3D Object Detection for Autonomous Driving: A Practical Survey

Álvaro Ramajo-Ballester[a], Arturo de la Escalera Hueso[b] and José María Armingol Moreno[c]
Intelligent Systems Lab, University Carlos III de Madrid, Spain

Keywords: 3D Object Detection, Autonomous Driving, Deep Learning.

Abstract: Autonomous driving has been one of the most promising research lines in the last decade. Although still far off the sought-after level 5, the research community shows great advancements in one of the most challenging tasks: the 3d perception. The rapid progress of related fields like Deep Learning is one the reasons behind this success. This enables and improves the processing algorithms for the input data provided by LiDAR, cameras, radars and such other devices used for environment perception. With such growing knowledge, reviewing and structuring the state-of-the-art solutions becomes a necessity in order to correctly address future research directions. This paper provides a comprehensive survey of the progress of 3D object detection in terms of sensor data, available datasets, top-performing architectures and most notable frameworks that serve as a baseline for current and upcoming works.

1 INTRODUCTION

The field of autonomous driving, which aspires to provide the ability for vehicles to move safely with little to no human intervention, has advanced quickly in recent years. With the potential to reduce human error and improve road safety, autonomous driving techniques have been widely used in a variety of situations, such as self-driving vehicles in controlled scenarios (Marin-Plaza et al., 2021), delivery robots, *etc*.

Environment perception, a crucial element of these kind of systems, aids autonomous vehicles in understanding their surroundings through sensory input. Perception systems typically use multi-modality data as input (RGB images from cameras, point clouds from LiDAR, etc.) to predict the geometry and semantic details of significative elements on the road. As deep learning techniques in computer vision have advanced (Ramajo-Ballester et al., 2022), 3D object detection algorithms have evolved quickly. The proposed solutions follow very diverse methodologies, data distributions and evaluation metrics.

To overcome this lack of common ground, this work aims to gather the most notable approaches to offer a broad view of the current state of the art. To achieve that, this survey tries to build on top of already existing ones (Qian et al., 2022; Mao et al., 2022), updating their results with the latest and best performing models accross several benchmarks.

Specifically in the last years, a new research approach is emerging in this context, which is detecting these 3D objects from an infrastructure point of view. With sensors mounted several meters above the ground, the field of view increases significantly while also reducing occlusions between the elements on the road and the environment. This, of course, comes with its own challenges, like vehicle to infrastructure communication, protocol standardization across different devices and other related tasks that have to be taken into account.

2 BACKGROUND

2.1 Foundations / Basic Concepts

3D object detection aims to predict the attributes of 3D objects in driving scenarios from sensory inputs. In that context, let $\mathcal{X}$ denote input data, LiDAR or RGB images, for instance, and $\mathcal{F}$ denote a detector parameterized by θ. A general formula for 3D object detection can be represented as follows:

$$\mathcal{B} = \mathcal{F}(\mathcal{X}; \theta) \tag{1}$$

where $\mathcal{B} = \{B_1, B_2, ..., B_n\}$ is a set of n 3D objects in a scene. In this task, representing a 3D object is critical

[a] https://orcid.org/0000-0001-9425-9408
[b] https://orcid.org/0000-0002-2618-857X
[c] https://orcid.org/0000-0002-3353-9956

Ramajo-Ballester, Á., de la Escalera Hueso, A. and Armingol Moreno, J.
3D Object Detection for Autonomous Driving: A Practical Survey.
DOI: 10.5220/0011748400003479
In *Proceedings of the 9th International Conference on Vehicle Technology and Intelligent Transport Systems (VEHITS 2023)*, pages 64-73
ISBN: 978-989-758-652-1; ISSN: 2184-495X

since it affects what data should be provided for the subsequent prediction and planning processes. Typically, a 3D item is shown as a 3D cuboid that contains this object. This enclosure can be described by its 8 corners (Chen et al., 2017), 4 corners and heights (Ku et al., 2018) or, usually, by the 7 parameters for an oriented bounding box (Lang et al., 2019; Weng and Kitani, 2019) in eq. 2

$$\mathcal{B} = \{x, y, z, l, w, h, \theta, c\} \tag{2}$$

where (x, y, z) is the center coordinates of that cuboid; (l, w, h) are its length, width and height, respectively; θ, yaw angle in the ground plane and c, the corresponding class, say, car, pedestrian, *etc*.

2.2 Sensors

Raw data for 3D object detection can be provided by a wide variety of sensors. The two sensor types that are most frequently used are cameras and LiDAR (Light Detection And Ranging) sensors.

Cameras are economical, accessible, and capable of capturing scene details from a specific angle. They produce images $\mathcal{X}_{cam} = \mathbb{R}^{W \times H \times 3}$, where W, H are the image width and height and each pixel has 3 RGB channels. Although cheap, cameras have inherent limitations when it comes to 3D object detection. First of all, they are only capable of capturing visual data so they cannot obtain the three-dimensional structure of a scene. To overcome this problem, stereo cameras use matching algorithms to align correspondences in both left and right pictures for depth recovery (Chang and Chen, 2018).

On the other hand, LiDAR sensors (Velodyne, Ouster, RoboSense) can be used to get fine-grained 3D structures of a scene by producing a great number of laser beams and measuring their reflective information. A range image $\mathcal{X}_{lid} \in \mathbb{R}^{m \times n \times 3}$ can be produced by a LiDAR sensor using m beams and n readings in one scan cycle. Each pixel has 3 channels, corresponding to range, azimuth and inclination in the spherical coordinate system. Through the conversion of spherical coordinates into Cartesian coordinates, range images can be further transformed into point clouds.

2.3 Datasets

As the data-driven era progresses, the accessibility of large-scale datasets has been continuously fostering the community. Some of the most notable and publicly accessible datasets related to autonomous driving have been included. KITTI (Geiger et al., 2012; Geiger et al., 2013), Waymo Open (Sun et al., 2020) and nuScenes (Caesar et al., 2020) datasets are the typical examples and stand out as the most popular ones, among others. An example of the last two is shown in fig. 1 and a more detailed comparison is presented in table 1.

Figure 1: Waymo (above) and nuScenes (below) datasets examples.

KITTI Dataset. This is the pioneer dataset when it comes to 3D object detection was released in 2012. It contains the LiDAR and visual information from 15k frames, with more than 200k 3D annotations from 8 classes (car, van, truck, pedestrian, person, cyclist, tram, misc), although just car, pedestrian and cyclist labels are considered for the online scoreboard evaluation. Three difficulty levels (Easy, Moderate, and Hard) are introduced depending on the height of 2D bounding boxes, the level of occlusion and truncation.
Waymo Open. Waymo annotates 12M 3d bounding boxes among more than 200k frames. It comprises 1150 sequences (or scenes) over 4 classes, out of which only 3 correspond to the KITTI classes.
nuScenes. It manually labels 1.4M boxes among 40K frames. It collects 1000 sequences over 23 classes but only 10 of those are considered for detection.

Only the training and validation labels are available, as none of them offer the testing ones. Researchers are required to submit their predictions to the online leaderboard server for assessing the test

set. It is important to note that, in contrast to KITTI, which only collects data on sunny days, both nuScenes and Waymo Open acquire their data under a variety of weather (such as rainy, foggy, and snowy conditions) and lighting (daytime and nighttime).

In the last years, many dataset publications have been released with infrastructure-mounted sensors, like Rope3D (Ye et al., 2022b), A9 Dataset (Creß et al., 2022), IPS300+ (Wang et al., 2022) and DAIR-V2X (Yu et al., 2022). The latter is one of the most complete ones because it released three versions: Vehicle Dataset (DAIR-V2X-V), with 22325 LiDAR and image frames; Infrastructure Dataset (DAIR-V2X-I), including 10084 point cloud and image frames and Cooperative Dataset (DAIR-V2X-C), with 18330 data frames from infrastructure and 20515 from vehicle for Vehicle-Infrastructure Cooperative (VIC) 3D object detection.

2.4 Evaluation Metrics

The primary performance metric for 3D object detection is Average Precision (AP), which shares the same philosophy as its 2D counterparts (Everingham et al., 2010). Before identifying the subtle connections and differences of dataset-specific AP adopted among widely used benchmarks, the vanilla form of the AP metric is defined as follows:

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

where $P(r)$ is the precision-recall curve; r', each posible recall value and r the recall variable over which the integral is calculated. When calculating precision and recall, the main distinction with the 2D AP metric is the matching criterion between predictions and ground truths. In that regard, KITTI proposes two commonly-used AP metrics: AP3D and APBEV. AP3D matches the predicted objects to the respective ground truths if the 3D Intersection over Union (3D IoU) of two cuboids is above a certain threshold whereas APBEV is based on the IoU of two cuboids from the bird's-eye view (BEV IoU).

However, accurately calculating this area is not trivial, PASCAL VOC (Everingham et al., 2010) proposed an alternative metric: the interpolated $AP|_{R_N}$. It is formulated as the mean precision computed for each recall subset (R) of N evenly spaced recall levels, from $r_0 = 0$ to $r_1 = 1$:

$$AP|_{R_N} = \frac{1}{N} \sum_{r \in R} P(r) \tag{4}$$

where $R = [r_0, r_0 + \frac{r_1 - r_0}{N-1}, r_0 + \frac{2(r_1 - r_0)}{N-1}, \ldots, r_1]$ and $P(r) = \max_{r':r' \geq r} P(r')$.

KITTI Benchmark (Geiger et al., 2012). KITTI used the interpolated $AP|_{R_{11}}$ before changing to $AP|_{R_{40}}$ as suggested in (Simonelli et al., 2019). That resulted in a more fair comparison of the scores. It holds two separate leaderboards for 3D object detection and BEV detection. As previously said, it differentiates three levels of difficulty: easy, moderate and hard, regarding the occlusion and height of the bounding boxes.

Waymo Benchmark (Sun et al., 2020). Similarly, it proposes interpolated AP_{R21} and Average Precision weighted by heading (APH). To calculate that, Waymo evaluates on 21 evenly spaced recall levels ($r_0 = 0, r_1 = 1, N = 21$), with IOU threslholds of 0.7 for vehicles and 0.5 for pedestrians and cyclists. For APH, the true positives are weighted by the heading accuracy:

$$w_h = \min(|\hat{\theta} - \theta|, 2\pi - |\hat{\theta} - \theta|)/\pi \tag{5}$$

where $\hat{\theta}$ and θ are the predicted and ground truth azimuth, respectively.Two levels of dificulty are included in the benchmark: L1 for bounding boxes with more than five lidar points and L2 for those with between one and five points.

nuScenes Benchmark (Caesar et al., 2020). It uses a custom metric called NuScenes Detection Score (NDS). In order to compute it, a set of error metrics are defined for measuring translation (ATE), scale (ASE), orientation (AOE), velocity (AVE) and attribute errors (AAE). All of them have a 2m center distance threshold. The errors (ε) are converted to scores as shown in eq. 6.

$$s_i = \max(1 - \varepsilon_i, 0) \tag{6}$$

These metrics are weighted afterwards to calculate de NDS:

$$NDS = \frac{1}{10}\left[5mAP + \sum_i s_i\right] \tag{7}$$

where mAP is calculated by a BEV center distance with thresholds 0.5m, 1m, 2m, 4m (Qian et al., 2022).

3 REVIEW AND ANALYSIS

The core section of this work brings together some of the most notable contributions for the task of 3D object detection in autonomous driving context. It starts with camera-based methods, which only make use of visual information from the surroundings, either with a single or multiple views. After that, LiDAR-based approaches are reviewed, showing a clearly superior

Table 1: Comparison between publicly available datasets for 3D object detection, sorted by year.

Dataset	LiDAR	Images	3D annot.	Cl.	Night/Rain	View
KITTI (Geiger et al., 2012)	15k	15k	200k	8	No/No	Onboard
Ko-FAS (Strigel et al., 2014)	39k	19.4k	-	-	-/-	Infrastructure
KAIST (Choi et al., 2018)	8.9k	8.9k	-	3	Yes/No	Onboard
ApolloScape (Huang et al., 2019)	20k	144k	475k	6	-/-	Onboard
H3D (Patil et al., 2019)	27k	83k	1.1M	8	No/No	Onboard
Lyft L5 (Kesten et al., 2019)	46k	323k	1.3M	9	No/No	Onboard
Argoverse (Chai et al., 2021)	44k	490k	993k	15	Yes/Yes	Onboard
A*3D (Pham et al., 2020)	39k	39k	230k	7	Yes/Yes	Onboard
A2D2 (Geyer et al., 2020)	12.5k	41.3k	-	14	-/-	Onboard
nuScenes (Caesar et al., 2020)	400k	1.4M	1.4M	23	Yes/Yes	Onboard
Waymo Open (Sun et al., 2020)	230k	1M	12M	4	Yes/Yes	Onboard
AIODrive (Weng et al., 2020)	250k	250k	26M	-	Yes/Yes	Virtual onboard
BAAI-VANJEE (Yongqiang et al., 2021)	2.5k	5k	74k	12	Yes/Yes	Infrastructure
PandaSet (Xiao et al., 2021)	8.2k	49k	1.3M	28	Yes/Yes	Onboard
KITTI-360 (Liao et al., 2022)	80k	300k	68k	37	-/-	Onboard
Argoverse 2 Sensor (Wilson et al., 2021)	∼150k	∼1M	-	30	Yes/Yes	Onboard
ONCE (Mao et al., 2021)	1M	7M	417k	5	Yes/Yes	Onboard
Cirrus (Wang et al., 2021)	6.2k	6.2k	-	8	-/-	Onboard
Rope3D (Ye et al., 2022b)	-	50k	1.5M	12	Yes/Yes	Infrastructure
A9 Dataset (Creß et al., 2022)	1.7k	5.4k	215k	8	Yes/Yes	Infrastructure
IPS300+ (Wang et al., 2022)	28k	57k	4.5M	7	Yes/-	Infrastructure
DAIR-V2X (Yu et al., 2022)	71k	71k	1.2M	10	-/-	Onboard / Infrast.

performance than its image-only counterpart. Fusioning both sensors, multi-modal techniques offer the best results. These metrics are finally compared in section 3.4.

3.1 Image-Based Methods

A simple method for monocular 3D object detection, inspired by 2D detection techniques, is to directly regress the 3D box parameters from images using a convolutional neural network. This regression techniques can be fully trained and normally incorporate designs from the 2D detection network architectures.

This methods range from anchor-based approaches, where pre-defined 3D bounding boxes are predicted for posterior filtering and refinement (Luo et al., 2021; Kumar et al., 2021); anchor-free models, which predict the 3D objects attributes passing the images through multiple and separate heads (Reading et al., 2021; Zhou et al., 2021); stereo-based techniques, where depth is estimated with pixel disparity across images (Chang and Chen, 2018; Hartley and Zisserman, 2003) or from temporal and multi-view images, where the temporal information is leveraged to improve 3D object detection (Liu et al., 2022a; Liu et al., 2022b; Rukhovich et al., 2022).

3.2 LiDAR-Based Methods

This section covers those models using point clouds or range images for detection. They differ mainly on the data representations: point, voxel and pillar.

3.2.1 Point-Based Methods

Point cloud sampling and feature learning are the two fundamental elements of a point-based 3D object detector. These methods typically inherit the success of deep learning techniques on point clouds (Mao et al., 2019; Chen et al., 2022b; Chen et al., 2022a) and propose a variety of architectures to detect 3D objects directly from raw points. Point clouds are first forwarded through a point-based backbone network, where they are gradually sampled and the operators learn the features. Then, using the points and features from the downsampled data, 3D bounding boxes are predicted. For this kind of representation, (Zhang et al., 2022) offers an uncertainty estimator to take into account in the optimization process.

3.2.2 Voxel-Based Methods

Voxels are 3D cubes and contain points inside voxel cells. Most voxel cells in the 3D space are empty and lack any points due to the sparse distribution of the point cloud. Practical applications only store and use the non-empty voxels for feature extraction

Table 2: Comparison of the state-of-the-art 3D detection mAP on KITTI test set, sorted by vehicle moderate difficulty.

	Vehicle			Pedestrian			Cyclist		
Model	Easy	Mod	Hard	Easy	Mod	Hard	Easy	Mod	Hard
SFD (Wu et al., 2022b)	91.73	**84.76**	77.92	-	-	-	-	-	-
CasA++ (Wu et al., 2022a)	90.68	84.04	**79.69**	**56.33**	**49.29**	**46.70**	**87.76**	**73.79**	**66.94**
GraR-VoI (Yang et al., 2022a)	**91.89**	83.27	77.78	-	-	-	-	-	-
GLENet-VR (Zhang et al., 2022)	91.67	83.23	78.43	-	-	-	-	-	-
VPFNet (Zhu et al., 2022)	91.02	83.21	78.20	-	-	-	-	-	-
BtcDet (Xu et al., 2022)	90.64	82.86	78.09	-	-	-	82.81	68.68	61.81
DVF-V (Mahmoud et al., 2022)	89.40	82.45	77.56	-	-	-	-	-	-
RDIoU (Sheng et al., 2022)	90.65	82.30	77.26	-	-	-	-	-	-
Focals Conv-F (Chen et al., 2022c)	90.55	82.28	77.59	-	-	-	-	-	-
SASA (Chen et al., 2022a)	88.76	82.16	77.16	-	-	-	-	-	-
VoxSeT (He et al., 2022)	88.53	82.06	77.46	-	-	-	-	-	-

Table 3: Comparison of the state-of-the-art 3D detection methods on Waymo test set, sorted by vehicle L2 mAPH.

	Vehicle				Pedestrian				Cyclist			
	L1		L2		L1		L2		L1		L2	
Model	mAP	mAPH	mAP	mAPH	mAP	mAPH	mAP	mAPH	mAP	mAPH	mAP	mAPH
MPPNetEns-MMLab (Chen et al., 2022b)	87.77	87.37	**81.33**	**80.93**	**87.92**	**85.15**	**82.86**	**80.14**	80.74	79.90	78.54	77.73
BEVFusion-TTA (Liu et al., 2022c)	**87.96**	**87.58**	81.29	80.92	87.64	85.04	82.19	79.65	82.53	81.67	80.17	79.33
LidarMultiNet-TTA (Ye et al., 2022a)	87.64	87.26	80.73	80.39	87.75	85.07	82.48	79.86	**82.77**	**81.84**	**80.50**	**79.59**
DeepFusion-Ens (Li et al., 2022b)	86.45	86.09	79.43	79.09	86.14	83.77	80.88	78.57	80.53	79.80	78.29	77.58
AFDetV2-Ens (Hu et al., 2022)	85.80	85.41	78.71	78.34	85.22	82.16	79.71	76.75	81.20	80.30	78.70	77.83
CenterFormer (Zhou et al., 2022)	85.36	84.94	78.68	78.28	85.22	82.48	80.09	77.42	76.21	75.32	74.04	73.17
VISTA (Deng et al., 2022)	81.70	81.30	74.40	74.00	81.40	78.30	75.50	72.50	74.90	73.90	72.50	71.60
Graph-CE (Yang et al., 2022a)	80.77	80.28	72.55	72.10	82.35	76.64	74.44	69.02	75.28	74.21	72.52	71.49
BtcDet (Xu et al., 2022)	78.58	78.06	70.10	69.61	-	-	-	-	-	-	-	-
RDIoU (Sheng et al., 2022)	78.40	78.00	69.50	69.10	-	-	-	-	-	-	-	-
GLENet-VR (Zhang et al., 2022)	77.32	76.85	69.68	68.97	-	-	-	-	-	-	-	-
TransFusion (Bai et al., 2022)	-	-	-	65.10	-	-	-	64.00	-	-	-	67.40
DVF-V (PV-RCNN) (Mahmoud et al., 2022)	67.62	67.09	62.66	62.17	-	-	-	-	-	-	-	-

Table 4: Comparison of the state-of-the-art 3D detection methods on nuScenes test set. CV stands for construction vehicle and TC, traffic cone.

Model	mAP	NDS	Car	Truck	Bus	Trailer	CV	Ped	Motor	Bycicle	TC	Barrier
DeepInteraction-e (Yang et al., 2022b)	**75.6**	**76.3**	88.3	64.3	**74.2**	66.0	**44.7**	**92.5**	**85.4**	**66.4**	**90.9**	**83.5**
BEVFusion-e (Liu et al., 2022c)	75.0	76.1	**90.5**	**65.8**	74.2	**67.4**	42.6	91.8	84.4	62.9	89.4	81.1
Focals Conv-F (Chen et al., 2022c)	70.1	73.6	87.5	60.0	69.9	64.0	32.6	89.0	81.1	59.2	85.5	71.8
BEVFusion (Liang et al., 2022)	71.3	73.3	88.5	63.1	72.0	64.7	38.1	90.0	75.2	56.5	86.5	78.3
MSMDFusion-T (Jiao et al., 2022)	70.8	73.2	87.9	61.6	70.0	64.4	38.1	89.7	73.9	56.6	87.1	79.0
MDRNet-L (Huang et al., 2022)	68.4	72.8	87.9	58.5	67.3	64.1	30.2	89.0	77.0	50.7	85.0	74.7
AutoAlignV2 (Chen et al., 2022d)	68.4	72.4	87.0	59.0	69.3	59.3	33.1	87.6	72.9	52.1	85.1	78.0
TransFusion (Bai et al., 2022)	68.9	71.7	87.1	60.0	68.3	60.8	33.1	88.4	73.6	52.9	86.7	78.1
LidarMultiNet (Ye et al., 2022a)	67.0	71.6	86.9	57.4	64.7	61.0	31.5	87.2	75.3	47.6	85.1	73.5
UVTR-Multi. (Li et al., 2022a)	67.1	71.1	87.5	56.0	67.5	59.5	33.8	86.3	73.4	54.8	79.6	73.0
VISTA (Deng et al., 2022)	63.7	70.4	84.7	54.2	64.0	55.0	29.1	83.6	71.0	45.2	78.6	71.8
AFDetV2 (Hu et al., 2022)	62.4	68.5	86.3	54.2	62.5	58.9	26.7	85.8	63.8	34.3	80.1	71.0
SASA (Chen et al., 2022a)	45.0	61.0	76.8	45.0	66.2	36.5	16.1	69.1	39.6	16.9	29.9	53.6

(Rukhovich et al., 2022; Wu et al., 2022a; Xu et al., 2022; Zhou et al., 2022; Huang et al., 2022). Some of the most recent contributions use an anchor-free approach (Hu et al., 2022), a transformer-based backbone (He et al., 2022), global context pooling (Ye et al., 2022a) or offer a plug-in module for rotation-decoupled IoU optimization (Sheng et al., 2022).

3.2.3 Pillar-Based Methods

In pillars, the voxel size is special since it is unlimited in the vertical direction. Through the use of a PointNet [207], pillar features can be merged from points and then dispersed again to create a 2D BEV image for feature extraction. The pillar representation was first introduced in PointPillars (Lang et al., 2019), and

Table 5: Development frameworks used in all the reviewed works. PC stands for point cloud; I, for monocular image and ST, for stereo image.

Model	Year	Mode	Code
SFD	2022	PC+I	OpenPCDet
CasA++	2022	PC	OpenPCDet
GraR-VoI	2022	PC+I	MMDetection3D + OpenPCDet
GLENet-VR	2022	PC	-
VPFNet	2022	PC+ST	OpenPCDet
BtcDet	2022	PC	OpenPCDet
DVF-V	2022	PC+I	-
RDIoU	2022	PC	OpenPCDet
Focals Conv-F	2022	PC+I	OpenPCDet + CenterPoint
SASA	2022	PC	OpenPCDet
VoxSeT	2022	PC	OpenPCDet
DeepInteraction-e	2022	PC+I	MMDetection3D
BEVFusion	2022	PC+I	MMDetection3D
BEVFusion (2)	2022	PC+I	MMDetection3D + CenterPoint
MSMDFusion-T	2022	PC+I	MMDetection3D + TransFusion
MDRNet-L	2022	PC	-
AutoAlignV2	2022	PC+I	MMDetection3D
TransFusion	2022	PC+I	MMDetection3D + CenterPoint
LidarMultiNet	2022	PC	-
TransFusion-L	2022	PC	-
UVTR-Multimodality	2022	PC+I	MMDetection3D + Det3D
VISTA	2022	PC	Det3D + OpenPCDet + CenterPoint
AFDetV2-Ens	2022	PC	-
MPPNet	2022	PC	OpenPCDet
DeepFusion-Ens	2022	PC+I	Lingvo
CenterFormer	2022	PC	CenterPoint

was then developed further in (Fan et al., 2022), as shown in (Mao et al., 2022).

3.3 Multi-Modal Methods

For 3D object detection, camera and LiDAR are two complementary sensor types. When compared to a LiDAR sensor, which specializes in 3D localization and provides rich information of 3D structures, a camera provides color information from which rich semantic features can be extracted. That is the reason why the data fusion from these two sensors is the most common in the state of the art and why many attempts have been made to accurately detect 3D objects using a combination of camera and LiDAR data. These approaches are primarily based on LiDAR-based 3D object detectors and try to incorporate image information into various stages of the detection pipeline, given that LiDAR-based detection methods outperform camera-based methods significantly. Combining the two modalities inevitably results in more computational overhead and inference time latency due to the complexity of both detection systems (Mao et al., 2022).

This technique has been tried extensively (Wu et al., 2022b; Bai et al., 2022). Some variations include intermediate stage fusion (Yang et al., 2022b; Chen et al., 2022d; Li et al., 2022a; Li et al., 2022b), multi-stage fusion (Yang et al., 2022a; Jiao et al., 2022), BEV representation fusion (Liu et al., 2022c; Liang et al., 2022), sequential fusion (Mahmoud et al., 2022), LiDAR and stereo images fusion (Zhu et al., 2022) or focal convolution to replace sparse CNN (Chen et al., 2022c).

3.4 Performance Comparison

Following the guideline of this survey and adhering to the principle of building on previous works, we compare the most recent and best performing solution across the main three different benchmarks: KITTI (Geiger et al., 2012; Geiger et al., 2013), as shown in table 2; Waymo (Sun et al., 2020), in table 3, and finally nuScenes (Caesar et al., 2020), displayed in table 4.

In addition, a comparison between the different methods and frameworks that lie underneath the reviewed works is shown in table 5. MMDetection (OpenMMLab, 2020a) and OpenPCDet (OpenMMLab, 2020b) remain as the most popular baseline for LiDAR and image 3D object detection projects. Both property of OpenMMLab, although developed by different teams, they accumulate more than 3k start on GitHub and almost a thousand forks each. MMDetection3D becomes useful for both LiDAR and image detection and it even has well documented tutorials on how to use it for custom datasets. On the other hand, OpenPCDet focuses on LiDAR-based detection, with custom cuda code for faster processing.

Apart from those, CenterPoint (Yin et al., 2021) is another framework of great use, which is based on the formers and Det3D (Zhu et al., 2019) and lies ground to TransFusion (Bai et al., 2022). Whereas all of them are based on PyTorch (Paszke et al., 2019), Lingvo (Li et al., 2022b) is written natively on TensorFlow (Abadi et al., 2015).

4 CONCLUSIONS

This papers covers a wide variety of different methods that try to solve 3D object detection tasks and performs an extensive review of the current technology vanguard. By starting from basics concepts and going through the details of the available datasets, evaluation metrics and the latest architectures for image, point cloud and sensor fusion processing, it gives a clear and broad view of this research field. As it can be extracted from the results, the fusion and LiDAR-based techniques offer a more robust solution, clearly surpassing the performance of their monocular image counterparts. Nonetheless, the sensor cost is an important factor that could make the latter a viable alternative in certain environments.

ACKNOWLEDGEMENTS

Grant PID2019-104793RB-C31 and PDC2021-121517-C31 funded by MCIN/AEI/10.13039/50110 0011033 and by the European Union “NextGenerationEU/PRTR” and the Comunidad de Madrid through SEGVAUTO-4.0-CM (P2018/EMT-4362).

REFERENCES

Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G. S., Davis, A., Dean, J., Devin, M., Ghemawat, S., Goodfellow, I., Harp, A., Irving, G., Isard, M., Jia, Y., Jozefowicz, R., Kaiser, L., Kudlur, M., Levenberg, J., Mané, D., Monga, R., Moore, S., Murray, D., Olah, C., Schuster, M., Shlens, J., Steiner, B., Sutskever, I., Talwar, K., Tucker, P., Vanhoucke, V., Vasudevan, V., Viégas, F., Vinyals, O., Warden, P., Wattenberg, M., Wicke, M., Yu, Y., and Zheng, X. (2015). TensorFlow: Large-scale machine learning on heterogeneous systems. Software available from tensorflow.org.

Bai, X., Hu, Z., Zhu, X., Huang, Q., Chen, Y., Fu, H., and Tai, C.-L. (2022). Transfusion: Robust lidar-camera fusion for 3d object detection with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1090–1099.

Caesar, H., Bankiti, V., Lang, A. H., Vora, S., Liong, V. E., Xu, Q., Krishnan, A., Pan, Y., Baldan, G., and Beijbom, O. (2020). nuscenes: A multimodal dataset for autonomous driving. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11621–11631.

Chai, Y., Sun, P., Ngiam, J., Wang, W., Caine, B., Vasudevan, V., Zhang, X., and Anguelov, D. (2021). To the point: Efficient 3d object detection in the range image with graph convolution kernels. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16000–16009.

Chang, J.-R. and Chen, Y.-S. (2018). Pyramid stereo matching network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5410–5418.

Chen, C., Chen, Z., Zhang, J., and Tao, D. (2022a). Sasa: Semantics-augmented set abstraction for point-based 3d object detection. In *AAAI Conference on Artificial Intelligence*, volume 1.

Chen, X., Ma, H., Wan, J., Li, B., and Xia, T. (2017). Multi-view 3d object detection network for autonomous driving. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 1907–1915.

Chen, X., Shi, S., Zhu, B., Cheung, K. C., Xu, H., and Li, H. (2022b). Mppnet: Multi-frame feature intertwining with proxy points for 3d temporal object detection. *arXiv preprint arXiv:2205.05979*.

Chen, Y., Li, Y., Zhang, X., Sun, J., and Jia, J. (2022c). Focal sparse convolutional networks for 3d object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5428–5437.

Chen, Z., Li, Z., Zhang, S., Fang, L., Jiang, Q., and Zhao, F. (2022d). Autoalignv2: Deformable feature aggregation for dynamic multi-modal 3d object detection. *arXiv preprint arXiv:2207.10316*.

Choi, Y., Kim, N., Hwang, S., Park, K., Yoon, J. S., An, K., and Kweon, I. S. (2018). Kaist multi-spectral day/night data set for autonomous and assisted driving. *IEEE Transactions on Intelligent Transportation Systems*, 19(3):934–948.

Creß, C., Zimmer, W., Strand, L., Fortkord, M., Dai, S., Lakshminarasimhan, V., and Knoll, A. (2022). A9-

dataset: Multi-sensor infrastructure-based dataset for mobility research. In *2022 IEEE Intelligent Vehicles Symposium (IV)*, pages 965–970. IEEE.

Deng, S., Liang, Z., Sun, L., and Jia, K. (2022). Vista: Boosting 3d object detection via dual cross-view spatial attention. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8448–8457.

Everingham, M., Van Gool, L., Williams, C. K., Winn, J., and Zisserman, A. (2010). The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88(2):303–338.

Fan, L., Pang, Z., Zhang, T., Wang, Y.-X., Zhao, H., Wang, F., Wang, N., and Zhang, Z. (2022). Embracing single stride 3d object detector with sparse transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8458–8468.

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

Geiger, A., Lenz, P., and Urtasun, R. (2012). Are we ready for autonomous driving? the kitti vision benchmark suite. In *2012 IEEE conference on computer vision and pattern recognition*, pages 3354–3361. IEEE.

Geyer, J., Kassahun, Y., Mahmudi, M., Ricou, X., Durgesh, R., Chung, A. S., Hauswald, L., Pham, V. H., Mühlegg, M., Dorn, S., et al. (2020). A2d2: Audi autonomous driving dataset. *arXiv preprint arXiv:2004.06320*.

Hartley, R. and Zisserman, A. (2003). *Multiple view geometry in computer vision*. Cambridge university press.

He, C., Li, R., Li, S., and Zhang, L. (2022). Voxel set transformer: A set-to-set approach to 3d object detection from point clouds. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8417–8427.

Hu, Y., Ding, Z., Ge, R., Shao, W., Huang, L., Li, K., and Liu, Q. (2022). Afdetv2: Rethinking the necessity of the second stage for object detection from point clouds. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 1, pages 969–979.

Huang, D., Chen, Y., Ding, Y., Liao, J., Liu, J., Wu, K., Nie, Q., Liu, Y., and Wang, C. (2022). Rethinking dimensionality reduction in grid-based 3d object detection. *arXiv preprint arXiv:2209.09464*.

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

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

Kesten, R., Usman, M., Houston, J., Pandya, T., Nadhamuni, K., Ferreira, A., Yuan, M., Low, B., Jain, A., Ondruska, P., Omari, S., Shah, S., Kulkarni, A., Kazakova, A., Tao, C., Platinsky, L., Jiang, W., and Shet, V. (2019). Level 5 perception dataset 2020. https://level-5.global/level5/data/.

Ku, J., Mozifian, M., Lee, J., Harakeh, A., and Waslander, S. L. (2018). Joint 3d proposal generation and object detection from view aggregation. In *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 1–8. IEEE.

Kumar, A., Brazil, G., and Liu, X. (2021). Groomed-nms: Grouped mathematically differentiable nms for monocular 3d object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8973–8983.

Lang, A. H., Vora, S., Caesar, H., Zhou, L., Yang, J., and Beijbom, O. (2019). Pointpillars: Fast encoders for object detection from point clouds. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12697–12705.

Li, Y., Chen, Y., Qi, X., Li, Z., Sun, J., and Jia, J. (2022a). Unifying voxel-based representation with transformer for 3d object detection. *arXiv preprint arXiv:2206.00630*.

Li, Y., Yu, A. W., Meng, T., Caine, B., Ngiam, J., Peng, D., Shen, J., Lu, Y., Zhou, D., Le, Q. V., et al. (2022b). Deepfusion: Lidar-camera deep fusion for multi-modal 3d object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17182–17191.

Liang, T., Xie, H., Yu, K., Xia, Z., Lin, Z., Wang, Y., Tang, T., Wang, B., and Tang, Z. (2022). Bevfusion: A simple and robust lidar-camera fusion framework. *arXiv preprint arXiv:2205.13790*.

Liao, Y., Xie, J., and Geiger, A. (2022). Kitti-360: A novel dataset and benchmarks for urban scene understanding in 2d and 3d. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Liu, Y., Wang, T., Zhang, X., and Sun, J. (2022a). Petr: Position embedding transformation for multi-view 3d object detection. *arXiv preprint arXiv:2203.05625*.

Liu, Y., Yan, J., Jia, F., Li, S., Gao, Q., Wang, T., Zhang, X., and Sun, J. (2022b). Petrv2: A unified framework for 3d perception from multi-camera images. *arXiv preprint arXiv:2206.01256*.

Liu, Z., Tang, H., Amini, A., Yang, X., Mao, H., Rus, D., and Han, S. (2022c). Bevfusion: Multi-task multi-sensor fusion with unified bird's-eye view representation. *arXiv preprint arXiv:2205.13542*.

Luo, S., Dai, H., Shao, L., and Ding, Y. (2021). M3dssd: Monocular 3d single stage object detector. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6145–6154.

Mahmoud, A., Hu, J. S., and Waslander, S. L. (2022). Dense voxel fusion for 3d object detection. *arXiv preprint arXiv:2203.00871*.

Mao, J., Niu, M., Jiang, C., Liang, H., Chen, J., Liang, X., Li, Y., Ye, C., Zhang, W., Li, Z., et al. (2021). One million scenes for autonomous driving: Once dataset. *arXiv preprint arXiv:2106.11037*.

Mao, J., Shi, S., Wang, X., and Li, H. (2022). 3d object detection for autonomous driving: A review and new outlooks. *arXiv preprint arXiv:2206.09474*.

Mao, J., Wang, X., and Li, H. (2019). Interpolated convolutional networks for 3d point cloud understanding. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1578–1587.

Marin-Plaza, P., Yagüe, D., Royo, F., de Miguel, M. Á., Moreno, F. M., Ruiz-de-la Cuadra, A., Viadero-Monasterio, F., Garcia, J., San Roman, J. L., and Armingol, J. M. (2021). Project ares: Driverless transportation system. challenges and approaches in an unstructured road. *Electronics*, 10(15):1753.

OpenMMLab (2020a). MMDetection3D: OpenMMLab next-generation platform for general 3D object detection. https://github.com/open-mmlab/mmdetection3d.

OpenMMLab (2020b). Openpcdet: An open-source toolbox for 3d object detection from point clouds. https://github.com/open-mmlab/OpenPCDet.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al. (2019). Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32.

Patil, A., Malla, S., Gang, H., and Chen, Y.-T. (2019). The h3d dataset for full-surround 3d multi-object detection and tracking in crowded urban scenes. In *2019 International Conference on Robotics and Automation (ICRA)*, pages 9552–9557. IEEE.

Pham, Q.-H., Sevestre, P., Pahwa, R. S., Zhan, H., Pang, C. H., Chen, Y., Mustafa, A., Chandrasekhar, V., and Lin, J. (2020). A* 3d dataset: Towards autonomous driving in challenging environments. In *2020 IEEE International Conference on Robotics and Automation (ICRA)*, pages 2267–2273. IEEE.

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

Ramajo-Ballester, Á., González Cepeda, J., and Armingol Moreno, J. M. (2022). Vehicle re-identification in road environments using deep learning techniques. In *ITS European Congress 2022*, pages 363–374.

Reading, C., Harakeh, A., Chae, J., and Waslander, S. L. (2021). Categorical depth distribution network for monocular 3d object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8555–8564.

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

Sheng, H., Cai, S., Zhao, N., Deng, B., Huang, J., Hua, X.-S., Zhao, M.-J., and Lee, G. H. (2022). Rethinking iou-based optimization for single-stage 3d object detection. *arXiv preprint arXiv:2207.09332*.

Simonelli, A., Bulo, S. R., Porzi, L., López-Antequera, M., and Kontschieder, P. (2019). Disentangling monocular 3d object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1991–1999.

Strigel, E., Meissner, D., Seeliger, F., Wilking, B., and Dietmayer, K. (2014). The ko-per intersection laserscanner and video dataset. In *17th International IEEE Conference on Intelligent Transportation Systems (ITSC)*, pages 1900–1901. IEEE.

Sun, P., Kretzschmar, H., Dotiwalla, X., Chouard, A., Patnaik, V., Tsui, P., Guo, J., Zhou, Y., Chai, Y., Caine, B., et al. (2020). Scalability in perception for autonomous driving: Waymo open dataset. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2446–2454.

Wang, H., Zhang, X., Li, Z., Li, J., Wang, K., Lei, Z., and Haibing, R. (2022). Ips300+: a challenging multimodal data sets for intersection perception system. In *2022 International Conference on Robotics and Automation (ICRA)*, pages 2539–2545. IEEE.

Wang, Z., Ding, S., Li, Y., Fenn, J., Roychowdhury, S., Wallin, A., Martin, L., Ryvola, S., Sapiro, G., and Qiu, Q. (2021). Cirrus: A long-range bi-pattern lidar dataset. In *2021 IEEE International Conference on Robotics and Automation (ICRA)*, pages 5744–5750. IEEE.

Weng, X. and Kitani, K. (2019). Monocular 3d object detection with pseudo-lidar point cloud. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pages 0–0.

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

Wilson, B., Qi, W., Agarwal, T., Lambert, J., Singh, J., Khandelwal, S., Pan, B., Kumar, R., Hartnett, A., Pontes, J. K., et al. (2021). Argoverse 2: Next generation datasets for self-driving perception and forecasting. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*.

Wu, H., Deng, J., Wen, C., Li, X., Wang, C., and Li, J. (2022a). Casa: A cascade attention network for 3-d object detection from lidar point clouds. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–11.

Wu, X., Peng, L., Yang, H., Xie, L., Huang, C., Deng, C., Liu, H., and Cai, D. (2022b). Sparse fuse dense: Towards high quality 3d detection with depth completion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5418–5427.

Xiao, P., Shao, Z., Hao, S., Zhang, Z., Chai, X., Jiao, J., Li, Z., Wu, J., Sun, K., Jiang, K., et al. (2021). Pandaset: Advanced sensor suite dataset for autonomous driving. In *2021 IEEE International Intelligent Transportation Systems Conference (ITSC)*, pages 3095–3101. IEEE.

Xu, Q., Zhong, Y., and Neumann, U. (2022). Behind the curtain: Learning occluded shapes for 3d object detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 3, pages 2893–2901.

Yang, H., Liu, Z., Wu, X., Wang, W., Qian, W., He, X., and Cai, D. (2022a). Graph r-cnn: Towards accurate 3d object detection with semantic-decorated local graph. *arXiv preprint arXiv:2208.03624*.

Yang, Z., Chen, J., Miao, Z., Li, W., Zhu, X., and Zhang, L. (2022b). Deepinteraction: 3d object detection via modality interaction. *arXiv preprint arXiv:2208.11112*.

Ye, D., Zhou, Z., Chen, W., Xie, Y., Wang, Y., Wang, P., and Foroosh, H. (2022a). Lidarmultinet: Towards a unified multi-task network for lidar perception. *arXiv preprint arXiv:2209.09385*.

Ye, X., Shu, M., Li, H., Shi, Y., Li, Y., Wang, G., Tan, X., and Ding, E. (2022b). Rope3d: The roadside perception dataset for autonomous driving and monocular 3d object detection task. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21341–21350.

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

Yongqiang, D., Dengjiang, W., Gang, C., Bing, M., Xijia, G., Yajun, W., Jianchao, L., Yanming, F., and Juanjuan, L. (2021). Baai-vanjee roadside dataset: Towards the connected automated vehicle highway technologies in challenging environments of china. *arXiv preprint arXiv:2105.14370*.

Yu, H., Luo, Y., Shu, M., Huo, Y., Yang, Z., Shi, Y., Guo, Z., Li, H., Hu, X., Yuan, J., et al. (2022). Dair-v2x: A large-scale dataset for vehicle-infrastructure cooperative 3d object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21361–21370.

Zhang, Y., Zhang, Q., Zhu, Z., Hou, J., and Yuan, Y. (2022). Glenet: Boosting 3d object detectors with generative label uncertainty estimation. *arXiv preprint arXiv:2207.02466*.

Zhou, Y., He, Y., Zhu, H., Wang, C., Li, H., and Jiang, Q. (2021). Monocular 3d object detection: An extrinsic parameter free approach. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7556–7566.

Zhou, Z., Zhao, X., Wang, Y., Wang, P., and Foroosh, H. (2022). Centerformer: Center-based transformer for 3d object detection. In *European Conference on Computer Vision*, pages 496–513. Springer.

Zhu, B., Jiang, Z., Zhou, X., Li, Z., and Yu, G. (2019). Class-balanced grouping and sampling for point cloud 3d object detection. *arXiv preprint arXiv:1908.09492*.

Zhu, H., Deng, J., Zhang, Y., Ji, J., Mao, Q., Li, H., and Zhang, Y. (2022). Vpfnet: Improving 3d object detection with virtual point based lidar and stereo data fusion. *IEEE Transactions on Multimedia*.