TechRxiv
Powered by IEEE

Deep Learning-based Image 3D Object Detection for Autonomous Driving: Review

This paper was downloaded from TechRxiv (https://www.techrxiv.org).

LICENSE

CC BY 4.0

SUBMISSION DATE / POSTED DATE

05-08-2022 / 23-08-2022

CITATION

Alaba, Simegnew; Ball, John (2022): Deep Learning-based Image 3D Object Detection for Autonomous Driving: Review. TechRxiv. Preprint. https://doi.org/10.36227/techrxiv.20442858.v2

DOI

10.36227/techrxiv.20442858.v2

Deep Learning-based Image 3D Object Detection for Autonomous Driving: Review

Simegnew Yihunie Alaba, *Member, IEEE*, AND John E. Ball, *Senior Member, IEEE*

***Abstract*— An accurate and robust perception system is key to understanding the driving environment of autonomous driving and robots. Autonomous driving needs 3D information about objects, including the object's location and pose, to understand the driving environment clearly. A camera sensor is widely used in autonomous driving because of its richness in color, texture, and low price. The major problem with the camera is the lack of 3D information, which is necessary to understand the 3D driving environment. Additionally, the object's scale change and occlusion make 3D object detection more challenging. Many deep learning-based methods, such as depth estimation, have been developed to solve the lack of 3D information. This survey presents the image 3D object detection 3D bounding box encoding techniques, feature extraction techniques, and evaluation metrics of 3D object detection. The image-based methods are categorized based on the technique used to estimate an image's depth information, and insights are added to each method. Then, state-of-the-art (SOTA) monocular and stereo camera-based methods are summarized. We also compare the performance of the selected 3D object detection models and present challenges and future directions in 3D object detection.**

***Index Terms*— Autonomous Driving, Camera, Deep Learning, 3D Object Detection.**

I. Introduction

AUTONOMOUS driving and robot navigation should obtain 3D information of objects to understand the environment clearly. For fully autonomous driving, the perception system, such as 3D object detection, needs to be robust to work in adverse weather, accurate to give precise information about the driving environment, and enable fast decision making for high-speed driving [1]. Although 2D object detection has shown significant performance improvement in the computer vision community due to the rapid growth of deep learning (DL), 3D object detection is still a challenging problem due to the lack of 3D information on sensors, scale changes, occlusions, and others. A robust perception system, including 3D object detection, contributes to the development of fully autonomous driving, reducing fatalities caused by reckless human drivers. Building a perception system that is accurate to give precise information about the driving environment, fast to decide high-speed driving, and robust to work in inclement weather is crucial to achieving the goal of fully autonomous driving [1].

There are different 3D sensors available for 3D object detection, such as Light Detection and Ranging (LiDAR), radio detection and ranging (radar), and depth sensors (RGB-D cameras) [2]. The LiDAR sensor is a good choice for distance measurement. It is also more robust to inclement weather than a camera. However, the LiDAR data is unstructured and sparse, making LiDAR processing more challenging. Additionally, LiDAR is poor for color-based detection, and it is expensive. Radar is another 3D sensor for distance measurement and velocity estimation and is suitable for use in bad weather and night driving. However, it has low resolution, so radar-based object detection is poor. The camera sensor is inexpensive and rich in color and texture information. The major problem with a camera is the lack of high accuracy depth information. Different DL-based methods have been developed to solve this problem. The monocular camera's lack of depth information can be partially solved using a stereo camera [3], [4] or structure from motion. Predicting stereo instance segmentation is another technique to solve the monocular depth problem for 3D object detection [5]. Additionally, a few works convert the image into pseudo-LiDAR representation to solve the lack of depth information [6] (see details in section IV).

The major contributions of the paper are summarized as follows:

1) We provide an in-depth analysis of monocular and stereo image 3D object detection methods.
2) We summarize 3D bounding box encoding techniques and object detection evaluation metrics.
3) We categorize image 3D object detection methods based on the depth estimation techniques.
4) We present SOTA image 3D object detection methods for autonomous driving.

The rest of the paper is organized as follows. Section II provides related work. Object detection, especially 3D object detection, including object detection categories, 3D bounding box encoding techniques, and 3D object detection evaluation metrics, are summarized in section III. Section IV summarizes the image 3D object detection methods and compares the selected ones. The challenges and future directions are presented in Section V. The last section summarizes the survey paper.

Simegnew Yihunie Alaba is with the Department of Electrical and Computer Engineering, Mississippi State University, Starkville, MS 39762, USA (E-mail: sa1724@msstate.edu)

John E. Ball is with the Department of Electrical and Computer Engineering, Mississippi State University, Starkville, MS 39762, USA (E-mail: jeball@ece.msstate.edu)

II. Related Work

The rapid growth of DL enables feature learning from images rather than hand-crafted feature extractors, improving performance and facilitating the training process of object detection models. This work reviews DL-based image 3D object detection models for autonomous driving. Most survey papers presented image 3D object detection models with other works, such as LiDAR 3D object detection methods. However, a tremendous number of papers are published each year. So, in this work, we present a detailed analysis of image 3D object detection methods for autonomous driving. Therefore, we present SOTA methods and a comprehensive image 3D object detection analysis for autonomous driving.

Kim and Hwang [7] reviewed a survey on Monocular 3D object detection, but the works are not specifically for autonomous driving. They presented deep DL-based monocular 3D object detection methods and datasets. Feng *et al.* [1] reviewed 2D and 3D object detection and semantic segmentation for autonomous driving. The commonly used datasets and 2D/3D methods were reviewed. Jiao *et al.* [8] presented DL-based object detection methods, but not limited to autonomous driving. Additionally, the survey focused more on 2D object detection methods. Arnold *et al.* [2] briefly reviewed 3D object detection methods including LiDAR and image-based for autonomous driving. Similarly, Rahman *et al.* [9] presented 3D object detection methods for autonomous driving. Li *et al.* [10] and Guo *et al.* [11] presented DL-based object detection, segmentation, and classification in autonomous driving. Fernandes *et al.* [12] also reviewed DL based object detection and semantic segmentation for autonomous driving. Recently, Qian *et al.* [13] published a 3D object detection method for autonomous driving. In addition to the current SOTA methods, we have included 3D bounding box encoding techniques and 3D object detection evaluation techniques not covered by those survey papers. Datasets and sensors are reviewed in those surveys. Therefore, we opt not to include them in this survey.

III. Overview of Object Detection

This section presents object detection categories, evaluation metrics for object detection, and 3D bounding box encoding techniques.

A. Object Detection categories

Image-based 3D object detection models use 2D object detection as a base model and use different techniques, such as regression, to extend to 3D object detection. Thus, we review 2D object detection models to understand 3D object detection fully. DL-based general object detection methods can be classified into two groups: two-stage and one-stage detection. A two-stage object detection network has a region of interest (ROI) network for region proposal generation and the subsequent network for bounding box regression and classification, as shown in Fig. 2. R-CNN [14], SPPNet [15], Fast R-CNN [16], Faster R-CNN [17], RFCN [18], and Mask R-CNN [19] are examples of two-stage 2D object detection models. Girshick *et al.* [14] proposed R-CNN, a two-stage 2D object detection network, as shown in Fig. 1. Selective search

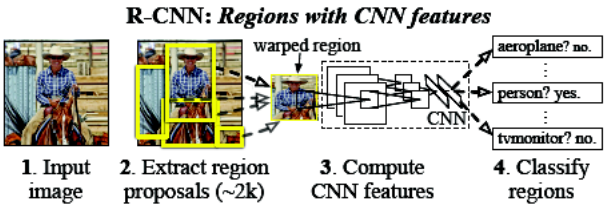

Fig. 1. R-CNN object detection system [14]. The system (1) takes an input image, (2) around 2000 bottom-up region proposals are extracted using selective search algorithm, (3) for each proposal, features are computed using CNN and feed to SVM classifier, and then (4) linear SVMs classifies each region.

algorithm [20] is used to generate 2,000 region proposals (candidate boxes), and then a CNN model is employed for feature extraction. The extracted features feed into support vector machines (SVM) to classify an object within the region proposals. The major limitation of R-CNN was the redundant generation of 2,000 bounding boxes from each image, increasing the network's computational burden. He *et al.* proposed Spatial Pyramid Pooling Networks (SPPNet) [15] to overcome this problem by introducing a spatial pyramid pooling layer, which generates a fixed-length representation of a region of interest (ROI). R-CNN and SPPNet train feature extraction and bounding box regression networks separately. So, the training takes a long time to process.

Girshick *et al.* proposed the Fast R-CNN [16] detector to solve the multistage training problem by simultaneously training the feature extraction and bounding box regression networks. Fast R-CNN also uses a selective search algorithm for proposal generations. The selective search algorithm increases the computational burden of the model because of the redundancy of proposal generation. So, Fast R-CNN's detection speed is low for real-time applications. To solve this problem, Faster R-CNN [17] uses a region proposal network instead of the selective search algorithm to generate region proposals. Many improvements have been made based on Faster R-CNN such as RFCN [18], Mask RCNN [19], Light head RCNN [21], Feature pyramid Network [22], etc. Mask RCNN network combines the Faster R-CNN and Fully Convolutional Network (FCN) in one architecture with an additional binary mask to show pixels of the object in the bounding box. There are also many 3D object detection networks, such as Mono3D [23] (see section IV for details on 3D object detection).

On the other hand, one-stage object detection networks directly learn the class probabilities and bounding box coordinates in a single pass through the network without generating region proposals for each image. The one-stage object detection general architecture is shown in Fig. 3. Redmon *et al.* developed You Only Look Once (YOLO) [24], which is the first one-stage DL object detector. The network uses a single neural network to divide the image into regions and simultaneously predict the bounding boxes and class probabilities for each region, as shown in Fig. 4. YOLO is fast compared to the two-stage object detection networks, but its accuracy is lower because of the class imbalance problem, a common problem for one-stage networks. YOLO struggles with small objects and groups

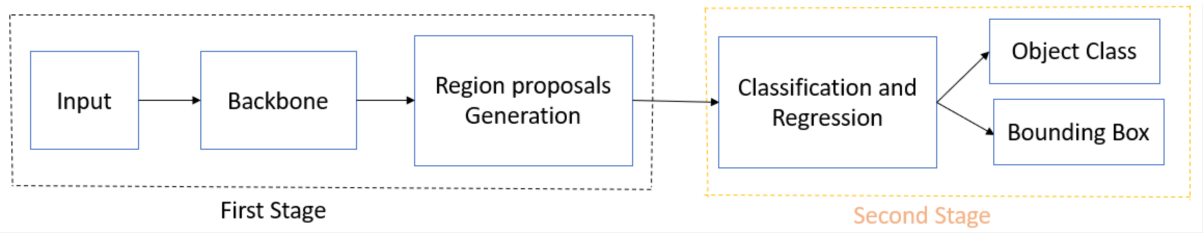

Fig. 2. Two-stage object detection architectural representation. The first stage generates the region of interest (ROI), and then the second stage predicts class probabilities and the bounding box for each object. The backbone network and RPN can be designed as one network.

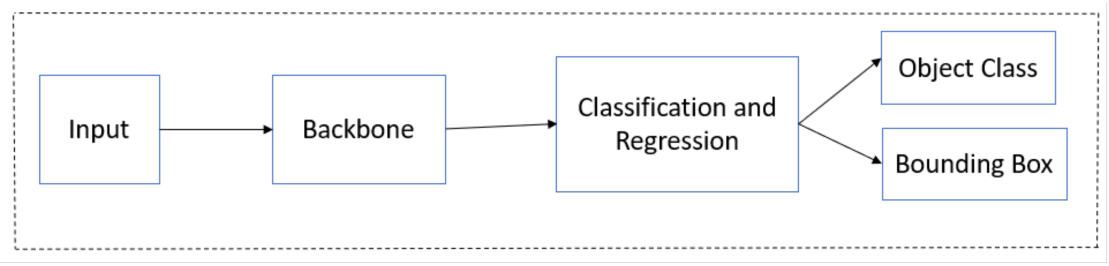

Fig. 3. One-stage object detection model architectural representation. The model learns the class probabilities and bounding box regression in a single pass through the network instead of two passes like the two-stage model.

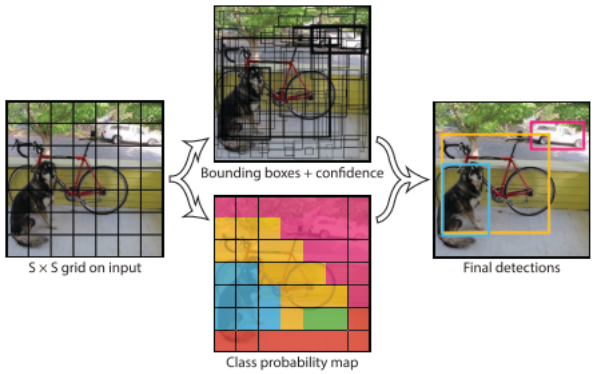

Fig. 4. The YOLO Model [24]. The model divides the image into an S×S grid. For each grid cell, the model predicts bounding boxes, a confidence score for those boxes, and class probabilities.

of object detection. YOLO version 2 [25] improves YOLO by adding batch normalization on convolutional layers, increasing the resolution of images from 224 × 224 to 448 × 448, using anchor boxes instead of fully connected layers to predict bounding boxes adopting multiscale training, and others. The next versions of YOLO [26] and [27] further improve detection speed and solve the accuracy bottlenecks. Similarly, Liu *et al.* put forth a Single Shot MultiBox Detector (SSD) [28], which is a one-stage detection network that improves the YOLO [24] accuracy bottlenecks and a small object detection problem by introducing aspect ratios and a multiscale feature map to detect objects at multiple scales. Then, Lin *et al.* [29] introduced RetinaNet to improve one-stage object detection by introducing focal loss (see the details from the paper [29]) as a classification loss function. The network's accuracy is comparable to the two-stage object detection while maintaining a high detection speed. Zhao *et al.* proposed M2det [30] a multilevel feature pyramid network that enables the construction of multiscale and multilevel features, which helps to detect objects of different scales. Zhang *et al.* introduced a RefineDet [31] to further increase the accuracy of one-stage object detection. MoVi-3D [32], [33], and AutoShape [34] are image 3D one-stage object detection networks (see the details in section IV).

One-stage object detection networks are fast, but their detection accuracy is lower than two-stage detectors due to class imbalance problems. On the other hand, two-stage detectors are slower than one-stage detectors; however, they have better detection accuracy. The RPN reduces redundant detections of two-stage detectors. However, one-stage detectors directly detect class probabilities and bounding box estimation in a single pass without RPN, so the redundancy reduces the detection accuracy.

B. 3D bounding box Encoding

One can estimate the 3D bounding box from the 2D bounding box using perspective projection. There are four commonly used 3D bounding box encoding techniques: the 8-corners method [35], 4-corner-2-height method [36], axis aligned 3D center offset method [37], and seven parameters method [38], [39] as shown in Fig. 5. Mousavian *et al.* [37] proposed an axis aligned 3D center offset 3D bounding box encoding technique that combines DL with geometric constraints. The 3D bounding box described by its center T = $[\Delta x, \Delta y, \Delta z]^T$, dimensions D = $[\Delta h, \Delta w, \Delta l]$, and orientation R $(\Delta\theta, \Delta\phi, \Delta\alpha)$, where $\Delta\theta, \Delta\phi, \Delta\alpha$, $\Delta h, \Delta w, \Delta l$ represents the azimuth angle, elevation angle, roll angle, height, width, and length of the box, respectively. The elevation and roll angles are considered zero. Therefore, we can represent the 3D bounding box as $[\Delta x, \Delta y, \Delta z, \Delta h, \Delta w, \Delta l, \Delta\theta]$. The eight corner box encoding method [35] regresses the oriented 3D boxes from eight corners of 3D proposals $(\Delta x0, \cdots, \Delta x7, \Delta y0, \cdots, \Delta y7, \Delta z0, \cdots, \Delta z7)$, which is a

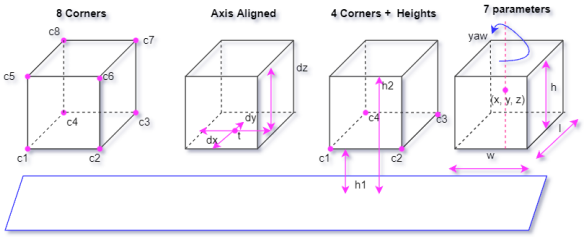

Fig. 5. A diagrammatic comparison between the 8 corner box encoding method [35], 4 Corners and 2 height encoding method [36], the axis aligned box encoding method [37], and seven parameters encoding method [38], [39].

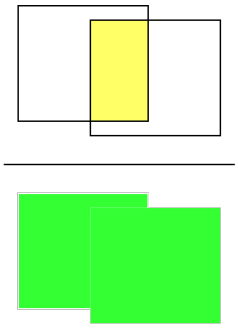

Fig. 6. Pictorial representation of IOU, best viewed in color. Top is intersection and bottom is union.

24-D vector representation. Then, Ku *et al.* [36] developed four corners and two heights, which represent the top and bottom corner offsets from the ground plane. The two heights are determined from the sensor height. Therefore, the 3D bounding box is represented as $(\Delta x1 \ ...\Delta x4, \Delta y1...\Delta y4, \Delta h1, \Delta h2)$. Although the eight corner encoding method gives better results than the axis-aligned method, it does not consider the physical constraints of a 3D bounding box [35]. Because of this, it forces the top corner of the bounding box to align with the bottom corners. The four-corners and two-heights encoding technique solves this problem by adding corner and height offset from the ground plane between the proposed bounding boxes and the ground truth boxes. Moreover, voxelnet [38] and SECOND [39] adopted the seven-point 3D bounding box encoding technique. The seven points are (*x, y, z, w, l, h, θ*), where *x, y*, and *z* are the center coordinates; *w, l*, and *h* are the width, length, and height, respectively. θ is the yaw rotation around the *z*-axis. The elevation and roll angles are considered zero. This encoding method is further adopted by pointpillars [40] and monocular 3d [23]. This technique is widely used in 3D object detection. The regression operation between ground truth and anchors using the seven-point technique can be defined as:

$$\Delta x = \frac{x^{gt}-x^a}{d^a}, \ \Delta y = \frac{y^{gt}-y^a}{d^a}, \Delta z = \frac{z^{gt}-z^a}{d^a}$$
$$\Delta w = \log\frac{w^{gt}}{w^a}, \ \Delta h = \log\frac{h^{gt}}{h^a}, \ \Delta l = \log\frac{l^{gt}}{l^a}$$
$$\Delta\theta = \sin(w^{gt} - w^a), \text{ where the superscripts}$$

gt and *a* represent the ground truth and the anchor boxes, respectively. $d^a = \sqrt{(w^a)^2 + (l^a)^2}$ is the diagonal of the anchor box.

C. Evaluation Metrics for Object detection

One commonly used evaluation metric for object detection is average precision (AP) [41], which is an average detection precision under different recalls for each object category. The mean average precision (mAP) is used as a final evaluation metric for performance comparison of overall object categories. The intersection over union (IOU) threshold value, a geometric overlap between the prediction and the ground truth bounding boxes, is used to measure the object localization accuracy. The graphical representation of IOU is shown in Fig. 6 (the yellow region represents the intersection of the predicted box and the ground truth bounding box, whereas the green region represents the union of the two). Equation (1) shows the mathematical expression of IOU. The representative threshold value may vary from object to object. For example, in the KITTI [42] dataset, a car's 3D bounding box requires an IOU of 0.7, and pedestrians and cyclists require an IOU of 0.5.

$$IOU = \frac{bbox_{pred} \cap bbox_{gt}}{bbox_{pred} \cup bbox_{gt}}, \tag{1}$$

where $bbox_{pred}$ is the predicted bounding box and $bbox_{gt}$ is the ground truth bounding box. Additionally, the F1 score and the Precision-Recall curve are used as evaluation metrics for classification. Precision shows the ratio of the true positives to the total dataset actual values, whereas the recall reveals the ratio of the true positives to the predicted values. The balance of the precision-recall is important for average precision (AP) and mAP. AP is the mean precision of 11 equally spaced recall levels [43] for the KITTI dataset.

$$AP = \frac{1}{11} \sum_{r\in\{0,0.1,...,1\}} P_{interp}(r), \tag{2}$$

$$p_{interp}(r) = maxP(\widetilde{r}), \widetilde{r} : \widetilde{r} \geq r, \tag{3}$$

where $p_{interp}(r)$ is the measured precision at recall $\widetilde{r}$. The mAP is calculated for the overall performance evaluation for 11 recall points. Some works, such as Monopair [44] used 40 recall points instead of eleven to calculate the mAP. The other common performance evaluating metrics are AP3D metric, Average Orientation Similarity (AOS) metrics [41], and the localization metrics (AP_{BV}) [35] for bird's-eye view representation. AOS measures the 3D orientation and detection performance by weighting the cosine similarity between the estimated and ground-truth orientations.

$$AOS = \frac{1}{11} \sum_{r\in\{0,0.1,...,1\}} s(\widetilde{r}), \widetilde{r} : \widetilde{r} \geq r, \tag{4}$$

where r = $\frac{TP}{TP+FN}$ is the recall based on PASCAL [43] dataset. TP is true positive and FN is false negative. The orientation similarity $\in [0, 1]$ at recall *r* is normalized by the cosine similarity.

$$s(r) = \frac{1}{|D(r)|} \sum_{i\in D(r)} \frac{1+\cos\Delta_\theta(i)}{2}\delta_i, \tag{5}$$

where D(r) denotes the set of all object detections at recall rate r, and $\Delta_\theta^{(i)}$ is the difference in angle between estimated and

ground truth orientation of detection i and $\delta(i)$ term penalizes multiple detections.

On the other hand, the nuScenes [45] AP method define a match by thresholding the 2D center distance d on the ground plane rather than IOU. This helps to decouple the effect of object size and orientation for detection.

$$mAP = \frac{1}{|C||D|} \sum_{c \in C} \sum_{d \in D} AP_{c,d}, \tag{6}$$

where D = {0.5, 1, 2, 4} meters, and C is the set of classes. For nuScenes dataset, they measure a set of true positives (TP) for each prediction matched with the ground truth box. Then, for each TP, the mean TP (mTP) is computed for over all classes.

$$mTP = \frac{1}{|C|} \sum_{c \in C} TP_c, \tag{7}$$

Finally, the nuScenes detection score (NDS) is computed.

$$NDS = \frac{1}{10}\left[5mAP + \sum_{mTP \in TP} (1 - min(1, mTP))\right], \tag{8}$$

The nuScenes detection score is an evaluation metric for nuScenes dataset. Most of the datasets in autonomy follow either the KITTI or nuScenes evaluation metric.

IV. Image 3D object Detection Methods and Comparison of Various Methods

Image-based object detection methods use images as input to do the detection task. In this section, we review the monocular image and stereo image-based methods. The 2D object detection is successfully implemented for many applications, but it is not enough for autonomous driving applications. The autonomous vehicle must clearly understand the driving environment for reliable driving. Because of the lack of accurate depth information, 3D object detection is more challenging for image-based methods. Different methods have been proposed to estimate depth from 2D images so that we can detect objects in 3D using the estimated depth. Some of these methods use two-stage object detection methods by first generating object proposals and performing regression for 3D bounding box detection and classification. The classic object detection methods use hand-crafted methods to generate 2D box proposals [47]–[50]. Others use the ability of deep neural networks to learn complex features from images to generate 2D box proposals [51], [52]. Similarly, the box proposals can be generated from geometric constraints [53], [54], pseudo-LiDAR [46], [55] or stereo depth estimation [3], [56]. We categorize image 3D object detection methods based on depth estimation technique into three: Pseudo-LiDAR-based methods, methods that generate depth information from Stereo-images, and methods for 3D proposal generation using geometric constraints.

A. Pseudo-LiDAR based Methods

Some works convert monocular or stereo images into a LiDAR representation called Pseudo-LiDAR to solve the lack of depth information, such as [6], [23], [46], [55], [57], [58]. Pseudo-LiDAR is a LiDAR representation of images by predicting the depth of each image pixel, called the depth map. Wang *et al.* [46] showed that the representation of the data plays a big role rather than the quality of the data on 3D object detection by converting monocular images into LiDAR representation (Pseudo-LiDAR). The stereo depth estimation was done by using pyramid stereo matching network (PSMNet) [59], DISP-NET [60], and SPS-STEREO [61], but they use DORN [62] as a monocular depth estimator. Then, the depth map is projected into a 3D point cloud to produce pseudo-LiDAR by mimicking the LiDAR signal as shown in Fig. 7. The LiDAR-based detectors can directly process the pseudo-LiDAR data. They experimented the pseudo-LiDAR representation with AVOD [63] and Frustum PointNet [64] LiDAR-based models. The result on the KITTI [42] dataset showed that pseudo-LiDAR representation is adequate for 3D object detection compared with image-only implementations. Similarly, Ma *et al.* [55] convert RGB images into pseudo-LiDAR and use pointNet as a backbone network to get objects' 3D locations, dimensions, and orientations for each region of interest (ROI). They also proposed a multimodal features fusion module to fuse the complementary RGB image cues and the generated point clouds to improve the performance. Although converting images to Pseudo-LIDAR takes extra processing, pseudo-LiDAR methods significantly improve performance over image-only methods.

Xu and Chen [65] developed a fusion-based model for 3D object detection by estimating the object class, 2D location, orientation, dimension, and 3D location based on a single monocular image. They use the MultiBin [53] architecture to obtain the pose of a 3D object and then compute the point cloud representation. The estimated depth is encoded as a front view feature and fused with the RGB image to improve the input. Finally, they combine features extracted from the original input and the point cloud to increase object detection performance. Weng and Kitani [23] proposed a two-stage detection network based on pseudo-LiDAR representation by using DORN [62] as a monocular depth estimator. They used instance mask 2D proposals rather than bounding boxes to reduce the number of points not belonging to the object in the point cloud. They train the network with an extended two-stage 3D LiDAR detection algorithm Frustum PointNets [64]. The 2D-3D bounding box consistency constraint was proposed to reduce noise in the Pseudo-LiDAR representation and handle local misalignment. The noise instance mask 2D proposal representation and 2D-3D bounding box consistency constraint improve the performance over [46] and [65] by 6 % and 21.2 %, respectively. In other work, Pseudo-LiDAR++ [6] is an end-to-end depth learning approach using a stereo depth estimation network rather than disparity estimation. The graph-based depth correction algorithm concatenates the learned dense stereo depth and the sparse LiDAR signal for further depth refinement. The result improves 3D object detection, especially the faraway object detection.

Vianney *et al.* [58] proposed a supervised and unsupervised preprocessing scheme to generate refined pseudo-LiDAR data from depth maps before feeding into a 3D object detection network. Qian *et al.* [57] put forth an end-to-end framework based on a differentiable change of representation (CoR)

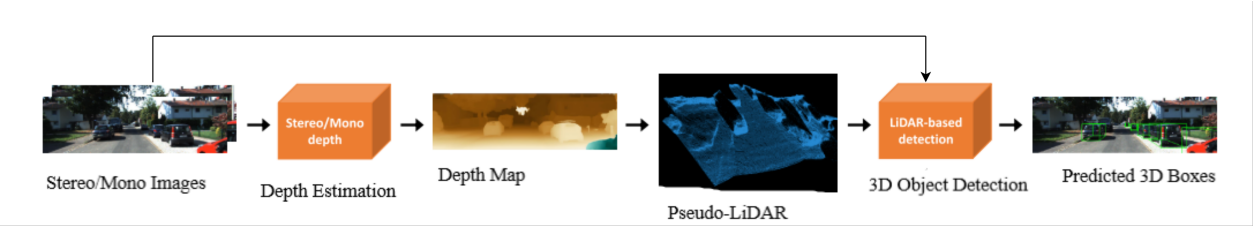

Fig. 7. Generating pseudo-LiDAR representation from given stereo or monocular images by predicting the depth map and projecting into a 3D point cloud coordinate system [46].

network to train the depth estimation and 3D object detection. Zhou *et al.* put forth SGM3D [66], a domain adaptation-based model that leverages the stereo representation to improve the performance of monocular 3D object detection. The authors used a pretrained stereo matching model PSMNet [59] for depth learning. Pixels are converted into 3D pseudo point clouds based on the estimated depth and camera instincts. They proposed a multi-granularity domain adaptation (MG-DA) module to get a consistent intermediate feature representation and the predictions per anchor between the output from the stereo-based and monocular approaches. An IOU matching-based alignment (IOU-MA) module is introduced to reduce the mismatches between stereo and monocular predictions. Experimental results on the KITTI [42] and Lyft [67] datasets show a performance improvement.

Reading *et al.* [68] put forth a categorical depth distribution network (CDDN) for Monocular 3D Object Detection. The frustum feature network projects image information into 3D space and constructs a frustum feature grid. Then, pointpillars [40] detection head performs 3D object detection. The model used the KITTI [42] and Waymo [69] datasets for the experiment. Chen *et al.* [70] proposed the Disp R-CNN 3D object detection model from stereo images, which has has three stages. In the first stage, Mask R-CNN [19] detects images' 2D bounding boxes and instance segmentation. The instance disparity estimation network (iDispNet) estimates an instance disparity map in the second stage. Finally, an instance point cloud is generated from the instance disparity map and is the input to the detector head for 3D bounding box regression. The experimental result on the KITTI [42] dataset shows a promising result.

Converting monocular or stereo images into Pseudo-LIDAR improves 3D object detection over image-only methods; however the performance is lower than LiDAR-based methods because of the error from image to LiDAR conversion. This method facilitates domain adaptation methods for 3D object detection in autonomous driving. Although converting image data into Pseudo-LiDAR representation takes extra processing, it is a good option when LIDAR data is not readily available.

B. Methods that Generate Depth Information from Stereo Images

These methods generate depth from stereo-images [3]–[5], [56], [71]–[74]. Mono3D [56] uses stereo images to estimate the depth and generate 3D bounding box object proposals by encoding object size priors, ground planes, a variety of depth-informed features, point cloud densities, and distance to the ground. The problem is articulated as an energy minimization function, and Markov Random Field (MRF) is used to score 3D bounding boxes for proposal generation. Fast R-CNN [16] is used to predict the class proposal, and objects' orientation is estimated using the top object candidates. Chen *et al.* [75] extended the previous work [56] to generate class-specific 3D object proposals (3DOP) with very high recall for various IOU thresholds by assuming that objects should be on the ground plane and using only a single monocular image. They use semantic and object instance segmentation, context, shape features, and location priors to score 3D bounding boxes. The limitation of 3DOP is that it should run separately for each object class to achieve a high recall. This operation increases the processing time because of many generated object proposals. To overcome this problem, Pham and Jeon [4] introduced a proposal reranking algorithm, DeepStereoOP, to rerank the generated 3D object proposals. This algorithm helps achieve high recall and good localization using only a few candidate proposals. The algorithm is a two-stream CNN that uses RGB features, depth features, disparity maps, and distance to the ground to rerank top-ranked candidates. The result shows the DeepStereoOP algorithm is superior to the Mono3D [75] algorithm to get high recall with fewer proposals. Chen *et al.* [3] presented a proposal generation algorithm using stereo imagery and contextual information. They generate 3D object proposals using an energy minimization function that encodes object size priors, ground plane information, and depth-informed features, such as free space, point cloud densities, and distance to the ground. The CNN scoring network uses appearance, depth, and context information to simultaneously predict 3D object proposals and object poses. The result outperforms the previous works, such as [4] and [75] on the KITTI dataset. Most of these methods formulated the object proposal as an energy minimization problem. Works, such as DeepStereoOP [4] proposed a reranking algorithm to reduce redundant proposals and use only a few proposals. Additionally, contextual information can be used together with stereo images for proposal generation.

Triangulation Learning Network (TLNet) [71] uses 3D anchors to construct object-level geometric correlation between stereo images. Then, the neural network learns the correspondence between stereo images to triangulate the target object near the anchor. The Channel reweighting method is also proposed to enhance informative features and weaken the noisy signals by measuring left-right coherence, which overcomes the high computational burden of generating disparity maps in mono3D [75]. Li *et al.* [76] developed an extended Faster R-CNN [16]

based 3D object detection method, Stereo R-CNN, to detect and associate objects in left and right images simultaneously by using the sparse, dense, semantic, and geometry information in stereo imagery. After generating left and right region of interests (ROIs) proposals, they concatenate the left-right ROI features of object classes and regress 2D stereo boxes, viewpoint, and 3D dimensions. They predicted a key point using only left features combined with 2D stereo boxes for 3D box estimation. Peng *et al.* [73] put forth an Instance-Depth-Aware module as a depth estimation method of a 3D bounding box's center using instance-depth awareness, disparity adaptation, and matching cost reweighting. The channel and cost reweighting methods are essential to enhance features and weaken noisy signals using left-right coherence.

DSGN [74] is a one-stage end-to-end stereo-based 3D object detection model that jointly estimates the depth and detects 3D objects. Stereo CenterNet [72] uses semantic and geometric information in stereo images to implement 3D object detection. They use the anchor-free 2D box association method by detecting only the objects in the left images and computing the left-right associations by predicting the distance between them. Königshof *et al.* [77] put forth 3D object detection method using stereo image and semantic information. The semantic map and optional bounding box suggestions are generated from the left image using ResNet-38 [78]. The model was trained and tested on the KITTI [42] dataset. Li and Chen proposed S-3D-RCNN [79], a two-stage joint stereo 3D object detection and shape estimation model from a pair of stereo RGB images. The authors presented a global-local framework to decouple object pose estimation from object shape. The model showed a significant performance improvement on the KITTI [42] dataset.

Liu *et al.* proposed YOLOStereo3D [80] 3D object detection model using stereo camera images. The authors described each anchor by 12 regressed parameters as $[x_{2d},\ y_{2d},\ w_{2d},\ h_{2d}]$ for the 2D bounding boxes; $[c_x,\ c_y,\ z]$ for the 3D centers of objects on the left image; $[w_{3d},\ h_{3d},\ l_{3d}]$ corresponds to the width, height, and length of the 3D bounding boxes, respectively. They concurrently applied photometric distortion augmentation [54] on binocular images and random flipping [76] during training. After extracting multi-scale features from binocular images, the features passed through a multiscale stereo matching and fusion module. Gao *et al.* [81] put forth an Efficient Geometry Feature Network (EGFN) for 3D object detection. The authors used ResNet-34 [82] to extract multi-scale feature maps. The proposed 3D geometry feature representation (EGFR) module generates multiscale 3D geometry features in 3D space with no 3D convolution. The experimental result on the KITTI dataset shows the EFGN model outperforms the YOLOStsereo3D [80] model. The YOLOStsereo3D model used photometric distortion augmentation and random flipping concurrently to generate multiscale features; however, EGFN used the EGFR module to generate multiscale features. Zhang *et al.* [83] extended CenterNet [72] as a flexible framework for monocular 3D object detection that explicitly decouples the truncated objects. The authors formulated the object depth estimation as an uncertainty-guided ensemble of multiple approaches and combined adaptively different key points to estimate the depth. The experimental results on the KITTI dataset [42] show the model outperforms SOTA models by the time, such as RTM3D [84] and MoVi3D [32]. Chen *et al.* [85] proposed Pseudo-Stereo 3D detection method for 3D object detection. The virtual view is generated from every single image to use as a stereo image with the input image. Three virtual view generation methods are proposed: image-level generation, feature-level generation, and feature-clone for detecting 3D objects from a single image. A disparity-wise dynamic convolution is proposed to filter the features adaptively from a single image for generating virtual image features. The model is trained and tested on the KITTI [42] dataset.

Stereo-image-based methods use 2D left and right boxes to predict the bounding boxes of objects in the 3D space. Photometric alignment is usually used to optimize the 3D bounding box position further. The object-level geometric correlation between left and right images can be constructed using different techniques, such as 3D anchors. The energy minimization function is also vital for generating 3D object proposals. Some of the stereo-image-based methods use stereo matching and stereo instance segmentation to match detection between left and right images on ROIs and estimate instance-level disparity only for regions that contain objects of interest. The following methods use either stereo matching or stereo instance segmentation to match detection or estimate the disparity of ROI.

ZoomNet [5] applied adaptive zooming to resize bounding boxes and adjust intrinsic camera parameters simultaneously to realize instance-level disparity estimation and construct point-cloud and pseudo-LiDAR from each object instance rather than the full image. The Pseudo-LiDAR-based object detection has poor performance on distant objects because a distant object has low resolution because of the small number of points, the difficulty in distinguishing the relative positions between stereo images, and occlusions. This adaptive zooming helps analyze the distant objects at larger resolutions, estimate better disparity, and have more uniform density point clouds. They also present pixel-wise part locations to help solve the occlusion detection problem. Similarly, Pon *et al.* [86] proposed an object-centric stereo matching network (OC Stereo), which solves the problems related to deep stereo matching methods. They developed an object-centric depth representation to help solve streaking artifacts, the ambiguity between the object or the background pixels, and the pixel imbalance problem between near and far objects. The authors presented a fast 2D box association algorithm to accurately match detection between left and right images by stereo matching on Regions of Interest (ROIs) and considering only pixels belonging to objects. Recently, Disp R-CNN [87] proposed an instance-level disparity estimation network (iDispNet) that estimates disparity only for regions that contain objects of interest rather than the entire image and learns a category-specific shape prior. This operation helps capture the smooth shape and sharp edges of object boundaries for more accurate 3D object detection.

The lack of depth of image-based methods can be partially solved using stereo images. The 3D object proposals

are generated from stereo images using different techniques. Some methods, such as TLNET [71] use cost and channel reweighting to enhance features and weaken noises. Others, for example, DeepStreoOP [4] use proposal reranking algorithms to reduce the redundant proposal generation.

C. Methods for 3D proposal generation using geometric constraints

These works create 3D proposals by adding additional geometric constraints including object shape, ground planes, and key points [32], [53], [54], [63], [75], [84], [88]–[96]. Mousavian *et al.* proposed Deep3DBox [53], a 3D object detection method by incorporating geometry constraints. A hybrid discrete-continuous loss is used to estimate the 3D object orientation and then apply regression on the 2D bounding box combined with the estimated geometric constraints to produce the object 3D bounding box. M3D-RPN [54] put forth a single end-to-end region proposal network for 3D object detection by using the correlation between 2D scale and 3D depth. The proposed depth-aware convolutional layer is used to improve the 3D parameter estimation that enhances 3D scene understanding. Mono3d++ [88] uses a joint method of predicting the vehicles's shape and pose using a 3D bounding box and morphable wireframe model from a single RGB image. The unsupervised monocular depth, a ground plane constraint, and vehicle shape priors optimize loss functions. The overall energy function integrates the loss, the vehicles'shape, and pose to improve vehicles' detection further. Integrating the loss function with the shape of vehicles may limit the model's performance because of the shape difference between vehicles.

GS3D [89] proposed an efficient approach to get a coarse cuboid for each predicted 2D box to determine the 3D bounding box by refinement. This method improves the 3D object detection and performs better than regression-based bounding box prediction. MonoGRNet [90] is a unified network for 3D object detection from monocular RGB images using geometric reasoning and instance-level depth estimation. ROI-10D [91] developed an end-to-end network for 3D object detection by lifting 2D into 3D to predict six degrees of freedom pose information (rotation and translation). The loss function measures the metrics misalignment of boxes and minimizes the error by comparing it with the ground truth 3D boxes. Barabanau *et al.* [92] also developed a combination of a key-point-based and geometric reasoning approach for 3D object detection from monocular images. Likewise, Cai *et al.* [93] modeled the 3D object detection task as a combination of a structured polygon prediction task and a depth estimation task. The depth estimation network uses an object's height to estimate the depth and then combines it with the structured polygon to obtain the 3D boxes. Finally, the fine-grained 3D box refinement is proposed in BEV to improve the accuracy of the 3D bounding box.

Similarly, Ku *et al.* [94] estimated the region proposal network through geometric constraints and applied regression further for 3D object detection. SMOKE [95] combined a single key-point estimate with regressed 3D variables to predict a 3D bounding box of individually detected objects rather than generating 2D region proposals. Roddick *et al.* [96] proposed a 3D object detection module by mapping image-based features into an orthographic 3D space. An orthographic feature transforms the RGB image into an orthographic bird's-eye-view feature map. RTM3D [84] predicted the nine-perspective key-points of a 3D bounding box and modeled the geometric relationship of 3D and 2D points to detect 3D objects from monocular images. MoVi-3D [32] is a one-stage deep architecture that leverages geometrical information to generate virtual views, using prior geometrical knowledge to control the scale variability of the object because of depth.

Ding *et al.* proposed a Depth-guided Dynamic Depthwise Dilated local convolution (D4LCN) [97] network where local filters learn specific geometry from each RGB image using a depth map that is applied locally to each pixel and channel of each image. Jorgensen *et al.* [33] proposed a one-stage detector network that comprises nonmaximal suppression and nonlinear least squares optimizer to generate perobject canonical 3D bounding box parameters. This method avoids processing the image multiple times and reduces the computational bottleneck of deep neural networks. Srivastava *et al.* [98] developed a 2D to 3D lifting method for autonomous vehicle's 3D object detection. They generate BEV images from a single RGB image using Generative Adversarial Networks (GAN) for image-to-image translation [99] and then do 3D object detection using the generated BEV images.

Garanderie *et al.* [100] proposed a 3D object detection model for autonomous vehicles by using 360 panoramic imagery. This method is important to avoid blind spots in driving. They tested their network using the CARLA [101] urban driving simulator and KITTI [42] dataset. Liu *et al.* [102] developed a deep fitting scoring network for monocular 3D object detection. The network generates 3D proposals using the object's anchor-based dimension and the orientation regression. Then, they use a fitting quality network (FQNet) to understand the spatial relationship between 3D proposals and objects only using 2D images. Chen *et al.* [44] proposed a pair-wise spatial relationship-based 3D object detection method. The object location is computed using uncertainty-aware predictions and 3D distances for the adjacent object pairs. Finally, nonlinear least squares jointly optimize the system. Recently, Bao *et al.* proposed MonoFENet [103] network for 3D object detection by estimating the disparity from a monocular image. The estimated disparity is transformed into a 3D dense point cloud to feed into a point feature enhancement (PointFE) network and fuse with the image features for the final 3D bounding box regression.

Bao *et al.* [104] proposed a two-stage object-aware 3D object detection model that uses both the region-wise appearance attention and the geometric projection distribution to vote the 3D centroid proposals for 3D object localization. The 2D region proposals are generated using RPN from Faster R-CNN [17], then 3D centroid proposals are estimated from generated ROIs grid coordinates. Based on the proposed object-aware voting module, which comprises region-wise appearance attention and the geometric projection distribution, the 3D centroid proposals are voted for 3D localization. Finally, 3D bounding boxes of objects are detected based on the proposed ROIs without learning the dense depth. Zhou *et al.* put forth IAFA

TABLE I

BEV AND 3D PERFORMANCE COMPARISON OF IMAGE-BASED 3D OBJECT DETECTION METHODS ON THE KITTI [42] VALIDATION BENCHMARK. R40 MEANS THE MAP IS CALCULATED FOR 40 RECALL POINTS INSTEAD OF 11 POINTS. E STANDS FOR EASY, M FOR MODERATE, AND H FOR HARD.

Methods	$AP_{BEV}(IOU=0.7)$									$AP_{3D}(IOU=0.7)$								
	car			pedestrians			cyclists			car			pedestrians			cyclists		
	E	M	H	E	M	H	E	M	H	E	M	H	E	M	H	E	M	H
3DOP [3]	12.63	9.49	7.59	-	-	-	-	-	-	6.55	5.07	4.10	-	-	-	-	-	-
Monopair [44] (r40)	24.12	18.17	15.76	-	-	-	-	-	-	16.28	12.30	10.42	-	-	-	-	-	-
TLNET [71]	29.22	21.88	18.83	-	-	-	-	-	-	18.15	14.26	13.72	-	-	-	-	-	-
Sun *et al.* [55]	43.75	28.39	23.87	-	-	-	-	-	-	32.23	21.09	17.26	-	-	-	-	-	-
Stereo RCNN [76]	68.50	48.30	41.47	-	-	-	-	-	-	54.11	36.69	31.07	-	-	-	-	-	-
IDA-3D [73]	70.68	50.21	42.93	-	-	-	-	-	-	54.97	37.45	32.23	-	-	-	-	-	-
Pseudo-LiDAR [46]	74.90	56.80	49.00	-	-	-	-	-	-	61.90	45.30	39.00	-	-	-	-	-	-
Disp R-CNN [87]	76.51	58.63	50.26	-	-	-	-	-	-	63.57	47.15	39.73	-	-	-	-	-	-
Sun *et al.* [86]	77.66	65.95	51.20	44.00	37.20	30.39	48.20	27.90	26.96	64.07	48.34	40.39	34.80	29.05	28.06	45.59	25.93	24.62
ZoomNet [5]	78.68	66.19	57.60	-	-	-	-	-	-	62.96	50.47	43.63	-	-	-	-	-	-

[105], an instance-aware feature aggregation model for 3D object detection from a single image. The model collects pixels that belong to the same object for contributing to the center classification and generates an attention map to aggregate useful information for each object. The authors used the coarse instance annotations from other networks as a supervision signal to generate the features aggregation attention maps. The model was trained with KITTI [42] dataset.

Lu *et al.* put forth Geometry Uncertainty Projection Network (GUP Net) [106] for monocular 3D object detection. The input images are processed by the 2D detection backbone, built on CenterNet [72], to get 2D bounding boxes (ROIs) and 3D bounding box information, i.e., angle, dimensions, and 3D projected center for each box. Then, GPU Net predicts the depth information and its corresponding uncertainty by combining mathematical priors and uncertainty modeling. An efficient Hierarchical Task Learning (HTL) strategy is proposed to reduce the instability caused by task dependency in geometry-based methods (error amplification). The error amplification causes amplification of the estimated depth. The HTL strategy controls the overall training process by making each task idle until its pre-tasks are well trained. The experimental result on the KITTI dataset [42] outperforms methods such as MoVi-3D [32] and RAR-net [107].

Wang *et al.* [108] proposed a graph-based depth-conditioned dynamic message propagation (DDMP) model for monocular 3D object detection. The model comprises two branches: the regression branch and the depth extraction branch. The regression branch receives the RGB images for feature extraction, and the depth extraction branch estimates the corresponding depth maps and extracts depth-aware features. The center-aware depth encoding (CDE) method is proposed to reduce the inaccurate depth prior issues. The context-aware and depth-aware features are integrated with a graph message propagation pattern via the DDMP module. Finally, 3D object boxes were achieved using a 3D detection head. The experimental result on the KITTI dataset [42] shows that the model outperforms the previous models, such as D4LCN [97]. Liu *et al.* presented AutoShape [34], a one-stage real-time shape-aware monocular 3D object detection model. The model employs geometry constraints for 3D keypoints and their 2D projections on images to enhance the detection performance. The proposed automatic annotation pipeline can autogenerate the shape-aware 2D/3D keypoints correspondences for each object. The model was evaluated with KITTI [42] car dataset.

Some works followed different approaches than we mentioned above to solve the 3D objection problem from the input of 2D images. Liu *et al.* put forth RAR-Net [107] a reinforced axial refinement network monocular 3D object detection model. The proposed model starts with an initial prediction, refines it gradually towards the ground truth, and only one 3D parameter is changed in each step. The ϵ- greedy policy, which maximizes the reward by selecting the action with the highest estimated reward, is implemented to get a reward after each action is taken and the refined 3D box of the monocular 3D detection network. At each step, information from the image and 3D space is fused and then projected the current detection into the image space to preserve information. This reinforcement learning-based learning can be used as a post-processing stage and integrated into an existing monocular 3D detection model to improve performance with some extra computational cost. The model trained with KITTI dataset [42] and showed promising performance. Mehtab *et al.* [109] proposed a 3D vehicle detection model using LiDAR and camera sensors. The autonomous vehicle's size and orientation of 3D bounding boxes are estimated from the RGB images, whereas the LiDAR point cloud is used for distance estimation. The authors used MobileNetV2 [110] as an image feature extractor. The model was trained and tested on the KITTI [42], and Waymo [69] datasets. Simonelli *et al.* [111] put forth self-supervised loss disentangling transformation for monocular 3D object detection. The loss separates the groups of parameter contributions into separate terms as the original loss. The authors also applied the loss function IOU for 2D detection and 3D bounding box predictions and detection confidence. The model was trained on the KITTI [42] dataset.

Image-based 3D object detection is more challenging due to the lack of depth information. Most depth estimation techniques can be categorized into Pseudo-LiDAR based, stereo-image based, or using geometric constraints, such as the object's shape and key points to estimate the depth. The Pseudo-LiDAR methods generate point cloud data from images and use 3D LiDAR-based methods for detection. Although these methods outperform image-only methods, their

accuracy is still lower than LiDAR-based methods because of the image-to-LiDAR generation error. The stereo image-based methods use the left and right image disparity to estimate the depth estimation. These methods also improve the 3D object detection performance than the single image methods. Some works also generate stereo images from a single image by generating a virtual image, which outperforms single-image methods. Other works use geometric constraints to estimate the depth information of a single image. Table I shows the BEV and 3D performance comparison of the image-based 3D object detection methods on KITTI [42] test data benchmarks.

V. Challenges and Future Directions

Camera images, especially monocular images, are rich in texture and color information, which are essential for color-related tasks, such as object classification and lane detection. However, they do not provide high accuracy depth information for a complete understanding of the surrounding environment. Autonomous driving needs to be robust to drive in different weather conditions, but cameras are affected by bad weather. Additionally, DL models evaluated on a different domain than trained perform poorly. We presented some challenges and future research direction in image-based 3D object detection for AVs.

1) **Semisupervised Learning:** One of the challenges of supervised learning is annotating and labeling data, which requires time and money. Data annotation and labeling problems can be solved using unsupervised learning. However, unsupervised models' detection and classification accuracy are lower than the supervised models. The potential solution to these problems is applying a semisupervised model using few labeled data and many unlabeled data to leverage the abundance of freely available images for different applications. Some teacher-student models, such as Zhang *et al.* [112] belong to a semi-supervised 3D object detection network for autonomous driving. The teacher model generates pseudo-labels in the teacher-student model, and the student model trains the pseudo-labels and the labeled dataset. Then, the teacher model may receive an update from the student model for better pseudo-label prediction. This model is mainly used in 2D object detection, but the 3D equivalents are limited.
2) **Multitask Learning:** The feature extractor part of DL networks can be common to multiple applications. Therefore, building a model with common feature extractor /lower architecture of the model with multiple decision layers to perform multiple tasks can save time, memory, and computational power. For example, [113] performs object detection, and segmentation multitask learning. We expect many multitask learning works for AVs.
3) **Domain Adaptive Models:** DL models should perform the same/equivalent when tested with a different domain than they were trained. However, most DL models poorly perform when the training domain changes. Domain adaptive models are essential for autonomous driving to avoid country-specific changes, such as traffic sign variability and corner issues. Therefore, we need domain adaptive models to learn the driving environment changes and respond quickly to the changes.
4) **Lightweight Models:** DL models in AVs should fulfill the following three criteria [1]:
   1) **Accurate** to precise information about the surrounding environments.
   2) **Robust** to work in different weather.
   3) **Real-time** to perform high-speed driving. To achieve the above criteria, DL models should be robust enough to work under different weather and lightweight to be deployed in low-power and low-memory embedded hardware devices. Most of the existing 3D object detection models are not lightweight as of 2D equivalents. There are relatively lightweight 2D object detection models, such as YOLO [114] and SSD [28] than 3D object detection models.
5) **Multisensor Fusion:** Cameras are suitable for color-related detection and reach in texture too. Although different methods have been developed to solve the lack of 3D information, 3D object detection using cameras is challenging. Additionally, cameras are not robust to adverse weather, which makes robust driving in different environmental weather challenging. Other sensors can provide better 3D information, such as LiDAR, and more robust to adverse weather, such as radar. Therefore, fusing the camera images with LiDAR and/or radar can improve 3D object detection by using the best out of different sensors.

VI. Conclusions

This survey presented DL-based monocular and stereo camera images for 3D object detection for autonomous driving. The 3D bounding box encoding methods and the corresponding evaluation metrics were summarized. The general object detection categories as one-stage and two-stage and depth estimation methods of 3D object detection are also reviewed. The depth estimation methods are grouped based on techniques, such as pseudo-LiDAR, stereo image, and geometric constraint methods. Although 3D object detection using camera images has shown significant performance improvement due to the rapid growth of DL, there are still issues to be solved for reliable and robust driving, such as driving in bad weather or at night. The camera sensor is rich in color and texture and is also inexpensive, but it cannot measure the distance from long range, cannot withstand bad weather, and does not give direct 3D information. 3D sensors, such as LiDAR and radar, provide 3D information about the driving environment and objects. LiDAR is more robust than a camera for inclement weather and a good choice for long-distance measurement and velocity estimation. However, it is not rich in color and texture. Similarly, radar is a robust sensor for inclement weather and the best choice for distance measurement and velocity estimation, but it has low resolution, making radar-based detection difficult. Additionally, there is a possibility of sensor failure during autonomous driving. Thus, using

multiple sensors for autonomous driving is essential to use redundant data from different sensors for reliable and robust driving to work under bad weather or sensor failure conditions. Lightweight and accurate 3D object detection models are necessary to improve the speed and accuracy of real-time processing. Finally, challenges and possible research directions were presented.

REFERENCES

[1] D. Feng, C. Haase-Schuetz, L. Rosenbaum, H. Hertlein, C. Glaeser, F. Timm, W. Wiesbeck, and K. Dietmayer, "Deep multi-modal object detection and semantic segmentation for autonomous driving: Datasets, methods, and challenges," *IEEE Transactions on Intelligent Transportation Systems*, 2020.

[2] E. Arnold, O. Y. Al-Jarrah, M. Dianati, S. Fallah, D. Oxtoby, and A. Mouzakitis, "A survey on 3d object detection methods for autonomous driving applications," *IEEE Transactions on Intelligent Transportation Systems*, vol. 20, no. 10, pp. 3782–3795, 2019.

[3] X. Chen, K. Kundu, Y. Zhu, H. Ma, S. Fidler, and R. Urtasun, "3d object proposals using stereo imagery for accurate object class detection," *IEEE transactions on pattern analysis and machine intelligence*, vol. 40, no. 5, pp. 1259–1272, 2017.

[4] C. C. Pham and J. W. Jeon, "Robust object proposals re-ranking for object detection in autonomous driving using convolutional neural networks," *Signal Processing: Image Communication*, vol. 53, pp. 110–122, 2017.

[5] Z. Xu, W. Zhang, X. Ye, X. Tan, W. Yang, S. Wen, E. Ding, A. Meng, and L. Huang, "Zoomnet: Part-aware adaptive zooming neural network for 3d object detection," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 07, 2020, pp. 12 557–12 564.

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

[7] S.-h. Kim and Y. Hwang, "A survey on deep learning based methods and datasets for monocular 3d object detection," *Electronics*, vol. 10, no. 4, p. 517, 2021.

[8] L. Jiao, F. Zhang, F. Liu, S. Yang, L. Li, Z. Feng, and R. Qu, "A survey of deep learning-based object detection," *IEEE access*, vol. 7, pp. 128 837–128 868, 2019.

[9] M. M. Rahman, Y. Tan, J. Xue, and K. Lu, "Recent advances in 3d object detection in the era of deep neural networks: A survey," *IEEE Transactions on Image Processing*, vol. 29, pp. 2947–2962, 2019.

[10] Y. Li, L. Ma, Z. Zhong, F. Liu, M. A. Chapman, D. Cao, and J. Li, "Deep learning for lidar point clouds in autonomous driving: a review," *IEEE Transactions on Neural Networks and Learning Systems*, 2020.

[11] Y. Guo, H. Wang, Q. Hu, H. Liu, L. Liu, and M. Bennamoun, "Deep learning for 3d point clouds: A survey," *IEEE transactions on pattern analysis and machine intelligence*, 2020.

[12] D. Fernandes, A. Silva, R. Névoa, C. Simões, D. Gonzalez, M. Guevara, P. Novais, J. Monteiro, and P. Melo-Pinto, "Point-cloud based 3d object detection and classification methods for self-driving applications: A survey and taxonomy," *Information Fusion*, vol. 68, pp. 161–191, 2021.

[13] R. Qian, X. Lai, and X. Li, "3d object detection for autonomous driving: A survey," *arXiv preprint arXiv:2106.10823*, 2021.

[14] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014.

[15] K. He, X. Zhang, S. Ren, and J. Sun, "Spatial pyramid pooling in deep convolutional networks for visual recognition," *IEEE transactions on pattern analysis and machine intelligence*, vol. 37, no. 9, pp. 1904–1916, 2015.

[16] R. Girshick, "Fast R-CNN," in *Proceedings of the International Conference on Computer Vision (ICCV)*, 2015.

[17] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *Neural Information Processing Systems (NIPS)*, 2015.

[18] J. Dai, Y. Li, K. He, and J. Sun, "R-fcn: Object detection via region-based fully convolutional networks," *arXiv preprint arXiv:1605.06409*, 2016.

[19] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask R-CNN," in *Proceedings of the International Conference on Computer Vision (ICCV)*, 2017.

[20] J. R. Uijlings, K. E. Van De Sande, T. Gevers, and A. W. Smeulders, "Selective search for object recognition," *International journal of computer vision*, vol. 104, no. 2, pp. 154–171, 2013.

[21] Z. Li, C. Peng, G. Yu, X. Zhang, Y. Deng, and J. Sun, "Light-head r-cnn: In defense of two-stage object detector," *arXiv preprint arXiv:1711.07264*, 2017.

[22] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 2117–2125.

[23] X. Weng and K. Kitani, "Monocular 3d object detection with pseudo-lidar point cloud," in *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, 2019, pp. 0–0.

[24] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 779–788.

[25] J. Redmon and A. Farhadi, "Yolo9000: better, faster, stronger," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 7263–7271.

[26] ——, "Yolov3: An incremental improvement," *arXiv preprint arXiv:1804.02767*, 2018.

[27] A. Bochkovskiy, C.-Y. Wang, and H.-Y. M. Liao, "Yolov4: Optimal speed and accuracy of object detection," *arXiv preprint arXiv:2004.10934*, 2020.

[28] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, "Ssd: Single shot multibox detector," in *European conference on computer vision*. Springer, 2016, pp. 21–37.

[29] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2980–2988.

[30] Q. Zhao, T. Sheng, Y. Wang, Z. Tang, Y. Chen, L. Cai, and H. Ling, "M2det: A single-shot object detector based on multi-level feature pyramid network," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 33, no. 01, 2019, pp. 9259–9266.

[31] S. Zhang, L. Wen, X. Bian, Z. Lei, and S. Z. Li, "Single-shot refinement neural network for object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 4203–4212.

[32] A. Simonelli, S. R. Bulò, L. Porzi, E. Ricci, and P. Kontschieder, "Towards generalization across depth for monocular 3d object detection," *arXiv preprint arXiv:1912.08035*, 2019.

[33] E. Jörgensen, C. Zach, and F. Kahl, "Monocular 3d object detection and box fitting trained end-to-end using intersection-over-union loss," *arXiv preprint arXiv:1906.08070*, 2019.

[34] Z. Liu, D. Zhou, F. Lu, J. Fang, and L. Zhang, "Autoshape: Real-time shape-aware monocular 3d object detection," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 15 641–15 650.

[35] X. Chen, H. Ma, J. Wan, B. Li, and T. Xia, "Multi-view 3d object detection network for autonomous driving," in *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 2017, pp. 1907–1915.

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

[37] S. Song and J. Xiao, "Deep sliding shapes for amodal 3d object detection in rgb-d images," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 808–816.

[38] Y. Zhou and O. Tuzel, "Voxelnet: End-to-end learning for point cloud based 3d object detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 4490–4499.

[39] Y. Yan, Y. Mao, and B. Li, "Second: Sparsely embedded convolutional detection," *Sensors*, vol. 18, no. 10, p. 3337, 2018.

[40] A. H. Lang, S. Vora, H. Caesar, L. Zhou, J. Yang, and O. Beijbom, "Pointpillars: Fast encoders for object detection from point clouds," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 12 697–12 705.

[41] A. Geiger, P. Lenz, and R. Urtasun, "Are we ready for autonomous driving? the kitti vision benchmark suite," in *2012 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2012, pp. 3354–3361.

[42] ——, "Are we ready for autonomous driving? the kitti vision benchmark suite," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012.

[43] M. Everingham, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes (voc) challenge," *International journal of computer vision*, vol. 88, no. 2, pp. 303–338, 2010.

[44] Y. Chen, L. Tai, K. Sun, and M. Li, "Monopair: Monocular 3d object detection using pairwise spatial relationships," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 12 093–12 102.

[45] H. Caesar, V. Bankiti, A. H. Lang, S. Vora, V. E. Liong, Q. Xu, A. Krishnan, Y. Pan, G. Baldan, and O. Beijbom, "nuscenes: A multimodal dataset for autonomous driving," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 11 621–11 631.

[46] Y. Wang, W.-L. Chao, D. Garg, B. Hariharan, M. Campbell, and K. Q. Weinberger, "Pseudo-lidar from visual depth estimation: Bridging the gap in 3d object detection for autonomous driving," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 8445–8453.

[47] C. L. Zitnick and P. Dollár, "Edge boxes: Locating object proposals from edges," in *European conference on computer vision*. Springer, 2014, pp. 391–405.

[48] P. Krahenbuhl and V. Koltun, "Learning to propose objects," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 1574–1582.

[49] T. Lee, S. Fidler, and S. Dickinson, "Learning to combine mid-level cues for object proposal generation," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 1680–1688.

[50] D. Eigen, C. Puhrsch, and R. Fergus, "Depth map prediction from a single image using a multi-scale deep network," *arXiv preprint arXiv:1406.2283*, 2014.

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

[52] Z. Lingtao, F. Jiaojiao, and L. Guizhong, "Object viewpoint classification based 3d bounding box estimation for autonomous vehicles," *arXiv preprint arXiv:1909.01025*, 2019.

[53] A. Mousavian, D. Anguelov, J. Flynn, and J. Kosecka, "3d bounding box estimation using deep learning and geometry," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 7074–7082.

[54] G. Brazil and X. Liu, "M3d-rpn: Monocular 3d region proposal network for object detection," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 9287–9296.

[55] X. Ma, Z. Wang, H. Li, P. Zhang, W. Ouyang, and X. Fan, "Accurate monocular 3d object detection via color-embedded 3d reconstruction for autonomous driving," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 6851–6860.

[56] X. Chen, K. Kundu, Y. Zhu, A. G. Berneshawi, H. Ma, S. Fidler, and R. Urtasun, "3d object proposals for accurate object class detection," in *Advances in Neural Information Processing Systems*. Citeseer, 2015, pp. 424–432.

[57] R. Qian, D. Garg, Y. Wang, Y. You, S. Belongie, B. Hariharan, M. Campbell, K. Q. Weinberger, and W.-L. Chao, "End-to-end pseudo-lidar for image-based 3d object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 5881–5890.

[58] J. M. U. Vianney, S. Aich, and B. Liu, "Refinedmpl: Refined monocular pseudolidar for 3d object detection in autonomous driving," *arXiv preprint arXiv:1911.09712*, 2019.

[59] J.-R. Chang and Y.-S. Chen, "Pyramid stereo matching network," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 5410–5418.

[60] N. Mayer, E. Ilg, P. Hausser, P. Fischer, D. Cremers, A. Dosovitskiy, and T. Brox, "A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 4040–4048.

[61] K. Yamaguchi, D. McAllester, and R. Urtasun, "Efficient joint segmentation, occlusion labeling, stereo and flow estimation," in *European Conference on Computer Vision*. Springer, 2014, pp. 756–771.

[62] H. Fu, M. Gong, C. Wang, K. Batmanghelich, and D. Tao, "Deep ordinal regression network for monocular depth estimation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 2002–2011.

[63] H.-N. Hu, Q.-Z. Cai, D. Wang, J. Lin, M. Sun, P. Krahenbuhl, T. Darrell, and F. Yu, "Joint monocular 3d vehicle detection and tracking," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 5390–5399.

[64] C. R. Qi, W. Liu, C. Wu, H. Su, and L. J. Guibas, "Frustum pointnets for 3d object detection from rgb-d data," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 918–927.

[65] B. Xu and Z. Chen, "Multi-level fusion based 3d object detection from monocular images," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 2345–2353.

[66] Z. Zhou, L. Du, X. Ye, Z. Zou, X. Tan, E. Ding, L. Zhang, X. Xue, and J. Feng, "Sgm3d: Stereo guided monocular 3d object detection," *arXiv preprint arXiv:2112.01914*, 2021.

[67] R. Kesten, M. Usman, J. Houston, T. Pandya, K. Nadhamuni, A. Ferreira, M. Yuan, B. Low, A. Jain, P. Ondruska *et al.*, "Lyft level 5 perception dataset 2020," 2019.

[68] C. Reading, A. Harakeh, J. Chae, and S. L. Waslander, "Categorical depth distribution network for monocular 3d object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 8555–8564.

[69] P. Sun, H. Kretzschmar, X. Dotiwalla, A. Chouard, V. Patnaik, P. Tsui, J. Guo, Y. Zhou, Y. Chai, B. Caine *et al.*, "Scalability in perception for autonomous driving: Waymo open dataset," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 2446–2454.

[70] L. Chen, J. Sun, Y. Xie, S. Zhang, Q. Shuai, Q. Jiang, G. Zhang, H. Bao, and X. Zhou, "Shape prior guided instance disparity estimation for 3d object detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021.

[71] Z. Qin, J. Wang, and Y. Lu, "Triangulation learning network: from monocular to stereo 3d object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 7615–7623.

[72] Y. Shi, Z. Mi, and Y. Guo, "Stereo centernet based 3d object detection for autonomous driving," *arXiv preprint arXiv:2103.11071*, 2021.

[73] W. Peng, H. Pan, H. Liu, and Y. Sun, "Ida-3d: Instance-depth-aware 3d object detection from stereo vision for autonomous driving," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 13 015–13 024.

[74] Y. Chen, S. Liu, X. Shen, and J. Jia, "Dsgn: Deep stereo geometry network for 3d object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 12 536–12 545.

[75] X. Chen, K. Kundu, Z. Zhang, H. Ma, S. Fidler, and R. Urtasun, "Monocular 3d object detection for autonomous driving," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 2147–2156.

[76] P. Li, X. Chen, and S. Shen, "Stereo r-cnn based 3d object detection for autonomous driving," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 7644–7652.

[77] H. Königshof, N. O. Salscheider, and C. Stiller, "Realtime 3d object detection for automated driving using stereo vision and semantic information," in *2019 IEEE Intelligent Transportation Systems Conference (ITSC)*. IEEE, 2019, pp. 1405–1410.

[78] Z. Wu, C. Shen, and A. Van Den Hengel, "Wider or deeper: Revisiting the resnet model for visual recognition," *Pattern Recognition*, vol. 90, pp. 119–133, 2019.

[79] S. Li and K.-T. Cheng, "Joint stereo 3d object detection and implicit surface reconstruction," *arXiv preprint arXiv:2111.12924*, 2021.

[80] Y. Liu, L. Wang, and M. Liu, "Yolostereo3d: A step back to 2d for efficient stereo 3d detection," *arXiv preprint arXiv:2103.09422*, 2021.

[81] A. Gao, Y. Pang, J. Nie, J. Cao, and Y. Guo, "Egfn: Efficient geometry feature network for fast stereo 3d object detection," *arXiv preprint arXiv:2111.14055*, 2021.

[82] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[83] Y. Zhang, J. Lu, and J. Zhou, "Objects are different: Flexible monocular 3d object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 3289–3298.

[84] P. Li, H. Zhao, P. Liu, and F. Cao, "Rtm3d: Real-time monocular 3d detection from object keypoints for autonomous driving," *arXiv preprint arXiv:2001.03343*, vol. 2, 2020.

[85] Y.-N. Chen, H. Dai, and Y. Ding, "Pseudo-stereo for monocular 3d object detection in autonomous driving," 2022.

[86] A. D. Pon, J. Ku, C. Li, and S. L. Waslander, "Object-centric stereo matching for 3d object detection," in *2020 IEEE International Con-*

ference on Robotics and Automation (ICRA). IEEE, 2020, pp. 8383–8389.
[87] J. Sun, L. Chen, Y. Xie, S. Zhang, Q. Jiang, X. Zhou, and H. Bao, "Disp r-cnn: Stereo 3d object detection via shape prior guided instance disparity estimation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 10 548–10 557.
[88] T. He and S. Soatto, "Mono3d++: Monocular 3d vehicle detection with two-scale 3d hypotheses and task priors," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, no. 01, 2019, pp. 8409–8416.
[89] B. Li, W. Ouyang, L. Sheng, X. Zeng, and X. Wang, "Gs3d: An efficient 3d object detection framework for autonomous driving," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 1019–1028.
[90] Z. Qin, J. Wang, and Y. Lu, "Monogrnet: A geometric reasoning network for monocular 3d object localization," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, no. 01, 2019, pp. 8851–8858.
[91] F. Manhardt, W. Kehl, and A. Gaidon, "Roi-10d: Monocular lifting of 2d detection to 6d pose and metric shape," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 2069–2078.
[92] I. Barabanau, A. Artemov, E. Burnaev, and V. Murashkin, "Monocular 3d object detection via geometric reasoning on keypoints," *arXiv preprint arXiv:1905.05618*, 2019.
[93] Y. Cai, B. Li, Z. Jiao, H. Li, X. Zeng, and X. Wang, "Monocular 3d object detection with decoupled structured polygon estimation and height-guided depth estimation," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 07, 2020, pp. 10 478–10 485.
[94] J. Ku, A. D. Pon, and S. L. Waslander, "Monocular 3d object detection leveraging accurate proposals and shape reconstruction," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 11 867–11 876.
[95] Z. Liu, Z. Wu, and R. Tóth, "Smoke: single-stage monocular 3d object detection via keypoint estimation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2020, pp. 996–997.
[96] T. Roddick, A. Kendall, and R. Cipolla, "Orthographic feature transform for monocular 3d object detection," *arXiv preprint arXiv:1811.08188*, 2018.
[97] M. Ding, Y. Huo, H. Yi, Z. Wang, J. Shi, Z. Lu, and P. Luo, "Learning depth-guided convolutions for monocular 3d object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2020, pp. 1000–1001.
[98] S. Srivastava, F. Jurie, and G. Sharma, "Learning 2d to 3d lifting for object detection in 3d for autonomous vehicles," *arXiv preprint arXiv:1904.08494*, 2019.
[99] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1125–1134.
[100] G. P. de La Garanderie, A. A. Abarghouei, and T. P. Breckon, "Eliminating the blind spot: Adapting 3d object detection and monocular depth estimation to 360 panoramic imagery," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 789–807.
[101] A. Dosovitskiy, G. Ros, F. Codevilla, A. Lopez, and V. Koltun, "Carla: An open urban driving simulator," in *Conference on robot learning*. PMLR, 2017, pp. 1–16.
[102] L. Liu, J. Lu, C. Xu, Q. Tian, and J. Zhou, "Deep fitting degree scoring network for monocular 3d object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 1057–1066.
[103] W. Bao, B. Xu, and Z. Chen, "Monofenet: Monocular 3d object detection with feature enhancement networks," *IEEE Transactions on Image Processing*, vol. 29, pp. 2753–2765, 2019.
[104] W. Bao, Q. Yu, and Y. Kong, "Object-aware centroid voting for monocular 3d object detection," 2020.
[105] D. Zhou, X. Song, Y. Dai, J. Yin, F. Lu, M. Liao, J. Fang, and L. Zhang, "Iafa: Instance-aware feature aggregation for 3d object detection from a single image," in *Proceedings of the Asian Conference on Computer Vision*, 2020.
[106] Y. Lu, X. Ma, L. Yang, T. Zhang, Y. Liu, Q. Chu, J. Yan, and W. Ouyang, "Geometry uncertainty projection network for monocular 3d object detection," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 3111–3121.
[107] L. Liu, C. Wu, J. Lu, L. Xie, J. Zhou, and Q. Tian, "Reinforced axial refinement network for monocular 3d object detection," in *European Conference on Computer Vision*. Springer, 2020, pp. 540–556.
[108] L. Wang, L. Du, X. Ye, Y. Fu, G. Guo, X. Xue, J. Feng, and L. Zhang, "Depth-conditioned dynamic message propagation for monocular 3d object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 454–463.
[109] S. Mehtab, W. Q. Yan, and A. Narayanan, "3d vehicle detection using cheap lidar and camera sensors," in *2021 36th International Conference on Image and Vision Computing New Zealand (IVCNZ)*. IEEE, 2021, pp. 1–6.
[110] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "Mobilenetv2: Inverted residuals and linear bottlenecks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 4510–4520.
[111] A. Simonelli, S. R. Bulo, L. Porzi, M. López-Antequera, and P. Kontschieder, "Disentangling monocular 3d object detection," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 1991–1999.
[112] J. Zhang, H. Liu, and J. Lu, "A semi-supervised 3d object detection method for autonomous driving," *Displays*, vol. 71, p. 102117, 2022.
[113] M. Liang, B. Yang, Y. Chen, R. Hu, and R. Urtasun, "Multi-task multi-sensor fusion for 3d object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 7345–7353.
[114] W. Ali, S. Abdelkarim, M. Zidan, M. Zahran, and A. El Sallab, "Yolo3d: End-to-end real-time 3d oriented object bounding box detection from lidar point cloud," in *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*, 2018, pp. 0–0.

Simegnew Yihunie Alaba received B.S. degree in Electrical Engineering from Arbaminch University and M.S. degree in Computer Engineering from Addis Ababa University. He is currently pursuing a Ph.D. degree in Electrical and Computer Engineering at Mississippi State University. His research interests include image processing, computer vision, deep learning, and autonomous driving. He is a member of IEEE.

John E. Ball received B.S. and Ph.D. degrees in Electrical Engineering from Mississippi State University in 1991 and 2007, respectively, and the M. S. in Electrical Engineering from the Georgia Institute of Technology in 1993. Currently, Dr. Ball is an Associate Professor and Endowed Chair in Electrical and Computer Engineering at Mississippi State University. His research interests include sensors, sensor processing, deep learning, and autonomous vehicles, especially in an unstructured environment. Dr. Ball is a senior member of IEEE and serves as an associate editor for IEEE Signal Processing Letters and the Journal of Applied Remote Sensing.