. For example, 3D FCN [48] and Vote3Deep [49] directly use a 3D convolution network to detect 3D bounding boxes of objects. However, the point cloud is sparse and the computation of 3D CNN is expensive. Additionally, affected by the receptive field, the traditional 3D convolution network cannot effectively learn the local features of different scales. To learn more effective spatial geometric representation from point cloud, some specific network frameworks have been proposed for point cloud, such as PointNet [50], PointNet++ [51], PointCNN [52], Dynamic Graph CNN [53], and Point-GNN [54]. PointNets [50,51] can directly process LiDAR point clouds and extract point cloud features through the MaxPooling symmetric function to solve the disorder problem of points. The network architecture of PontNet is illustrated in Figure 10. Thanks to the networks, the performance of 3D object detection is improved but the computation of point-based methods is expensive, especially when the information of large scenes are captured by using Velodyne LiDAR HDL-64E and there are more than 100K points in one scan. Therefore, some preprocessing operations need to be conducted, such as downsampling.

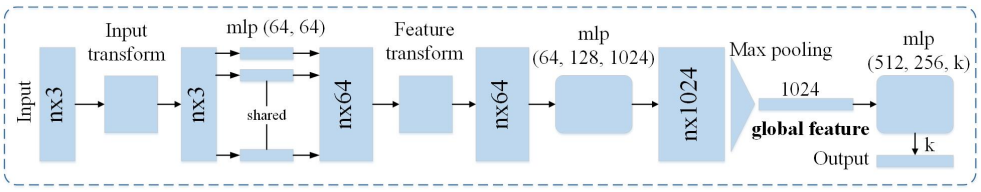

Figure 10. Architecture of PointNet.

After the point cloud features learning models are proposed, PointRCNN [55] constructs a PointNet++-based architecture to detect 3D objects, which is simply illustrated in Figure 11. Through the bottom-up 3D PRN, the subnetwork is used to transform the proposals information into standard coordinates to learn better local spatial features. By combining with the global semantic features of each point, the accuracy of the detected bounding boxes is improved. Similarly, Yang et al. [56] add a proposal generation module based on spherical anchor, which uses PointNet++ as the backbone network to extract semantic context features for each point. At the same time, in the second stage of boxes prediction, an IoU estimation branch is added for postprocessing, which further improves the accuracy of object detection.

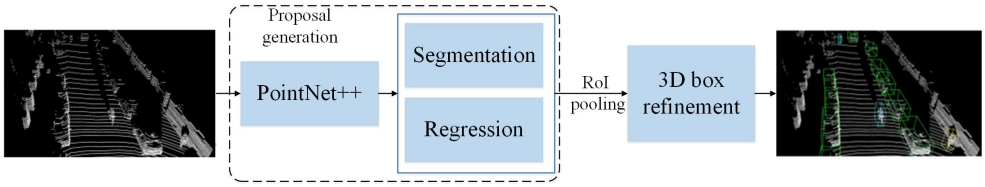

Figure 11. Architecture of PointRCNN.

Benifical from multi-task learning, LiDARMTL [57] utilizes a encoder–decoder architecture to predict perception parameters for 3D object detection and road understanding, which can be leveraged for online localization. Although the location accuracy of the object is improved compared with the previous methods, the calculation burden is heavy due to the large scale of the point cloud. To deal with the drawback, AFDet [58] adopts an anchor-free and Non-Maximum Suppresion-free single-stage framework to detect objects, which has the advantage in embedded systems.

Segmentation -based methods divide the point cloud into segments with spatial relationships, which implement voxelization in 3D space and extract features in grouped voxels by using 3D convolution [59]. The voxelization is showed in Figure 12. Based on the development and improvement of PointNet++, the VoxelNet [60] is proposed and widely used for 3D object detection [61]. The network first divides the 3D point cloud into a certain number of voxels. Then, after random sampling and normalization of points, the local features of each non empty voxel are extracted, and the geometric space representations of the objects are obtained. The RPN is utilized for classification and regression of 3D bounding boxes. Shi et al. [62] observe that the ground truth of the 3D bounding box not only provides the segmentation mask, but also provides the relative position information of its internal points, as is shown in Figure 13. Therefore, they propose a detection method that conducts sparse convolution and Voxel Set Abstraction to learn features. The segmentation mask and position information generated in the first stage are used as the features of the second stage, and then the ROI proposals are utilized for fine mapping of objects by sparse 3D convolution. PV-RCNN [63] deeply integrates both 3D voxel CNN and PointNet-based set abstraction to learn more discriminative point cloud features. The voxel CNN generates 3D proposals and RoI-grid pooling is leveraged to abstract proposal-specific features from the keypoint to the RoI-grid points via keypoint set abstraction, which encode rich context information. These methods improve the detection performance, but the quality of segmentation results affectes the detection results.

With the help of point cloud data, the performance of 3D object detection is significantly improved. In general, the accuracy of 3D bounding boxes with image-based methods is much less than point cloud-based methods. Currently, LiDAR point cloud-based 3D object detection has become a main trends, but point cloud cannot provide texture information to efficiently discrimate categories of objects. Moreover, the density of points decreases when the distance between object and LiDAR increases, which affects the performance of detectors, while images can still capture faraway objects. Therefore, multi-sensor fusion-based methods are proposed to improve the overall performance.

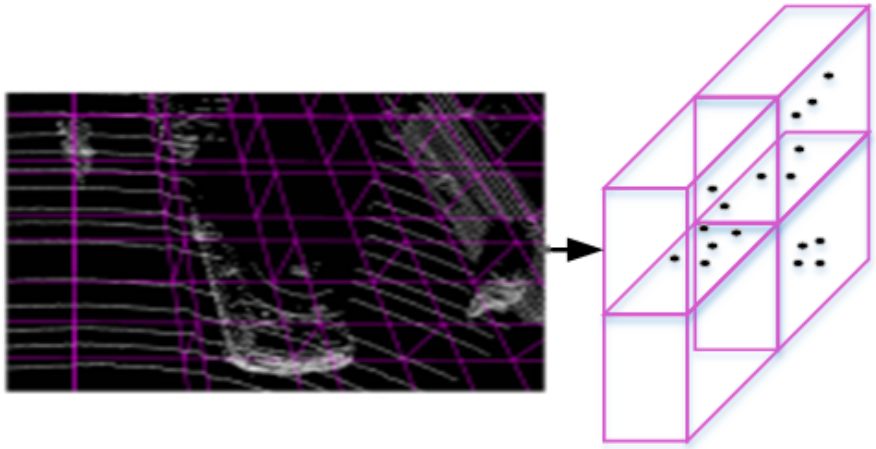

Figure 12. Voxelization of a point cloud.

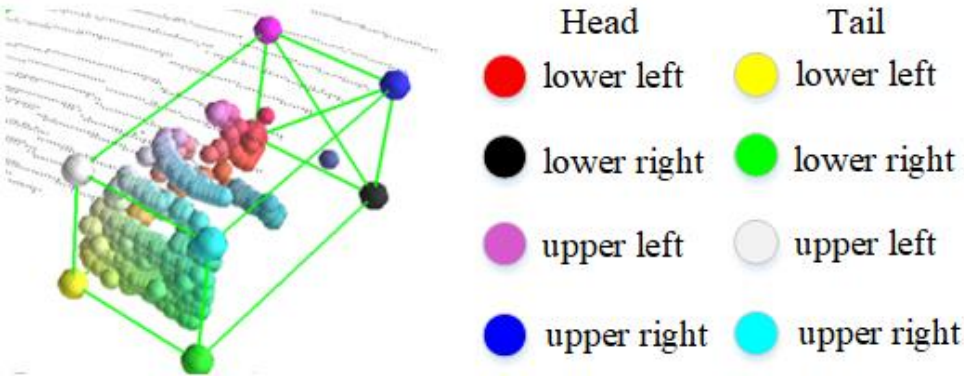

Figure 13. Intra-object part location and segmentation masks.

4. Multi-Sensor Fusion-Based 3D Object Detection Methods

Considering the advantages and disadvantages of image-based and point cloud-based methods, some methods try to apply fuse both modalities with different strategies. The fusion of LiDAR point cloud and images is done to conduct a projection transformation of the point cloud, and then to integrate the multi-view projected plane with the image by different feature fusion schemes, such as MV3D [64], AVOD [65], etc. There are three fusion schemes, including early fusion, late fusion, and deep fusion, which are illustrated in Figure 14. MV3D aggregates features by using a deep fusion scheme, where feature maps can hierarchically interact with others. AVOD is the first approach to introduce early fusion. The features of each modality proposal are merged and a FC layer is followed to output category and coordinates of 3D box for each proposal. These methods lose space information in the projection transformation process and the detection performance of small targets is poor. In addition, ROI feature fusion only uses advanced features, and the sparsity of LiDAR point cloud limits the fusion-based methods.

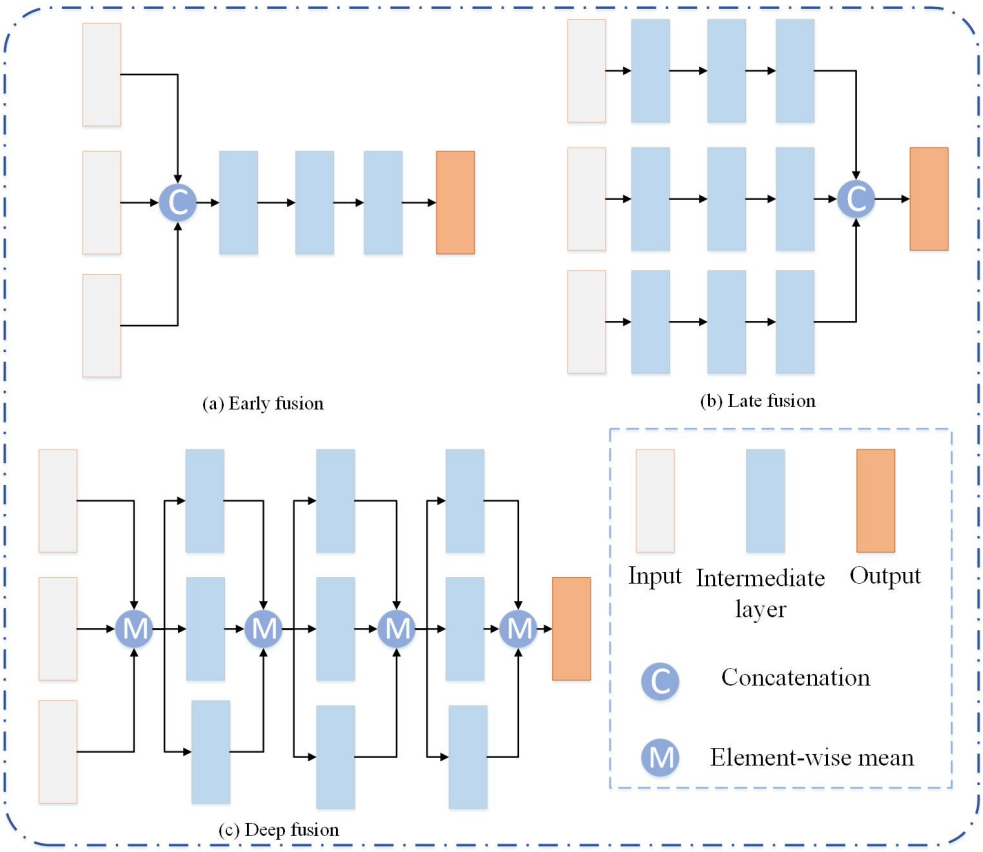

Figure 14. Architecture of different fusion schemes.

To solve these problems, Liang et al. [66] propose a feature fusion method combining point- and ROI-level features. Different from multi-view fusion, with the introduction of PointNets, the PointNets based fusion detection frameworks are proposed. F-PointNet [67] uses the detection results of mature 2D object detection to get the frustum spaces of objects. Then, it integrates PointNet++ to reduce the three-dimensional spaces of frustums, and returns the normalized coordinates. Finally, PointNet++ is used for the second time to realize the regression of the relevant parameters of the 3D bounding boxes of the objects. PointFusion [68] combines the advantages of AVOD and F-PointNet that extracts the features of RGB image blocks and corresponding raw point cloud, respectively. These features are fused and used to predict 3D bounding boxes with dense anchors. RoarNet [69] is similar to F-PointNet that apply two-stage strategy for 3D object detection, in which 3D proposals are first generated based on monucular image and then RoarNet_3D is used to directly process point cloud to estimate parameters of 3D bounding boxes. The method can deal with the asynchronous situation between LiDAR and camera. However, the detection accuracy depends on the recall of proposals and the undetected objects proposals can not be recovered at the second step.

Compared with projection transformation, some methods process raw point cloud directly. Gong et al. [70] also use a frustum model to integrate visual information and 3D spatial information, and combine visual and distance information into a probability framework. The method solves the problems of sparse and noise in LiDAR SLAM data, but it is not robust enough to dynamic objects. Similarly, SEG-VoxelNet [71] also uses mature 2D detection technology. The difference is that framework uses the current mature segmentation model to segment the image and integrates the semantic features obtained from segmentation and point cloud features based on VoxelNet. The 3D detection results of these network frameworks depend on the mature 2D detection methods, and the feature fusion is not enough. Therefore, Sindagi et al.[72] propose a multimodal information fusion

method, which combines early fusion of point features and late fusion of voxel features to fully integrate the LiDAR point cloud and image information.

To address the problem of information loss, 3D-CVF [73] combines the features of camera and LiDAR by using the cross-view spatial feature fusion strategy. Autocalibrated projection is applied to transform the image features to a smooth spatial feature map with the highest correspondence to the LiDAR features in the BEV domain. A gated feature fusion network is used mix the features appropriately. Additionally, the fusion methods based on BEV or voxel format are not accurate enough. Thus, PI-RCNN [74] proposes a novel fusion method named Point-based Attentive Cont-conv Fusion module to fuse multi-sensor features directly on 3D points. Except for continuous convolution, Point-Pooling and Attentive Aggregation are used to fuse features expressively.

In the process of 3D object detection, inconsistency between the localization and classification confidence is a critical issue [75]. To solve the problem, a consistency enforcing loss is utilized to increase the consistency of both the localization and classification in EPNet [76]. Moreover, the point features is enhanced with semantic image features in a point-wise manner without image annotations.

Besides fusion of camera and LiDAR, radar data are also used for 3D object detection [77,78]. CenterFusion [77] first associates radar detections to corresponding objects in the 3D space. Then, these radar detections are mapped into image plane to complement features of images in a middle-fusion method.

5. Evaluation

5.1. Datasets

A widely used dataset of 3D object detection for autonomous driving is KITTI [1], which provides RGB images, 3D velodyne point clouds, and GPS coordinates. These data are collected by a car equipped with a 64-channel LiDAR, 4 cameras, and a combined GPS/IMU system. The dataset is composed of 20 scenes, including cities, residential areas, and roads, as shown in Figure 15. In particular, the 3D object detection benchmark of KITTI consists of 7481 training images and 7518 test images as well as the corresponding point clouds, comprising a total of 80,256 labeled objects. It also contains sensor calibration information and annotated 2D and 3D bounding boxes of interested objects. The annotation of each object is classified as “easy”, “moderate”, and “hard” cases according to different difficulties.

Figure 15. Images of the KITTI dataset.

The nuScenes dataset [79] is anthor dataset for autonomous driving, and its scale is larger than KITTI. The datatset contains 700 scenes for training, 150 scenes for validation, and 150 scenes for testing. These data are collected by 6 cameras and a 32-beam LiDAR in Boston and Singapore. There are 23 classes in 360 degree field of view for 3D annotations. In the process of 3D object detection, some rare classes with few samples are removed

and 10 classes are retained for the task. There are 1000 driving scenes with dense traffic and greatly challenging driving situations. Some images captured front camera are shown in Figure 16. Moreover, annotations of objects contain some attributes such as visibility, activity, pose, etc.

Figure 16. Images of nuScenes dataset.

With the development of autonomous driving, datasets are developing very rapidly and many other datasets are also established, such as Waymo Open dataset [80], ApolloScape [81], H3D [82], AIO Drive [83], etc.

5.2. Metrics

IoU: Intersection-over-Union (IoU) is a common evaluation index. IoU is the overlap of predicted boxes and ground truth boxes, which is defined as follows:

$$IoU = \frac{\boldsymbol{B}_{pred} \cap \boldsymbol{B}_{gt}}{\boldsymbol{B}_{pred} \cup \boldsymbol{B}_{gt}}, \tag{1}$$

where $\boldsymbol{B}_{pred}$ represents the predicted 3D boxes, and $\boldsymbol{B}_{gt}$ is the ground truth.

AP: Generally, the average precision (AP) is selected to evaluate the performance of the algorithm. The definition of average precision is

$$AP = \frac{1}{N} \sum_{n=1}^{N} \max_{n \le i \le N} p(r_i), \tag{2}$$

where $p(r_i)$ represents a precision when recall is r_i. N is set as 11.

To compare detection performance of some methods on nuScenes datasets, other metrics also considered, including True Positive (TP)'s average translation, scale, orientation, velocity, and attribute error with ground-truth, denoted by ATE, ASE, AVE, and AAE, respectively. The final metrics is derived from a weighted sum of mAP and errors, which is a comprehensive comparison standard of detection performance.

5.3. Performance Comparison

We compare the detection results of some discussed methods in three difficulties of three categories (car, pedestrain, and cyclist). Table 2 shows the comparison results of the state-of-the-art methods on the KITTI object detection test set, in which accuracy and runtime are presented. Table 3 shows the detection results of the state-of-the-art methods on the nuScenes test set. Currently, safety is mainly taken into consideration, so

the final detection accuracy and inference efficiency is utilized to evaluate performance of existing methods.

Table 2. Performance comparison of 3D object detection on KITTI test set (Average precision in %).

Methods	M	Runtime (ms)/ Hardware	Car			Pedestrian			Cyclist		
			Easy	Moderate	Hard	Easy	Moderate	Hard	Easy	Moderate	Hard
Mono3D [15]	R	-/-	2.53	2.31	2.31	-	-	-	-	-	-
Deep3DBox [84]	R	-/-	5.84	4.09	3.83	-	-	-	-	-	-
OFT-Net [85]	R	-/-	3.28	2.50	2.27	1.06	1.11	1.06	0.43	0.43	0.43
MonoPair [86]	R	-/4 GPU@2.5 Ghz	13.04	9.99	8.75	-	-	-	-	-	-
E2E-PL P-RCNN [87]	S	490/1 GPU	64.8	43.9	38.1	-	-	-	-	-	-
DSGN [88]	S	113/Tesla V100	73.50	52.18	45.14	-	-	-	-	-	-
VoxelNet [18]	L	66/Titan X@1.7 Ghz	77.47	65.11	57.73	-	-	-	-	-	-
SECOND [28]	L	50/1 1080Ti@2.5 Ghz	83.34	72.55	65.82	-	-	-	-	-	-
PointPillars [29]	L	16.2/1 1080Ti@Intel i7	82.58	74.32	68.99	51.45	41.92	38.89	77.10	58.65	51.92
PointRCNN [55]	L	100/Titan XP@2.5 Ghz	86.96	75.64	70.70	47.98	39.37	36.01	74.96	58.82	52.53
TANet [89]	L	34.75/1 Titan V@2.5 Ghz	83.81	75.38	67.66	54.92	46.67	42.42	73.84	59.86	53.46
Voxel-FPN [90]	L	20/1 1080Ti@Intel i7	85.64	76.70	69.44	-	-	-	-	-	-
Fast PointRCNN [91]	L	65/Tesla P40@2.5 Ghz	85.29	77.40	70.24	-	-	-	-	-	-
Pathches [92]	L	150/1 1080Ti@2.5 Ghz	88.67	77.20	71.82	-	-	-	-	-	-
Part A^2 [62]	L	80/Tesla V100@2.5 Ghz	87.81	78.49	73.51	-	-	-	-	-	-
Point-GNN [54]	L	643/1070@2.5 Ghz	88.33	79.47	72.29	51.92	43.77	40.14	78.60	63.48	57.08
STD [56]	L	80/1 Titan V@2.5 Ghz	87.95	79.71	75.09	-	-	-	-	-	-
SA-SSD [93]	L	40/1 2080Ti@Intel i7	88.75	79.79	74.16	-	-	-	-	-	-
RangeIoUDet [46]	L	22/Tesla V100	88.60	79.80	76.76	-	-	-	83.12	67.77	60.26
Voxel R-CNN [94]	L	40/1 2080Ti@3.0 Ghz	90.90	81.62	77.06	-	-	-	-	-	-
MV3D [64]	R & L	240/Titan X	74.97	63.63	54.00	-	-	-	-	-	-
AVOD [65]	R & L	100/Titan XP	83.07	71.76	65.73	50.46	42.27	39.04	63.76	50.55	44.93
F-PointNet [67]	R & L	170/1 1080Ti	82.19	69.79	60.59	50.53	42.15	38.08	72.27	56.12	49.01
UberATG-ContFuse [95]	R & L	60/GPU@2.5 Ghz	82.54	66.22	64.04	-	-	-	-	-	-
MVX-Net [72]	R & L	-/-	83.2	72.7	65.2	-	-	-	-	-	-
RoarNet [69]	R & L	100/Titan X	83.95	75.79	67.88	-	-	-	-	-	-
UberATG-MMF [66]	R & L	80/GPU@2.5 Ghz .	88.40	77.43	70.22	-	-	-	-	-	-
3D-CVF [73]	R& L	75/1 1080Ti@2.5 Ghz	89.20	80.05	73.11	-	-	-	-	-	-

'M': Modality; 'R': RGB; 'S': Stereo; 'L': LiDAR.

Through comparison and analysis, image-based methods demonstrate low performance on 3D detection metrics due to the absence of depth information. Point cloud-based methods achieve significant improvement of performance on the task. Single-stage detection methods can achieve a fast inference but their accuracies cannot satisfy requirements of AVs. Two-stage methods can achieve high detection accuracy but the their efficiencies need to be improved. Additionally, in the current 3D object detection methods of autonomous driving, most fusion-based 3D object detection performance is lower than that based on point cloud due to the lack of mature and effective multi-sensor fusion strategy. Therefore, effective and robust fusion strategy is urgent and meaningful.

Table 3. Performance comparison on the nuScenes test set.

Method	mAP	NDS	Car	Truck	C.V.	Bus	Trailer	Barrier	Motor.	Bicycle	Ped.	T.C.
PointPillar [29]	30.5	45.3	68.4	23.0	4.1	28.2	23.4	38.9	27.4	1.1	59.7	30.8
3DSSD [96]	42.6	56.4	81.2	47.2	12.6	61.4	30.5	47.9	36.0	8.6	70.2	31.1
PointPainting [97]	46.4	58.1	77.9	35.8	15.8	36.2	37.3	60.2	41.5	24.1	73.3	62.4
CBGS [98]	52.8	63.3	81.1	48.5	10.5	54.9	42.9	65.7	51.5	22.3	80.1	70.9
CenterPoint [99]	60.3	67.3	85.2	53.5	20.0	63.6	56.0	71.1	59.5	30.7	84.6	78.4
FusionPainting [100]	66.3	70.4	86.3	58.5	27.7	66.8	59.4	70.2	71.2	51.7	87.5	84.2
PointAugmenting [101]	66.8	71.0	87.5	57.3	28.0	65.2	60.7	72.6	74.3	50.9	87.9	83.6

6. Conclusions

We review the current mainstream 3D object detection techniques and analyze the advantages and disadvantages of using RGB image, LiDAR point cloud, and image and point cloud fusion for 3D object detection. The performances of these methods are com-

pared on the KITTI public benchmark dataset. The directly use of LiDAR point cloud for detection provides a simple and effective solution for 3D object detection. However, due to the sparsity of point cloud and the lack of color information, it is necessary to use multi-modal information to overcome the problem of insufficient and incomplete single-modal information. Among the existing state-of-the-art methods, there is still a lack of mature multimodal information fusion detection frameworks. The research on multimodal fusion based detection methods is urgent and meaningful. By introducing the visual information, the representation of features can be enhanced to improve the recognition capacity of different objects. In addition, most of the current 3D object detection methods are based on a single frame. It is incomplete and insufficient due to the occlusion of objects. Therefore, it is meaningful to fuse the temporal information. The uncertainty of the information can be reduced and the detection accuracy can be improved by fusing the context information.

Author Contributions: Conceptualization, D.D. and J.W.; methodology, D.D.; formal analysis, D.D. and J.W.; investigation, D.D.; resources, D.D.; data curation, D.D.; writing—original draft preparation, D.D.; writing—review and editing, J.W., P.B.; visualization, D.D.; supervision, Z.C.; project administration, Z.C.; funding acquisition, Z.C. All authors have read and agreed to the published version of the manuscript.

Funding: This work is supported by the National Natural Science Foundation of China (Grant No. 91848111).

Data Availability Statement: Please refer to public KITTI 3D object detection benchmark (accessed on 5 August 2020) at http://www.cvlibs.net/datasets/kitti/eval_object.php?obj_benchmark=3d and public nuScenes detection leaderboard (accessed on 5 August 2020) at https://www.nuscenes.org/object-detection?externalData=all&mapData=all&modalities=Any.

Conflicts of Interest: All authors have read and approved the final version submitted. All authors listed have approved the manuscript and agree with its submission to special issue of World Electric Vehicle Journal. No conflict, financial or other, exists.

References

1. Geiger, A.; Lenz, P.; Urtasun, R. Are we ready for autonomous driving? The kitti vision benchmark suite. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Providence, RI, USA, 16–21 June 2012; pp. 3354–3361.
2. Agarwal, S.; Vora, A.; Pandey, G.; Williams, W.; Kourous, H.; McBride, J. Ford multi-AV seasonal dataset. *Int. J. Robot. Res.* **2020**, *39*, 1367–1376. [CrossRef]
3. Choi, Y.; Kim, N.; Hwang, S.; Park, K.; Yoon, J.S.; An, K.; Kweon, I.S. KAIST multi-spectral day/night data set for autonomous and assisted driving. *IEEE Trans. Intell. Transp. Syst.* **2018**, *19*, 934–948. [CrossRef]
4. Elmquist, A.; Negrut, D. *Technical Report TR-2016-13*; Simulation-Based Engineering Lab, University of Wisconsin-Madison: Madison, WI, USA, 2017.
5. Lin, T.Y.; Dollár, P.; Girshick, R.; He, K.; Hariharan, B.; Belongie, S. Feature pyramid networks for object detection. In Proceedings of the 30th IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 2117–2125.
6. Redmon, J.; Farhadi, A. Yolov3: An incremental improvement. *arXiv* **2018**, arXiv:1804.02767.
7. Wang, X.; Girshick, R.; Gupta, A.; He, K. Non-local neural networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 7794–7803.
8. Girshick, R. Fast r-cnn. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 7–13 December 2015; pp. 1440–1448.
9. Deng, J.; Dong, W.; Socher, R.; Li, L.J.; Li, K.; Fei-Fei, L. Imagenet: A large-scale hierarchical image database. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Miami, FL, USA, 20–25 June 2009; pp. 248–255.
10. Everingham, M.; Van Gool, L.; Williams, C.K.I.; Winn, J.; Zisserman, A. The pascal visual object classes (voc) challenge. *Int. J. Comput. Vis.* **2010**, *88*, 303–338. [CrossRef]
11. Lin, T.Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollar, P.; Zitnick, C.L. Microsoft coco: Common objects in context. In *European Conference on Computer Vision, Proceedings of the Computer Vision—ECCV 2014, 13th European Conference, Zurich, Switzerland, 6–12 September 2014*; Springer: Cham, Switzerland, 2014; pp. 740–755.
12. Chen, X.; Kundu, K.; Zhu, Y.; Berneshawi, A.; Ma, H.; Fidler, S.; Urtasun, R. 3d object proposals for accurate object class detection. In *Advances in Neural Information Processing Systems, Proceedings of the Neural Information Processing Systems 2015, Montreal, QC, Canada, 7–12 December 2015*; DBLP: Trier, Germany, 2015; pp. 424–432.
13. Song, S.; Xiao, J. Deep sliding shapes for amodal 3d object detection in rgb-d images. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 808–816.

14. Deng, Z.; Jan Latecki, L. Amodal detection of 3d objects: Inferring 3d bounding boxes from 2d ones in rgb-depth images. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 5762–5770.
15. Chen, X.; Kundu, K.; Zhang, Z.; Ma, H.; Fidler, S.; Urtasun, R. Monocular 3d object detection for autonomous driving. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 2147–2156.
16. Pham, C.C.; Jeon, J.W. Robust object proposals re-ranking for object detection in autonomous driving using convolutional neural networks. *Signal Process. Image Commun.* **2017**, *53*, 110–122. [CrossRef]
17. Xiang, Y.; Choi, W.; Lin, Y.; Savarese, S. Data-driven 3d voxel patterns for object category recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 1903–1911.
18. Xiang, Y.; Choi, W.; Lin, Y.; Savarese, S. Subcategory-aware convolutional neural networks for object proposals and detection. In Proceedings of the IEEE Winter Conference on Applications of Computer Vision (WACV), Santa Rosa, CA, USA, 24–31 March 2017; pp. 924–933.
19. Hu, H.N.; Cai, Q.Z.; Wang, D.; Lin, J.; Sun, M.; Krahenbuhl, P.; Darrell, T; Yu, F. Joint monocular 3D vehicle detection and tracking. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Korea, 27 October–2 November 2019; pp. 5390–5399.
20. Tang, Y.; Dorn, S.; Savani, C. Center3D: Center-based monocular 3D object detection with joint depth understanding. In *DAGM German Conference on Pattern Recognition, Proceedings of the DAGM GCPR 2020: Pattern Recognition, 42nd DAGM German Conference, DAGM GCPR 2020, Tübingen, Germany, 28 September–1 October 2020*; Proceedings 42; Springer International Publishing: Cham, Switzerland, 2021; pp. 289–302.
21. Duan, K.; Bai, S.; Xie, L.; Qi, H.; Huang, Q.; Tian, Q. Centernet: Keypoint triplets for object detection. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Korea, 27 October–2 November 2019; pp. 6569–6578.
22. Lahoud, J.; Ghanem, B. 2d-driven 3d object detection in rgb-d images. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 4622–4630.
23. Rahman, M.M.; Tan, Y.; Xue, J.; Shao, L.; Lu, K. 3D object detection: Learning 3D bounding boxes from scaled down 2D bounding boxes in RGB-D images. *Inf. Sci.* **2019**, *476*, 147–158. [CrossRef]
24. Simonelli, A.; Bulo, S.R.; Porzi, L.; Lopez-Antequera, M.; Kontschieder P. Disentangling monocular 3d object detection. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Korea, 27 October–2 November 2019; pp. 1991–1999.
25. Fu, H.; Gong, M.; Wang, C.; Batmanghelich, K.; Tao, D. Deep ordinal regression network for monocular depth estimation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 2002–2011.
26. Chang, J.R.; Chen, Y.S. Pyramid stereo matching network. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 5410–5418.
27. 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. *arXiv* **2019**, arXiv:1906.06310.
28. Yan, Y.; Mao, Y.; Li, B. Second: Sparsely embedded convolutional detection. *Sensors* **2018**, *18*, 3337. [CrossRef]
29. 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 Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 12697–12705.
30. Minemura, K.; Liau, H.; Monrroy, A.; Kato, S. LMNet: Real-time Multiclass Object Detection on CPU Using 3D LiDAR. In Proceedings of the 3rd Asia-Pacific Conference on Intelligent Robot Systems (ACIRS), Singapore, 21–23 July 2018; pp. 28–34.
31. Yu, F.; Koltun, V. Multi-scale context aggregation by dilated convolutions. *arXiv* **2015**, arXiv:1511.07122.
32. Beltrán, J.; Guindel, C.; Moreno, F.M.; Cruzado, D.; Garcia, F.; Escalera, A.D.L. Birdnet: A 3d object detection framework from lidar information. In Proceedings of the 21st International Conference on Intelligent Transportation Systems (ITSC), Maui, HI, USA, 4–7 November 2018; pp. 3517–3523.
33. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster r-cnn: Towards real-time object detection with region proposal networks. *Adv. Neural Inf. Process. Syst.* **2015**, *28*, 91–99. [CrossRef]
34. Barrera, A.; Guindel, C.; Beltrán, J.; Garcia, F. Birdnet+: End-to-end 3d object detection in lidar bird's eye view. In Proceedings of the IEEE 23rd International Conference on Intelligent Transportation Systems (ITSC), Rhodes, Greece, 20–23 September 2020; pp. 1–6.
35. Zeng, Y.; Hu, Y.; Liu, S.; Ye, J.; Han, Y.; Li, X.; Sun, N. Rt3d: Real-time 3-d vehicle detection in lidar point cloud for autonomous driving. *IEEE Robot. Autom. Lett.* **2018**, *3*, 3434–3440. [CrossRef]
36. Dai, J.; Li, Y.; He, K.; Sun, J. R-fcn: Object detection via region-based fully convolutional networks. *Adv. Neural Inf. Process. Syst.* **2016**, *29*, 379–387.
37. 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; pp. 779–788.
38. Shafiee, M.J.; Chywl, B.; Li, F.; Wong, A. Fast YOLO: A fast you only look once system for real-time embedded object detection in video. *arXiv* **2017**, arXiv:1709.05943.

39. Simony, M.; Milzy, S.; Amendey, K.; Gross, H.M. Complex-yolo: An euler-region-proposal for real-time 3d object detection on point clouds. In Proceedings of the European Conference on Computer Vision (ECCV) Workshops, Munich, Germany, 8–14 September 2018.
40. 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; pp. 7652–7660.
41. Zhou, Y.; Sun, P.; Zhang, Y.; Anguelov, D.; Gao, J.; Ouyang, T.; Guo, J.; Ngiam, J.; Vasudevan, V. End-to-end multi-view fusion for 3d object detection in lidar point clouds. In Proceedings of the Conference on Robot Learning, Virtual, 16–18 November 2020; pp. 923–932. Available online: https://proceedings.mlr.press/v100/zhou20a.html (accessed on 1 May 2020).
42. Wang, J.G.; Zhou, L.B. Traffic light recognition with high dynamic range imaging and deep learning. *IEEE Trans. Intell. Transp. Syst.* **2018**, *20*, 1341–1352. [CrossRef]
43. Bewley, A.; Sun, P.; Mensink, T.; Anguelov, D.; Sminchisescu, C. Range conditioned dilated convolutions for scale invariant 3d object detection. *arXiv* **2020**, arXiv:2005.09927.
44. Liang, Z.; Zhang, M.; Zhang, Z.; Zhao, X.; Pu, S. Rangercnn: Towards fast and accurate 3d object detection with range image representation. *arXiv* **2020**, arXiv:2009.00206.
45. Milioto, A.; Vizzo, I.; Behley, J.; Stachniss, C. Rangenet++: Fast and accurate lidar semantic segmentation. In Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Macau, China, 3–8 November 2019; pp. 4213–4220.
46. Liang, Z.; Zhang, Z.; Zhang, M.; Zhao X.; Pu, A. RangeIoUDet: Range Image Based Real-Time 3D Object Detector Optimized by Intersection Over Union. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Virtual, 19–25 June 2021; pp. 7140–7149.
47. Laddha, A.; Gautam, S.; Palombo, S.; Pandey, S.; Vallespi-Gonzalez, C. MVFuseNet: Improving End-to-End Object Detection and Motion Forecasting through Multi-View Fusion of LiDAR Data. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Virtual, 19–25 June 2021; pp. 2865–2874.
48. Li, B. 3d fully convolutional network for vehicle detection in point cloud. In Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Vancouver, BC, Canada, 24–28 September 2017; pp. 1513–1518.
49. Engelcke, M.; Rao, D.; Wang, D.Z.; Tong, C.H.; Posner, I. Vote3deep: Fast object detection in 3d point clouds using efficient convolutional neural networks. In Proceedings of the IEEE International Conference on Robotics and Automation (ICRA), Singapore, 29 May–3 June 2017; pp. 1355–1361.
50. 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; pp. 652–660.
51. Qi, C.R.; Yi, L.; Su, H.; Guibas, L.J. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. In *Advances in Neural Information Processing Systems, Proceedings of the NIPS 2017, Long Beach, CA, USA, 4–9 December 2017*; DBLP: Trier, Germany, 2017; pp. 5099–5108.
52. Li, Y.; Bu, R.; Sun, M.; Wu, W.; Di, X.; Chen, B. PointCNN: Convolution On X-Transformed Points. *Adv. Neural Inf. Process. Syst.* **2018**, *31*, 828–838.
53. Wang, Y.; Sun, Y.; Liu, Z.; Sarma, S.E.; Bronstein, M.M. Dynamic graph cnn for learning on point clouds. *arXiv* **2018**, arXiv:1801.07829.
54. Shi, W.; Rajkumar, R. Point-gnn: Graph neural network for 3d object detection in a point cloud. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Virtual, 14–19 June 2020; pp. 1711–1719.
55. Shi, S.; Wang, X.; Li, H. Pointrcnn: 3d object proposal generation and detection from point cloud. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 770–779.
56. Yang, Z.; Sun, Y.; Liu, S.; Shen, X.; Jia, J. Std: Sparse-to-dense 3d object detector for point cloud. In Proceedings of the IEEE International Conference on Computer Vision, Seoul, Korea, 27 October–2 November 2019; pp. 1951–1960.
57. Feng, D.; Zhou, Y.; Xu, C.; Tomizuka, M.; Zhan, W. A Simple and Efficient Multi-task Network for 3D Object Detection and Road Understanding. *arXiv* **2021**, arXiv:2103.04056.
58. Ge, R.; Ding, Z.; Hu, Y.; Wang, Y.; Chen, S.; Huang, L.; Li, Y. Afdet: Anchor free one stage 3d object detection. *arXiv* **2020**, arXiv:2006.12671.
59. Shi, S.; Jiang, L.; Deng, J.; Wang, Z.; Guo, C.; Shi, J.; Wang, X.; Li, H. PV-RCNN++: Point-Voxel Feature Set Abstraction with Local Vector Representation for 3D Object Detection. *arXiv* 2021, arXiv:2102.00463.
60. 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; pp. 4490–4499.
61. Ye, M.; Xu, S.; Cao, T. Hvnet: Hybrid voxel network for lidar based 3d object detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Virtual, 14–19 June 2020; pp. 1631–1640.
62. Shi, S.; Wang, Z.; Wang, X.; Li, H. Part-A^2 Net: 3D Part-Aware and Aggregation Neural Network for Object Detection from Point Cloud. *arXiv* **2019**, arXiv:1907.03670.
63. Shi, S.; Guo, C.; Jiang, L.; Wang, Z.; Shi, J.; Wang, X.; Li, H. Pv-rcnn: Point-voxel feature set abstraction for 3d object detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Virtual, 14–19 June 2020; pp. 10529–10538.
64. Chen, X.; Ma, H.; Wan, J.; Li, B.; Xia, T. Multi-view 3d object detection network for autonomous driving. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 1907–1915.

65. Ku, J.; Mozifian, M.; Lee, J.; Harakeh, A.; Waslander, A.L. 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; pp. 1–8.
66. Liang, M.; Yang, B.; Chen, Y.; Hu, R.; Urtasun, R. Multi-task multi-sensor fusion for 3d object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 7345–7353.
67. 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, 18–23 June 2018; pp. 918–927.
68. 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.
69. 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 (IV), Paris, France, 9–12 June 2019; pp. 2510–2515.
70. Gong, Z.; Lin, H.; Zhang, D.; Luo, Z.; Zelek, J.; Chen, Y.; Nurunnabi, A.; Wang, C.; Li, J. A Frustum-based probabilistic framework for 3D object detection by fusion of LiDAR and camera data. *ISPRS J. Photogramm. Remote Sens.* **2020**, *159*, 90–100. [CrossRef]
71. Dou, J.; Xue, J.; Fang, J. SEG-VoxelNet for 3D Vehicle Detection from RGB and LiDAR Data. In Proceedings of the International Conference on Robotics and Automation (ICRA), Montreal, QC, Canada, 20–24 May 2019; pp. 4362–4368.
72. Sindagi, V.A.; Zhou, Y.; Tuzel, O. MVX-Net: Multimodal voxelnet for 3D object detection. In Proceedings of the International Conference on Robotics and Automation (ICRA), Montreal, QC, Canada, 20–24 May 2019; pp. 7276–7282.
73. 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. In *European Conference on Computer Vision, Proceedings of the ECCV 2020: Computer Vision—ECCV 2020, 16th European Conference, Glasgow, UK, 23–28 August 2020*; Proceedings, Part XXVII 16; Springer International Publishing: Cham, Switzerland, 2020; pp. 720–736.
74. 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. In Proceedings of the AAAI Conference on Artificial Intelligence, New York, NY, USA, 7–12 February 2020; Volume 34, pp. 12460–12467.
75. Dai, D.; Wang, J.; Chen, Z.; Zhao, H. Image guidance based 3D vehicle detection in traffic scene. *Neurocomputing* **2021**, *428*, 1–11. [CrossRef]
76. Huang, T.; Liu, Z.; Chen, X.; Bai, X. Epnet: Enhancing point features with image semantics for 3d object detection. In *European Conference on Computer Vision, Proceedings of the ECCV 2020: Computer Vision—ECCV, 16th European Conference, Glasgow, UK, 23–28 August 2020*; Springer: Cham, Switzerland, 2020; pp. 35–52.
77. 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, Virtual, 5–9 January 2021; pp. 1527–1536.
78. Long, Y.; Morris, D.; Liu, X.; Castro, M.; Chakravarty, P.; Narayanan, P. Radar-Camera Pixel Depth Association for Depth Completion. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Virtual, 19–25 June 2021; pp. 12507–12516.
79. Caesar, H.; Bankiti, V.; Lang, A.H.; Vora, S.; Liong, V.E.; Xu, Q.; Krishnan, A.; Pan, Y.; Baldan, G.; Beijbom, O. nuscenes: A multimodal dataset for autonomous driving. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Virtual, 14–19 June 2020; pp. 11621–11631.
80. Sun, P.; Kretzschmar, H.; Dotiwalla, X.; Chouard, A.; Patnaik, V.; Tsui, P.; Guo, J.; Zhou, Y.; Chai, Y.; Caine, B.; et al. Scalability in perception for autonomous driving: Waymo open dataset. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Virtual, 14–19 June 2020; pp. 2446–2454.
81. 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.* **2019**, *42*, 2702–2719. [CrossRef]
82. Patil, A.; Malla, S.; Gang, H.; Chen, Y.T. The h3d dataset for full-surround 3d multi-object detection and tracking in crowded urban scenes. In Proceedings of the International Conference on Robotics and Automation (ICRA), Montreal, QC, Canada, 20–24 May 2019; pp. 9552–9557.
83. Weng, X.; Man, Y.; Cheng, D.; Park, J.; O'Toole, M.; Kitani, K. All-In-One Drive: A Large-Scale Comprehensive Perception Dataset with High-Density Long-Range Point Clouds. 2020, in submission. Available online: http://www.aiodrive.org/ (accessed on 5 August 2020).
84. Mousavian, A.; Anguelov, D.; Flynn, J.; Kosecka, 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; pp. 7074–7082.
85. Roddick, T.; Kendall, A.; Cipolla, R. Orthographic feature transform for monocular 3d object detection. *arXiv* **2018**, arXiv:1811.08188.
86. Chen, Y.; Tai, L.; Sun, K.; Li, M. Monopair: Monocular 3d object detection using pairwise spatial relationships. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Virtual, 14–19 June 2020; pp. 12093–12102.
87. Qian, R.; Garg, D.; Wang, Y.; You, Y.; Belongie, S.; Hariharan, B.; Campbell, M.; Weinberger, K.Q.; Chao, W.L. End-to-end pseudo-lidar for image-based 3d object detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Virtual, 14–19 June 2020; pp. 5881–5890.
88. 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, Virtual, 14–19 June 2020; pp. 12536–12545.

89. Liu, Z.; Zhao, X.; Huang, T.; Hu, R.; Zhou, Y.; Bai, X. Tanet: Robust 3d object detection from point clouds with triple attention. In Proceedings of the AAAI Conference on Artificial Intelligence, New York, NY, USA, 7–12 February 2020; Volume 34, pp. 11677–11684.
90. Kuang, H.; Wang, B.; An J.; Zhang, M.; Zhang, Z. Voxel-FPN: Multi-scale voxel feature aggregation for 3D object detection from LIDAR point clouds. *Sensors* **2020**, *20*, 704. [CrossRef]
91. Chen, Y.; Liu, S.; Shen, X.; Jia, J. Fast point r-cnn. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Korea, 27 October–2 November 2019; pp. 9775–9784.
92. Lehner, J.; Mitterecker, A.; Adler, T.; Hofmarcher, M.; Nessler, B.; Hochreiter, S. Patch Refinement—Localized 3D Object Detection. *arXiv* **2019**, arXiv:1910.04093.
93. He, C.; Zeng, H.; Huang, J.; Hua, X.; Zhang, L. Structure aware single-stage 3d object detection from point cloud. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Virtual, 14–19 June 2020; pp. 11873–11882.
94. Deng, J.; Shi, S.; Li, P.; Zhou, W.; Zhang, Y.; Li, H. Voxel R-CNN: Towards High Performance Voxel-based 3D Object Detection. *arXiv* **2020**, arXiv:2012.15712.
95. 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.
96. Yang, Z.; Sun, Y.; Liu, S.; Jia, J. 3dssd: Point-based 3d single stage object detector. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Virtual, 14–19 June 2020; pp. 11040–11048.
97. 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, Virtual, 14–19 June 2020; pp. 4604–4612.
98. Zhu, B.; Jiang, Z.; Zhou, X.; Li, Z.; Yu, G. Class-balanced grouping and sampling for point cloud 3d object detection. *arXiv* **2019**, arXiv:1908.09492.
99. Yin, T.; Zhou, X.; Krahenbuhl, P. Center-based 3d object detection and tracking. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Virtual, 19–25 June 2021; pp. 11784–11793.
100. Xu, S.; Zhou, D.; Fang, J.; Yin, J.; Bin, Z.; Zhang, L. FusionPainting: Multimodal Fusion with Adaptive Attention for 3D Object Detection. *arXiv* **2021**, arXiv:2106.12449.
101. Wang, C.; Ma, C.; Zhu, M.; Yang, X. PointAugmenting: Cross-Modal Augmentation for 3D Object Detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Virtual, 19–25 June 2021; pp. 11794–11803.