Vision based 3D Object Detection using Deep Learning: Methods with Challenges and Applications towards Future Directions

A F M Saifuddin Saif[1]
School of Engineering
Aalto University
Finland

Zainal Rasyid Mahayuddin[2]
Faculty of Information Science and Technology
Universiti Kebangsaan Malaysia
Selangor, Malaysia

***Abstract*—For autonomous intelligent systems, 3D object detection can act as a basis for decision making by providing information such as object's size, position and direction to perceive information about surrounding environment. Successful application using robust 3D object detection can hugely impact robotic industry, augmented and virtual reality sectors in the context of Fourth Industrial Revolution (IR4.0). Recently, deep learning has become potential approach for 3D object detection to learn powerful semantic object features for various tasks, i.e., depth map construction, segmentation and classification. As a result, exponential development in the growth of potential methods is observed in recent years. Although, good number of potential efforts have been made to address 3D object detection, a depth and critical review from different viewpoints is still lacking. As a result, comparison among various methods remains challenging which is important to select method for particular application. Based on strong heterogeneity in previous methods, this research aims to alleviate, analyze and systematize related existing research based on challenges and methodologies from different viewpoints to guide future development and evaluation by bridging the gaps using various sensors, i.e., cameras, LiDAR and Pseudo-LiDAR. At first, this research illustrates critical analysis on existing sophisticated methods by identifying six significant key areas based on current scenarios, challenges, and significant problems to be addressed for solution. Next, this research presents strict comprehensive analysis for validating 3D object detection methods based on eight authoritative 3D detection benchmark datasets depending on the size of the datasets and eight validation matrices. Finally, valuable insights of existing challenges are presented for future directions. Overall extensive review proposed in this research can contribute significantly to embark further investigation in multimodal 3D object detection.**

***Keywords*—3D object detection; deep learning; vision; depth map; point cloud**

I. Introduction

3D object detection provides precise representation of objects in the format of semantically meaningful 3D bounding boxes. 3D object detection aims to categorize and localize objects from various sensors data, i.e., monocular and stereo cameras [1, 2], LiDAR point clouds [3], to understand the 3D visual world and associated semantic labels for objects in 3D scenes, has attracted increasing attention from vision community. In addition, advances of deep learning facilitate the rapid progress of 3D object detection indicates strong application demands which can serve numerous applications, i.e., robotics, autonomous driving, augmented reality, virtual reality, robot navigation, enabling systems to understand their environment and react accordingly. As a result, there has been a surge of interest for developing improved 3D object detection pipeline. Although current methods show impressive performance despite the facts that various problems form different viewpoints were observed and illustrated by this research.

This research identified four major approaches along with deep neural networks for 3D object detection, i.e., monocular images, stereo images, LiDAR and Pseudo LiDAR based approaches. 3D object detection from monocular frames is a fertile research area due to potentially vast impact, ubiquity of cameras, low expense, easy implicated solution with one camera [4]. However, estimation of depth from single monocular images is an ill-posed inverse problem causes accuracy of 3D detection from only monocular images is lower than that from LiDAR or stereo images [2, 5]. In addition, loss of significant information during calibration is another reason for degraded performance on the same 3D object detection benchmarks comparing with LiDAR and stereo methods. Advances of deep neural network facilitate immensely the progress of 3D object detection using LiDAR and stereo based methods. The inclusion of depth information allows capturing the three-dimensional structure of the object's environment, is a key feature for ensuring robustness while maintaining high accuracy. Modern LiDAR acquisition sensors provide meaningful information not only for avoiding imminent collisions, but also to perceive the environment as good as image-based data and even surpass it under poor lighting conditions. For stereo images, calibration issues between two camera rigs [6] and occupation of more pixels for nearby objects than far way objects during perspective projection are considered as major challenges. Besides, high cost of LiDAR sensor encourages researchers to look for alternatives such as Pseudo LiDAR based approaches. In this context, Pseudo LiDAR based approaches uses pre-trained depth network to compute an intermediate point cloud representation to mimic LiDAR data and then fed to a 3D detection network [7]. The strength of Pseudo LiDAR based approaches is that they monotonically improve with depth estimation quality although for long distance object Pseudo

LiDAR approaches could not provide expected detection outcome. However, accountability of the depth estimation is the prime gap between Pseudo-LiDAR and LiDAR based 3D object detection. In this context, simpler end-to-end monocular 3D detectors shows strong promise as an option where lack of same scalability from unsupervised pre-training for their one stage nature is the main drawbacks in this context. Besides, existing datasets for validating 3D object detection was not generalized well for different weather conditions and geographical locations, i.e., any method trained on Waymo datasets [8] suffer from dramatic performance drop on KITTI dataset [1]. Therefore, approaches to effectively adaptive 3D object detection method are highly demanded for practical applications where environment varies significantly. In this context, recent success for various 3D object methods mostly depends on larger datasets of 3D scenes where annotations are done carefully and remains as main bottleneck in the context of available datasets.

In summary, the contributions of this work are:

1) Six key areas are identified to analysis existing 3D object detection methods from different viewpoints to guide future development and evaluation by bridging the gaps using various sensors, i.e., cameras, LiDAR and Pseudo- LiDAR.

2) This research demonstrates extensive experimental analysis based on existing research depending on three major aspects, i.e., comprehensive insights on hardware and software, analysis and challenges of using six datasets with details specification, illustration of eight performance metrics required for validation of 3D object detection methods.

3) Based on comprehensive previous research investigation on existing 3D object detection methods, six key observations are elaborated for improving future 3D object detection methods.

II. Background

3D object detection methods mostly depend on four types of input pattern, i.e., monocular image, stereo image, LiDAR point cloud and pseudo-LiDAR signal. Research in [9] also used monocular RGB images as input and mapped image features into an orthographic 3D space by describing deep learning architecture for estimating 3D bounding boxes. They performed front-end feature extractor to extract features from monocular RGB images to form 3D features space and the transformation from 2D features maps to 3D features map was termed as Orthographic Feature Transform (OFT). However, validation on large variations of scales and distances could provide real time usage of their proposed method. Research in [10] used monocular RGB images as input to predict 3D human locations by learning data ambiguity without supervision which leads to predict confidence intervals with point estimation. They used Laplace loss to model Aleatoric Uncertainty and multivariate Gaussians to model Epistemic Uncertainty. However, loss of multiview visual characteristics and spatial structure characteristics from 2D human poses for 3D localization might be the reason for low accuracy compared with other research results.

Research in [1] used stereo data as input to produce set of 3D object proposals which run through convolutional neural network for high quality 3D object detection. Their proposed method generated 3D proposals using energy minimization function to encode object size priors, context of ground plane and features for depth information. However, they used 3D integral images which are not suitable for real time 3D object detection. Research in [4] used stereo data to predict sparse key point for estimating 3D bounding box. They proposed Stereo R-CNN to improve overall performance. However, recent advancement of R-CNN like Faster R-CNN should be implicated to justify the effectiveness of their proposed method. Research in [2] used stereo imagery as input to estimate depth using convolutional neural net (CNN) [11] to compute point clouds. They used proposal generation problem as inference in Markov Random Field (MRF) to encode high density in the point cloud. However, overall performance of their proposed method depends on accurate depth estimation.

Research in [2] used LiDAR point cloud and RGB images as input to generate 3D proposals and projected them to multiple views for feature extraction. They used region-based fusion network to deeply integrate Multiview information for classification of 3D proposals. However, region based fusion network works-based region using convolutional neural network [12] requires more computation overheads for real time detection. In addition, LiDAR sensor is expensive whereas optical camera could be a potential alternative for their proposed method. Research in [7] generated 3D proposal by using raw point cloud instead of generating proposals from RGB image or projecting point cloud to bird's view or voxels. They used PointNet++ to learn point-wise features for describing the raw point clouds which later used for foreground point cloud segmentation. However, for sparse convolutions, alternative point-cloud network structures, such as VoxelNet could be investigated as their backbone network. Research in [13] used RGBD data as input for raw point clouds and proposed framework for RGB-D data-based 3D object detection called Frustum PointNets. Their frustum region is based on 2D region proposal indicates that no 3D object will be detected without 2D region proposal or 2D detection. However, aggregation of image feature after extracting features using based on their backbone network could improve overall 3D object detection performance. Although, point clouds can provide detailed geometry and capture 3D structure of the scene, on the other hand, point clouds are irregular, which cannot be processed by powerful deep learning models, such as convolutional neural networks directly [14]. Research in [5] used both monocular images and stereo images and estimated depth and disparity for monocular and stereo images respectively to generate point clouds. They proposed two step approach by first extracting dense pixel depth from stereo or monocular imagery followed by back-projecting pixels into a 3D point cloud to view the representation as pseudo-LiDAR signal. However, for long distance objects pseudo-LiDAR signal could not provide expected detection results comparing with short distance objects.

III. Analysis of Methods based on Six Aspects

This research identified six significant key areas to analysis existing methods for 3D object detection mentioned in Fig. 1 and comprehensively illustrated in the next subsequent sections.

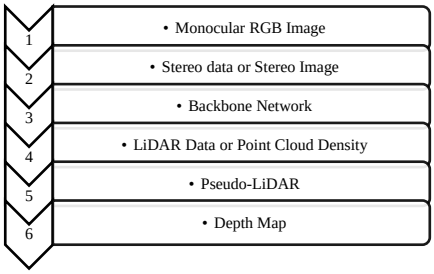

Fig. 1. Six Key Areas to Analysis Existing Methods for 3D Object Detection.

A. Monocular rgb Images

Although 3D point cloud extracted from LiDAR provides superior performance, previously less effective alternative was monocular RGB images that were collected from single view RGB image as input for monocular 3D object detection [15]. Usage of monocular RGB images can facilitate domain adaptation for multimodal big data management and significantly aids for 3D model retrieval and classification when compared with multiview image sets [16]. Research in [4] used a single monocular image to generate class specific object proposals to run through standard CNN pipelines for 3D object detection. However, they assumed that objects should be on a ground plane which makes the overall proposed methodology uncomfortable to be used for other mediums such as detection from UAV, UGV or Krane. Research in [10] used monocular RGB images to tackle ill-posed problem of 3D human localization. They used Laplace distribution to address ambiguity by predicting confidence intervals of 3D bounding boxes. However, they used 2D human poses for 3D localization which might be the reason for low accuracy compared with other research results.

Research in [5] used both monocular and stereo images to mimic LiDAR signal and depth map constructions using the Pseudo-Signal hence called Pseudo-LiDAR and provided a milestone instead of using LiDAR sensor for point cloud generation. However, Pseudo-LiDAR approach required additional post processing steps due to the range issue of objects from the source of the camera platform.

If these pose processing can be eradicated, then Pseudo-LiDAR can be considered as a potential option to generate point cloud and depth map. Research in [9] used orthographic feature transforms from monocular image as part of an end-to-end deep learning architecture to map image-based features into an orthographic 3D space. However, fruitful experimentation on large variations of scales and distances can provide real time usage of their proposed method. There are several challenges exists for using monocular images for 3D object detection, i.e., significant information loss, object pose inconsistency, complex background of monocular images [16], accurate depth information [6]. Monocular images lose significant information during calibration such as multiview visual characteristics and spatial structure characteristics. Overall scenarios of using monocular images are mentioned in Fig. 2.

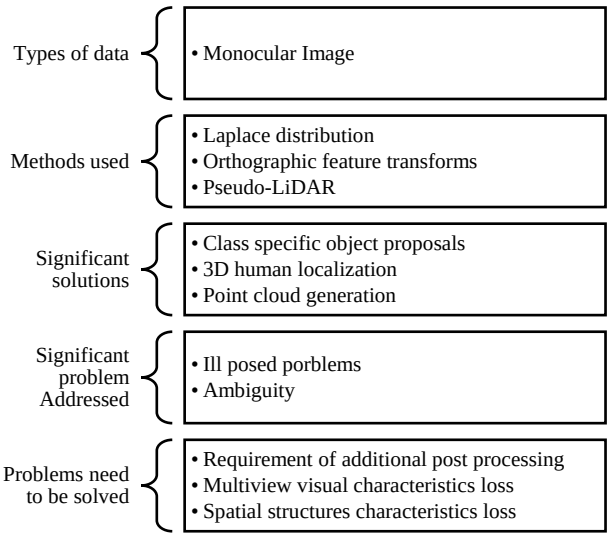

Fig. 2. Scenarios of using Monocular RGB Images for 3D Object Detection.

B. Stereo rgb Images

Stereo cameras work in a manner like human binocular vision which is cost less and have higher resolutions for which they have gathered significant attention in academia and industry [17]. Disparity cues are provided by stereo images to enable better depth estimation compared to monocular images [18]. Research in [1] exploited stereo imagery for high quality 3D object proposals using energy minimization function to encode object size, localization of objects on the depth informed features to reason free space, point cloud densities and distance to the ground. They used CNN to exploit depth information for regressing 3D bounding box coordinates and object pose. However, they used 3D integral images which might cause additional computational cost during training. Research in [4] used stereo R-CNN for detecting associated objects in the stereo images. To achieve the aim, they proposed a 3D box estimator to exploit stereo box key points and dense region-based photometric alignment method to improve 3D object localization accuracy. However, due to later advancement of R-CNN like Faster R-CNN could be better option to investigate in their proposed method. Research in [5] proposed a twostep approach using stereo imagery data, i.e., estimation of depth map from stereo and usage of existing LiDAR-based 3D object detection pipelines. Although they used both monocular and stereo images, it is not clear based on the elaboration whether they used combined depth maps for both types of images or not. Research in [2] used contextual models by exploiting stereo information by reasoning and placing in 3D proposals in the form of 3D bounding boxes. Although, they implicated their proposed method for autonomous driving, they claimed better performance on other object classes such as Cyclist and Pedestrians. Several other shortcomings for using stereo images are observed by this research although stereo image based detection shown promising results, i.e. calibration issues between two camera rigs [2], memory expensive 3D cost

volume learning in image space due to the image's high resolution and requirement of additional dimension for features [18], depth estimation in image space while downstream detection is performed in 3D space [18], existence of many pixels among nearby objects due to perspective projection than faraway objects which lead to biased depth estimation with degraded long-range detection performance [18], dependency on anchor-based 2D detectors with association approach [17].

For stereo images, in image space, 3D cost volume learning is costly in memory due to high computation and extra dimension of the features. Besides, previous lest computation time for one frame was 0.5 seconds which can be challenging for critical applications such as collision avoidance [17, 20, 21]. Besides, in stereo images, depth estimation is performed in 2D image space whereas detection tasks take place in 3D space. Existence of more pixels in nearby objects than faraway objects due to perspective projection led to imbalance biased depth map. Overall scenarios of using stereo images are mentioned in Fig. 3.

C. LiDAR Data/ Point Cloud Density

Research in [1] extended 3D object proposal with class independent variant and neural network to grab both appearance and depth features. Later, they used point clouds obtained via LiDAR followed by giving comparison of the stereo, LIDAR and hybrid settings. However, class specific module they proposed, needs further elaboration to justify their claim. Research in [3] used sensory-fusion framework to take both LiDAR point cloud and RGB images as input. They encoded sparse 3D point cloud for compact multiview representation.

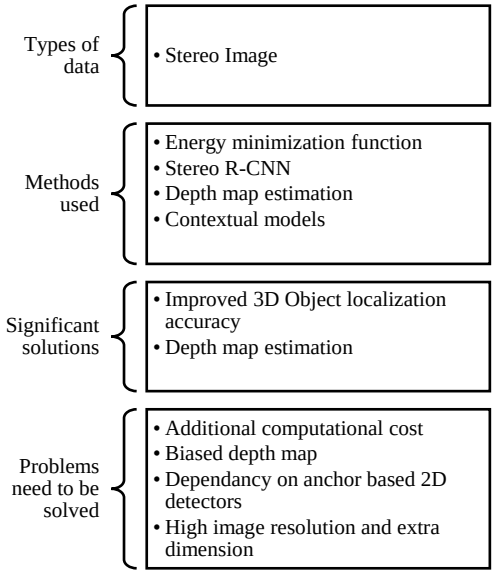

Fig. 3. Scenarios of using Stereo Images for 3D Object Detection.

Their proposed deep learning network was composed of two subnetworks, i.e., one for 3D object proposal generation and other one for multiview features fusion. For multiple views and interactions between intermediary layers they used deep fusion schemes to combine these two parts. However, region based fusion network works-based region using convolutional neural network requires more computation overheads for real time detection. In addition, LiDAR sensor is expensive whereas optical camera could be a potential alternative for their proposed method. Research in [4] used raw point cloud for generating small number of high-quality 3D proposals via segmenting the point clouds of the whole scene into foreground points and background. They performed 3D box refinement later by combing local spatial features with global semantic features [22] for each point cloud. However, later version of RCNN such as Fast RCNN and Faster RCNN could be aligned later to find the suitability of the proposed method with most recent deep learning architectures in case of using CNN for raw point cloud achieved from LiDAR.

D. Pseudo-LiDAR Generation

Previous LiDAR methods formulated LiDAR point cloud as point [7], pillar [23] and voxel representation [29,30]. Despite the remarkable performance of these methods, LiDAR sensors are expensive sensor and for the far distance, LiDAR sensor-based detection can potentially make 3D object detection tasks difficult which initiate the need for some approaches to mimic LiDAR point cloud due to high accuracy called Pseudo-LiDAR based point cloud representation. Research in [26] proposed self-supervised learning schemes by mimicking latent spatial features representation based on point-based module. However, their proposed method depends on 2D-3D detection pairs. Accurate depth map is one of the key challenges for 3D object detection from monocular image as lack of prior information is the main issue and eradicated recently by deep learning approaches [25, 27]. By constructing robust depth map, Pseudo-LiDAR point cloud can be constructed to mimic LiDAR point cloud based on pre-calibrated intrinsic camera parameters [28,29]. However, heavy computation is the bottleneck for this research. Research in [30] generated pseudo-ground without the need of LiDAR point clouds by proposing a statistical shape model to address the challenge of disparity annotations in training. However, for monocular 3D object detection, disparity map generation will not be possible for their proposed method. In addition, their proposed research depends on shape analysis which may provide poor performance in case of obstacles such as shadow. Research in [31] combined 2D object detection with Pseudo-LiDAR point cloud data generated from stereo images to investigate the boost of performance with existing six different 3D object detectors. However, distance calculation for their proposed method can be costly. Besides, they did not investigate high consistency issue for the point clouds generated by Pseudo-LiDAR approach. Research in [32] constructed a pseudo-LiDAR feature volume (PLUME) to estimate depth map and 3D object detection in 3D metric space. The main purpose for their proposed PLUME is to avoid biased depth estimation with degraded long-range detection for stereo images. However, for low light and extreme weather conditions, their proposed method may encounter due to the need of accurate depth map estimation.

E. Depth Map

Recent algorithms for stereo depth estimation can produce surprisingly accurate depth maps [33]. However, inferring depth of pedestrians from monocular images is a fundamentally ill-posed problem [10]. This additional

challenge is due to human variation of height. If every pedestrian has the same height, there would be no ambiguity [10]. LiDAR point clouds has been dominated approach in the existing research for 3D object detection due to the availability of accurate depth map [3, 13, 34, 35, 36, 37, 38, 39]. However, performance of image-only methods lacks in absolute depth information of LiDAR. In this context, optical cameras are highly affordable than LiDAR, operate at a high frame rate, and provide a dense depth map rather than 64 or 128 sparse rotating laser beams that LiDAR signal is inherently limited to [5]. Research in [40] proposed first study for estimating depth map in outdoor environments where they combined single RGB image and LiDAR point cloud. However, due to large learning network, powerful architectures for training were needed to use their proposed method. Research in [5] converted estimated depth map from stereo or monocular imagery into a 3D point cloud referred to as pseudo-LiDAR as it mimics the LiDAR signal. Then they took the advantage of existing LiDAR-based 3D object detection pipelines [13, 34] to train directly on the pseudo-LiDAR representation using deep convolutional neural networks to obtain unprecedented increase in accuracy of image-based 3D object detection algorithms. However, usage of LiDAR based method may not be suitable for other image-based classifier for overall 3D object detection. Besides, their proposed pseudo-LiDAR fails to detect far-away objects precisely due to inaccurate depth estimation. Research in [10] detected 2D joints using PifPaf and used MonoDepth [41] to estimate depth for a set of 9 pixels around each key point followed by consideration of minimum depth as their reference value. Later, they calculated distance from normalized image coordinates of the centre of the bounding box using the estimated minimum depth rather than using the average one. However, their proposed method needs further investigation for monocular image as their method worked only for stereo images. Overall scenarios for depth map estimation using different sensors and images are shown in Fig. 4.

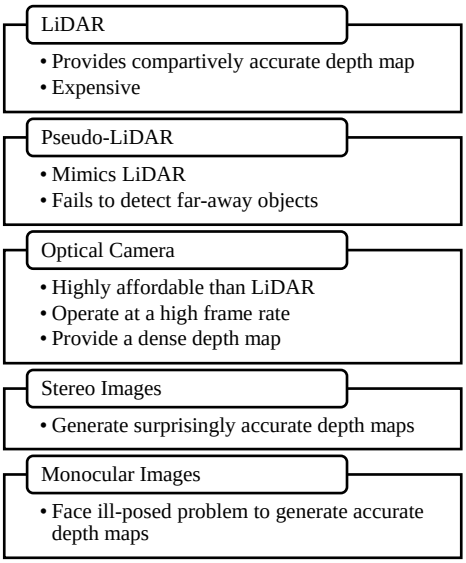

Fig. 4. Overall Scenario for Depth Map Estimation using Different Sensors and Types of Images.

F. Backbone Network

For 3D object detection various backbone networks has been used for feature representation shown in Fig. 5. Research in [42] preferred image features in ResNet-101 [4, 17, 43] for block 1 to maintain a high spatial resolution and avoid redundancy of same features. However, ResNet as backbone network contains too many layers and are not very efficient [4]. Darknet-53 [44] as a backbone network are used with fewer floating-point operations and more speed. Darknet-53 is better than ResNet-101 and 1.5× faster. In addition, Darknet-53 has similar performance to ResNet-152 and is 2× faster. Darknet-53 also achieves the highest measured floating-point operations per second which means the network structure better utilizes the GPU, making it more efficient to evaluate and thus faster.

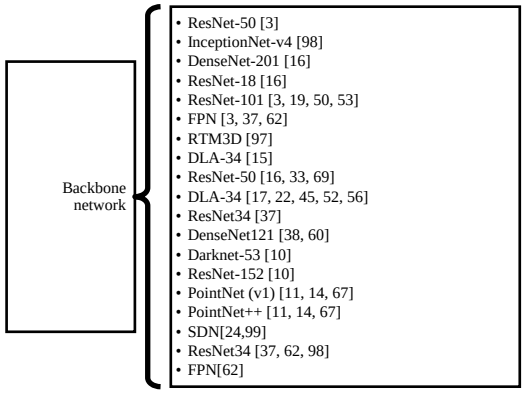

Fig. 5. Backbone Networks for 3D Object Detection.

Research in [13] used PointNet (v1) and PointNet++ (v2) backbone due to much cleaner in design. While out of the scope for their work, sensor fusion in terms with aggregation of image feature for 3D object detection after extracting features using PointNet (v1) and PointNet++ could further improve their results. Research in [7] extracted point-wise features encoded by PointNet++ as backbone point cloud network, they appended one segmentation head for estimating the foreground mask and one box regression head for generating 3D proposals. However, alternative point-cloud network structures, such as VoxelNet with sparse convolutions could be investigated as their backbone network. Research in [16] used ResNet-50 as the backbone for feature extraction followed by a fully connected neural network [45] with one or two hidden layers. The number of hidden nodes is tuned with 64, 100, 128, and 192 to generate the candidate results which cause additional processing due to too many layers. Research in [63] used SDN [46] as backbone network to estimate dense depth map and fine-tuned on the KITTI datasets. However, their predicted depth was not accurate enough since the ground truth is very sparse. Research in [47] used ResNet34 [48, 49] with Feature Pyramid Network (FPN) [48] as backbone network. They replaced BatchNorm+ReLU layers with the synchronized version of InPlaceABN activated with LeakyReLU with negative slope 0.01 as proposed in [50] to

free up a significant amount of GPU memory for scaling up the batch size or input resolution. However, top down and bottom-up features fusion in feature pyramid network might create additional processing due to combine ResNet34 and FPN. Existing methods mostly used ResNet-50 as the backbone network for both monocular and stereo images due to contain too many layers [4]. InceptionNet-v4 [49] was previously used for visual feature extraction, however, in case of 3D models for multiview mages, InceptionNet-v4 needs to be further investigated for 3D object detection overall pipeline. DenseNet-201 [16] was used to extract the visual features of monocular images and ResNet-18 [16] was for multiview feature extraction of 3-D models. However, combination two backbone network at two different phases, may increase overall computational overheads. In case of monocular images and multiview fusion, four key aspects can ensure efficient use of backbone network for 3D object detection mentioned in Fig. 6 [16].

Domain adaptation is the first concerns for backbone network to be used with any method or strategy to solve cross domain problem. In this context, improved datasets developed with the concerns about cross domain ease the task for backbone network. For monocular and multiview aspects different fusion strategies for multiview views are needed for the extracted features using backbone network, such as pooling, concatenation. As monocular images lose multiview visual characteristic and spatial structure characteristic causes significant information loss, various function needs to be designed during features learning process. Efficient implication of the above three aspects helps for efficient pair wise similarity measurement.

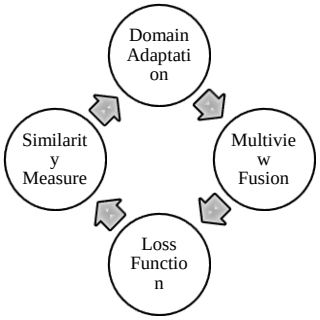

Fig. 6. Four Key Aspects for Efficient Use of Backbone Network for 3D Object Detection.

IV. Review on Experimental Analysis

A. Hardwares and Software Specification for Experimentation

Previously researcher used sensors such as Grayscale cameras [51], optical cameras [5], color cameras [51] and LiDAR like Velodyne HDL-64E LIDAR [5, 51] for own datasets development for validating 3D object detection evaluation. For training, various GPUs were used mentioned in Fig. 7. PyTorch machine learning framework [42, 52, 53, 54, 55, 56] have been mostly used by existing research.

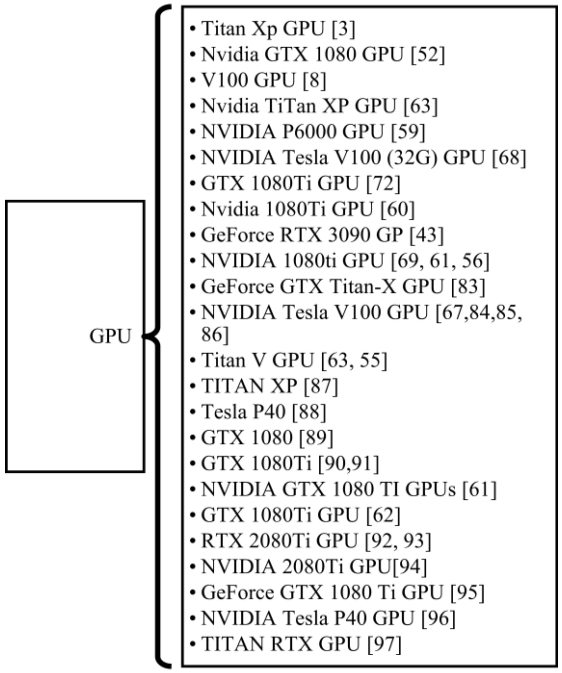

Fig. 7. GPUs Used in Previous Research.

B. Datasets

Domain gaps for different datasets depend on object size, weather condition, specific locations, and orientation [57]. Overall gaps can be categorized in to two categories mentioned below.

1) Content gap such as object, weather condition due to locations during data capture depending on time.

2) Point distribution gap owing to different LiDAR types such as number of beam ways, beam range, vertical inclination, horizontal and vertical angularity estimation of LiDAR. Existing datasets used for 3D object detection purpose are mentioned in Fig. 8.

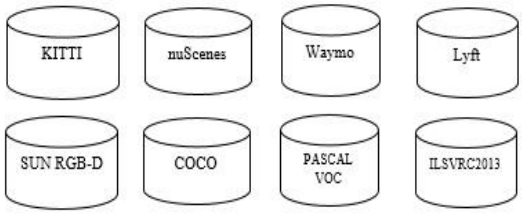

Fig. 8. Datasets for 3D Object Detection.

1) KITTI dataset: KITTI is the most used dataset for 3D object detection. However, existing datasets such as nuScenes, Waymo, Lyft, SUN RGB-D, COCO, PASCAL VOC and ILSVRC2013 shown in Fig. 8 has been used by the research for more robust validation.

The KITTI 3D dataset [58, 59] is the most used benchmark in the 3D object detection task and it provides left camera images, calibration files, annotations for 3D detection [45].

KITTI datasets are the widely used benchmark for validating 3D object detection includes 2D object detection, Average Orientation Similarity (AOS), Bird's Eye View (BEV) [13, 60, 61]. Samples in KITTI datasets include 3D point clouds, images and Camera-LiDAR calibration data [52]. Images in KITTI datasets are captured in the same city using same camera [10]. 3D bounding boxes for various object classes are provided in KITTI datasets which includes cars, vans, trucks, pedestrians and cyclists labelled manually in 3D point clouds depending on calibrated camera's information [62]. Number of training and test images or point clouds in KITTI datasets are 7481 and 7518 images respectively containing three classes, i.e., car, pedestrian and cyclist [1,2, 8, 42, 52, 54, 60, 61, 63, 64, 65, 66, 75]. Each class is annoted by camera Field of Vision (FOV) with the 3D bounding boxes [64]. Evaluation for each class depends into three categories, i.e., easy, moderate, hard according three aspects, i.e., object size, occlusion state and maximum truncation levels of objects [10, 42, 53, 55]. For ranking the completion of the methods, moderate category is used in the benchmark [2]. Easy object is indicated with minimum pixel height as 40px within 28m as vehicles correspondence. Besides, 25px are the limit for moderate and hard level objects within minimum distance of 47m [62].

In another way, there are three commonly used data splits in the KITTI dataset, i.e., split for testing, validation category 1, and validation category 2 mentioned in Fig. 9.

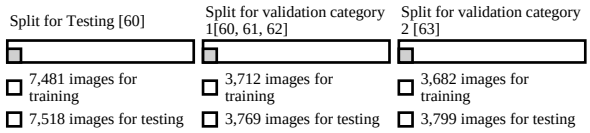

Fig. 9. Split of Images for Training and Testing in KITTI Datasets.

In KITTI datasets, object detection validation is estimated mostly through average precision (AP) and IOU (Intersection over union) with threshold 0.7 for car class [8, 53, 42, 67, 62], 0.5 for pedestrian [60, 67, 62] and 0.5 for cyclist [60, 67, 62]. Six illustrative factors for using KITTI datasets is shown in Fig. 10.

In addition, this research identified ten challenges to use KITTI dataset mentioned in Fig. 11 although KITTI dataset is the commonly used dataset for 3D object detection performance validation.

2) nuScenes dataset: nuScenes dataset contains 1000 segments of 20 seconds each for 3D object detection where 750, 150 and 150 segments for training, validation and testing, respectively [10, 68, 69, 57, 70]. Annotation rate is 2Hz for which 28k, 6k and 6k annotated frames are available in these datasets for training, validation and testing respectively. This dataset contains more classes comparing with KITTI datasets which is 10 and evaluation metrics are mean average precision and nuScenes detection score (NDS) [68]. BEV center distance is the true positive metric for this dataset instead of IoU which is another significant difference with KITTI datasets. However, camera extrinsic information is not also available in this dataset like in KITTI datasets [69].

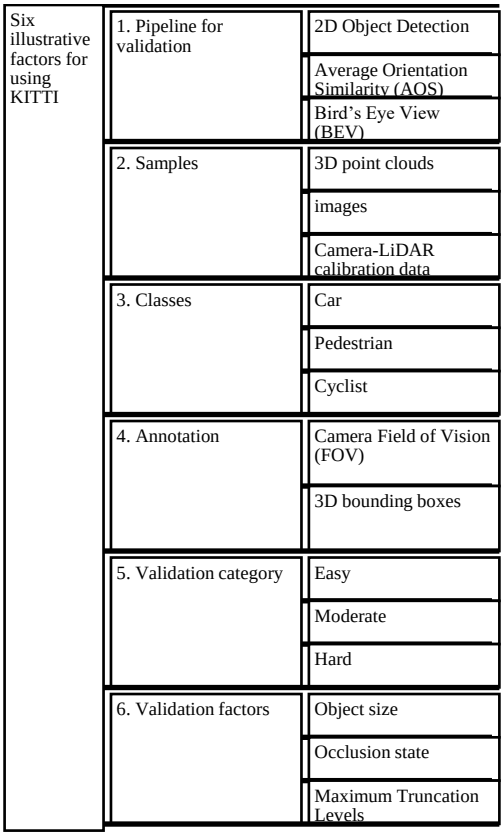

Fig. 10. Six Illustrative Factors to use KITTI Datasets for 3D Object Detection.

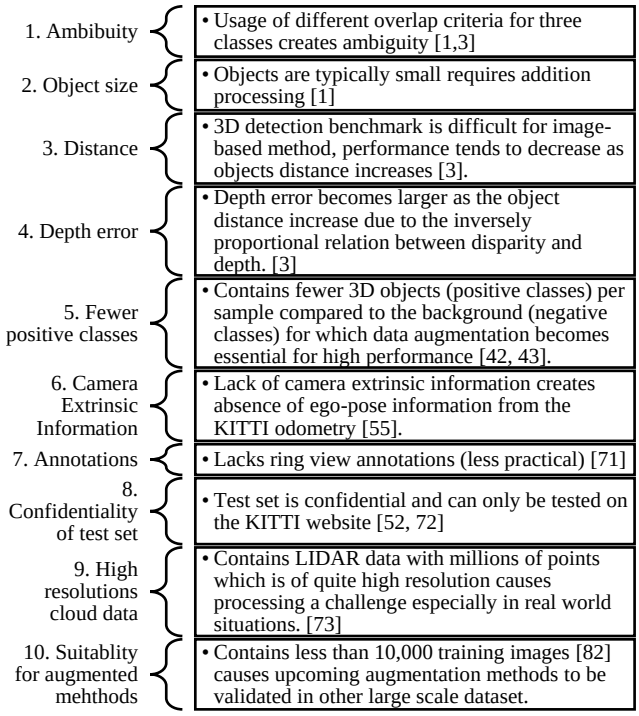

Fig. 11. Ten Challenges to use KITTI Datasets for 3D Object Detection Validation Performance.

3) Waymo dataset: Waymo Open Dataset is more recently released datasets consists of 798 training sequences and 202 validation sequences [8, 53, 71, 42, 72, 57]. Waymo dataset provides object labels in the full 360◦ field of view with a multi-camera rig which is the advantage over KITTI and NuSceneses dataset. However, Waymo dataset includes only 150 test sequences without ground truth data. In addition, there is no published depth results on Waymo open dataset [15].

In addition, with KITTI, nuScenes and Lyft datasets, for 3D object detection other datasets such Lyft [68, 57], PASCAL VOC [73, 74], ILSVRC2013 [73, 74] dataset was previously used for validation. However, these datasets use different annotation rules for validation, i.e., large number of objects outside the road were not annotated for validation.

C. Validation Metrics

Overall framework for 3D object detection has been validated based on two tasks, i.e., object detection and then object detection with orientation estimation [75,103]. To validate accurate object detection, Average Precision (AP) metric was used by most of the existing research, Average Orientation Similarity (AOS) was used mostly for object detection and orientation estimation. The true positive metric is based on 2D/3D IoU. This research identified eight major performance metrics for validating any method for 3D object detection mentioned in Fig. 12.

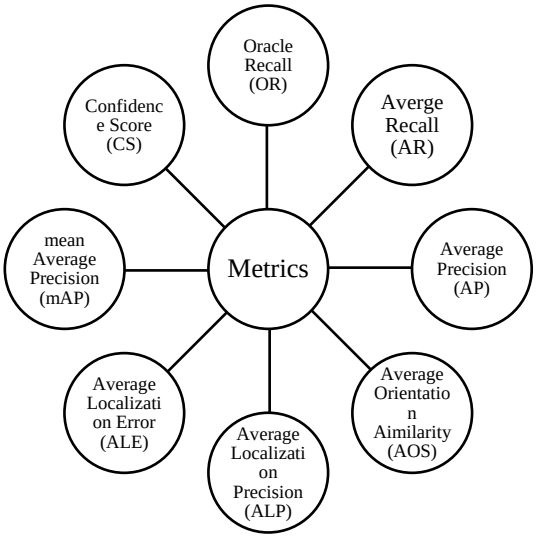

Fig. 12. Major Eight Performance Metrics for Validation for 3D Object Detection.

1) Oracle Recall: Oracle recall computes the percentage of the recalled ground truth objects to receive recall rate [75, 2, 76, 104,105]. For certain threshold, if at least one proposal overlaps with IoU, then ground truth object is said to be recalled [1]. Recall is measured for both 2D bounding box and later for 3D bounding box for overall performance.

2) Average Recall (AR): For 3D object detection performance, Average Recall (AR) [32] is highly correlated metrics that needs to be measured for both 2D bounding box and 3D bounding box for overall performance [1,4,75,32]. For stereo AR metrics, Average Recall needs to be measured for both left and right images [3].

3) Average Precision (AP): For various sampled points, average precision (AP) extracts average value of precision at various recall threshold values [62].

In other words, Average Precision (AP) indicates the average precision value for recall over 0 to 1. For precision and recall, ideal value is 1 [62]. However, for the real time scenarios, any method is assumed to be good enough if the precision and recall metrics gets closer to 1. In this context, this research observed that there is a trade-off between precision and recall, i.e. if more optimizations can be done on precision, recall gets lower, oppositely if recall can be improved, precision value becomes lower. So, this research recommends to balance at the point of fixing threshold point. Average Precision (AP) is used for 2D and 3D object detection for both monocular and stereo images [1, 4, 8, 53, 62, 63,75,77, 83,102]. Like AR, stereo AP metric needs to be evaluated for both on left and right images.

4) Average Orientation Similarity (AOS): Average Orientation Similarity (AOS) indicates perfect prediction between 0 and 1 [1,2, 75]. For 3D object detection, AOS has been used for orientation estimation task for 2D object detection performance towards 3D object detection [75, 15].

5) Average Localization Precision (ALP) [1,10]: Average precision and recall are the requirement to calculate Average Localization Precision (ALP) which can be computed similarly to AP except that 3D localization precision needs to be replaced in pace of bounding box overlap [1]. In other words, ALP provides a prediction to be correct depending on the error between predicted distance and ground truth is smaller than threshold [10]. In this context, predicted 3D location is to be correct if the distance to the ground truth 3D location is smaller than certain threshold [1].

6) Average Localization Error (ALE) [10]: For misaligned bounding box, Average Localization Error (ALE) is estimated from the target category [10]. In other words, ALE provides variation of the actual and estimated value of each location in the localization process.

7) mean Average Precision (mAP): mean Average Precision (mAP) is calculated from Average Precision (AP) from all classes for the IoU thresholds depending on various problems scenario [42,78, 101].

8) Confidence Score: Confidence score indicates the optimum threshold to categorize false positives to ensure the predicted bounding box contains minimum standard score and often used for model performance evaluation [19,79,106]. Non optimal settings for any proposed model requires more minimized confidence score for precise bounding box detection for 3D object detection. 3D box confidence estimation for 3D object detection realized by the previous research. Research in [80] calculated 3D IoU from the predicted 3D box and ground truth involves 3D object dimensions. Research in [47] used 3D box loss to represent 3D detection. Research in [81] introduced self-balancing

confidence loss for generating confidence score from relatively achievable samples. However, all these methods considered loss function for confidence score. To overcome this problem, research in [82] considered the relationship between 3D objects and associated 2D boxes to decompose confidence mechanisms. However, in case of weak transformation for 2D to 3D, their proposed confidence decomposition may result in weakness for their proposed methodology. Research in [10] addressed ill posed problems for predicting confidence intervals to account aleatoric and epistemic uncertainties. They estimated pose to obtain 2D joints which later were used as input to feedforward network and output the 3D location along with a confidence interval. However, their proposed method is suited only for small training data [100]. Research in [83] propagated information from the labelled to unlabeled training set in the form of pseudo-labels contains significant noise for which they introduced confidence-based filtering mechanism for 3D object detection. Their confidence proposals were based on predicted objectless and class probability to filter low quality pseudo labels. However, their proposed confidence intervals depend on category specific thresholds.

V. Observations and Future Research Directions

1) There is a tradeoff between precision and recall, i.e. if more optimizations can be done on precision, recall gets lower, oppositely if recall can be improved, precision value becomes lower. So, this research recommends to balance at the point of fixing threshold point for IoU.

2) Design of experiments should cover practical 3D domain adaptation scenarios mentioned below:

a) Adaptation from label rich domains to label insufficient domains,

b) Adaptation across domains with different data collection locations and time (e.g., Waymo → KITTI, nuScenes → KITTI), and

c) Adaptation across domains with a different number of the LiDAR beams (i.e., Waymo → nuScenes and nuScenes → KITTI). Therefore, domain adaptive evaluation needs to be done for validating 3D object detection models on the following four adaptation tasks: Waymo → KITTI, Waymo → Lyft, Waymo → nuScenes and nuScenes → KITTI.

3) Some ill-posed settings that is not suitable for evaluation needs to rule out. For example, KITTI datasets lacks in ring view annotations (less practical) and Lyft uses very different annotation rules (i.e., many objects outside the road are not annotated).

4) Comparing with stereo images, monocular images lose multiview visual characteristic and spatial structure characteristic causes significant information loss demands for robust depth construction for 3D object detection which can lead to 3D localization.

5) Due to the significant visual gap between object-centric images (various texture with complex background) and multiview images of 3-D models (gray model appearance with clean background), monocular and stereo image-based 3D object detection toward localization or tacking domain adaptation is a challenging task requires further investigation. In lieu of current existing datasets, development of novel dataset for object centric monocular and stereo image based is required to advocate the use of 3D object detection towards 3D localization for real world applications. In this context, few possible query images should be the primary key which can be the significant implication for domain adaptation for overall 3D object detection pipeline.

6) To choose the appropriate backbone network depending on problem specific 3D object detection, appropriate fusion strategies need to be designed in lieu of robust loss function to ensure efficient similarity measurement for final classification tasks of 3D object detection.

VI. Conclusion

3D object detection is the basis of many autonomous intelligent applications. This research demonstrates comprehensive and critical reviews on existing 3D object detection methods using RGB images and other fusion based detection methodology based on LiDAR and Pseudo-LiDAR. Some existing methods detected objects with 2D bounding box to recognize position of the objects which is not sufficient for perfect autonomous system. Therefore, predicting 3D object's position is similarly important as determining the 2D position of object in the image. In this research, sensor modality for the overall review is categorized in four types, i.e., monocular image, stereo image, point clouds obtained from LiDAR and Pseudo-LiDAR and fusion of both where advantages and disadvantages were addressed for each type. Depth summary with relative challenges for eight datasets are critically highlighted by this research. In this context, KITTI benchmark are not suitable for monocular methods for 3D object detection due to lack of depth information and prevents accurate 3D positioning which encourages to use maximum number of datasets to ensure robustness for any 3D object detection method. Besides, comprehensive details for eight evaluation metrics are illustrated to evaluate 3D object detection methods. This research observed that 3D object detection is not matured as 2D object due to large gap existing between them. Existing methods still did not achieve the benchmark performance for real time autonomous applications initiates the need for fast and reliable 3D object detection system for wide range of real time applications. Besides, recent trend for using point cloud processing was observed by this research provides effective solution for 3D object detection but LiDAR is an expensive sensor and further geometrical relationship needs to be discovered among points. Besides, some fusion based methods, i.e., RGB images either with LiDAR point cloud or depth images from RGB-D data could not confirm their superiority than other methods in multimodal datasets to ensure robust validation which indicates that more focus is needed to develop multimodal methods for 3D object detection. In addition, lack of large scale annoted training data, more datasets and fusion methods are expected in near future for indoor and outdoor scenarios to form a unified 3D object detection framework. From this study, this research remarks that 3D object detection has

gained many successes, but remains as potential and fertile research problem which requires more exploration. Demonstrated critical review by this research is expected to serve as a supportive significant reference and forms an important endorsement to the related research community.

ACKNOWLEDGMENTS

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

REFERENCES

[1] X. Chen, K. Kundu, Y. Zhu, H. Ma, S. Fidler, R. Urtasun, 3d object proposals using stereo imagery for accurate object class detection, IEEE transactions on pattern analysis and machine intelligence, 40 (2017) 1259-1272.

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

[3] X. Chen, H. Ma, J. Wan, B. Li, 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.

[4] P. Li, X. Chen, 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.

[5] Y. Wang, W.-L. Chao, D. Garg, B. Hariharan, M. Campbell, 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.

[6] Z. Liu, D. Zhou, F. Lu, J. Fang, L. Zhang, Autoshape: Real-time shape-aware monocular 3d object detection, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 15641-15650.

[7] S. Shi, X. Wang, H. Li, Pointrcnn: 3d object proposal generation and detection from point cloud, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2019, pp. 770-779.

[8] H. Sheng, S. Cai, Y. Liu, B. Deng, J. Huang, X.-S. Hua, M.-J. Zhao, Improving 3d object detection with channel-wise transformer, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 2743-2752.

[9] T. Roddick, A. Kendall, R. Cipolla, Orthographic feature transform for monocular 3d object detection, arXiv preprint arXiv:1811.08188, (2018).

[10] L. Bertoni, S. Kreiss, A. Alahi, Monoloco: Monocular 3d pedestrian localization and uncertainty estimation, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 6861-6871.

[11] A.S. Saif, Z.R. Mahayuddin, Moment Features based Violence Action Detection using Optical Flow, International Journal of Advanced Computer Science and Applications, 11 (2020).

[12] A.S. Saif, Z.R. Mahayuddin, Robust Drowsiness Detection for Vehicle Driver using Deep Convolutional Neural Network, International Journal of Advanced Computer Science and Applications, 11 (2020).

[13] C.R. Qi, W. Liu, C. Wu, H. Su, 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.

[14] X. Pan, Z. Xia, S. Song, L.E. Li, G. Huang, 3d object detection with pointformer, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 7463-7472.

[15] L. Wang, L. Zhang, Y. Zhu, Z. Zhang, T. He, M. Li, X. Xue, Progressive coordinate transforms for monocular 3d object detection, Advances in Neural Information Processing Systems, 34 (2021).

[16] D. Song, W.-Z. Nie, W.-H. Li, M. Kankanhalli, A.-A. Liu, Monocular Image-Based 3-D Model Retrieval: A Benchmark, IEEE Transactions on Cybernetics, (2021).

[17] Y. Chen, S. Liu, X. Shen, 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. 12536-12545.

[18] Y. Wang, B. Yang, R. Hu, M. Liang, R. Urtasun, PLUMENet: Efficient 3D Object Detection from Stereo Images, in: 2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), IEEE, pp. 3383-3390.

[19] D. Rukhovich, A. Vorontsova, A. Konushin, 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, 2022, pp. 2397-2406.

[20] C. Li, J. Ku, S.L. Waslander, Confidence guided stereo 3D object detection with split depth estimation, in: 2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), IEEE, 2020, pp. 5776-5783.

[21] A.S. Saif, Z.R. Mahayuddin, H. Arshad, Vision-Based Efficient Collision Avoidance Model Using Distance Measurement, in: Soft Computing Approach for Mathematical Modeling of Engineering Problems, CRC Press, 2021, pp. 191-202.

[22] A.S. Saif, Z.R. Mahayuddin, Edge Feature based Moving Object Detection Using Aerial Images: A Comparative Study, in: Edge Feature based Moving Object Detection Using Aerial Images: A Comparative Study, IEEE Press, 2021.

[23] A.H. Lang, S. Vora, H. Caesar, L. Zhou, J. Yang, 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. 12697-12705.

[24] Y. Zhou, 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.

[25] S. Shi, Z. Wang, J. Shi, X. Wang, H. Li, From points to parts: 3d object detection from point cloud with part-aware and part-aggregation network, IEEE transactions on pattern analysis and machine intelligence, (2020).

[26] D. Feng, S. Han, H. Xu, X. Liang, X. Tan, Point-Guided Contrastive Learning for Monocular 3-D Object Detection, IEEE Transactions on Cybernetics, (2021).

[27] Z. Liu, D. Zhou, F. Lu, J. Fang, L. Zhang, Autoshape: Real-time shape-aware monocular 3d object detection, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 15641-15650.

[28] R. Qian, D. Garg, Y. Wang, Y. You, S. Belongie, B. Hariharan, M. Campbell, K.Q. Weinberger, 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.

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

[30] L. Chen, J. Sun, Y. Xie, S. Zhang, Q. Shuai, Q. Jiang, G. Zhang, H. Bao, X. Zhou, Shape Prior Guided Instance Disparity Estimation for 3D Object Detection, IEEE Transactions on Pattern Analysis and Machine Intelligence, (2021).

[31] F. Negahbani, O.B. Töre, F. Güney, B. Akgun, Frustum Fusion: Pseudo-LiDAR and LiDAR Fusion for 3D Detection, arXiv preprint arXiv:2111.04780, (2021).

[32] Y. Wang, B. Yang, R. Hu, M. Liang, R. Urtasun, PLUMENet: Efficient 3D Object Detection from Stereo Images, in: 2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), IEEE, pp. 3383-3390.

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

[34] J. Ku, M. Mozifian, J. Lee, A. Harakeh, 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.

[35] L. Peng, F. Liu, S. Yan, X. He, D. Cai, OCM3D: Object-Centric Monocular 3D Object Detection, arXiv preprint arXiv:2104.06041, (2021).

[36] S. Wirges, T. Fischer, C. Stiller, J.B. Frias, Object detection and classification in occupancy grid maps using deep convolutional networks, in: 2018 21st International Conference on Intelligent Transportation Systems (ITSC), IEEE, 2018, pp. 3530-3535.

[37] X. Du, M.H. Ang, S. Karaman, D. Rus, A general pipeline for 3d detection of vehicles, in: 2018 IEEE International Conference on Robotics and Automation (ICRA), IEEE, 2018, pp. 3194-3200.

[38] K. Minemura, H. Liau, A. Monrroy, S. Kato, LMNet: Real-time multiclass object detection on CPU using 3D LiDAR, in: 2018 3rd Asia-Pacific Conference on Intelligent Robot Systems (ACIRS), IEEE, 2018, pp. 28-34.

[39] J. Beltrán, C. Guindel, F.M. Moreno, D. Cruzado, F. Garcia, A. De La Escalera, Birdnet: a 3d object detection framework from lidar information, in: 2018 21st International Conference on Intelligent Transportation Systems (ITSC), IEEE, 2018, pp. 3517-3523.

[40] J. Qiu, Z. Cui, Y. Zhang, X. Zhang, S. Liu, B. Zeng, M. Pollefeys, Deeplidar: Deep surface normal guided depth prediction for outdoor scene from sparse lidar data and single color image, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 3313-3322.

[41] C. Godard, O. Mac Aodha, G.J. Brostow, Unsupervised monocular depth estimation with left-right consistency, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 270-279.

[42] C. Reading, A. Harakeh, J. Chae, 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.

[43] Y. Liu, Y. Yixuan, M. Liu, Ground-aware monocular 3d object detection for autonomous driving, IEEE Robotics and Automation Letters, 6 (2021) 919-926.

[44] J. Redmon, A. Farhadi, Yolov3: An incremental improvement, arXiv preprint arXiv:1804.02767, (2018).

[45] A.S. Saif, Z.R. Mahayuddin, Vision based 3D Gesture Tracking using Augmented Reality and Virtual Reality for Improved Learning Applications, International Journal of Advanced Computer Science and Applications, 12 (2022) 631-638.

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

[47] A. Simonelli, S.R. Bulo, L. Porzi, M. López-Antequera, P. Kontschieder, Disentangling monocular 3d object detection, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 1991-1999.

[48] A. Simonelli, S.R. Bulo, L. Porzi, M.L. Antequera, P. Kontschieder, Disentangling monocular 3d object detection: From single to multi-class recognition, IEEE Transactions on Pattern Analysis and Machine Intelligence, (2020).

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

[50] S.R. Bulo, L. Porzi, P. Kontschieder, In-place activated batchnorm for memory-optimized training of dnns, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 5639-5647.

[51] C. Chen, L.Z. Fragonara, A. Tsourdos, RoIFusion: 3D Object Detection From LiDAR and Vision, IEEE Access, 9 (2021) 51710-51721.

[52] A. Paigwar, D. Sierra-Gonzalez, Ö. Erkent, C. Laugier, Frustum-pointpillars: A multi-stage approach for 3d object detection using rgb camera and lidar, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 2926-2933.

[53] J. Yang, S. Shi, Z. Wang, H. Li, X. Qi, ST3D: Self-training for Unsupervised Domain Adaptation on 3D Object Detection, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 10368-10378.

[54] A. Kumar, G. Brazil, X. Liu, GrooMeD-NMS: Grouped Mathematically Differentiable NMS for Monocular 3D Object Detection, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 8973-8983.

[55] J. Noh, S. Lee, B. Ham, HVPR: Hybrid Voxel-Point Representation for Single-stage 3D Object Detection, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 14605-14614.

[56] L. Peng, F. Liu, S. Yan, X. He, D. Cai, OCM3D: Object-Centric Monocular 3D Object Detection, arXiv preprint arXiv:2104.06041, (2021).

[57] J. Yang, S. Shi, Z. Wang, H. Li, X. Qi, ST3D++: Denoised Self-training for Unsupervised Domain Adaptation on 3D Object Detection, arXiv preprint arXiv:2108.06682, (2021).

[58] A. Geiger, P. Lenz, 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.

[59] T. Guan, J. Wang, S. Lan, R. Chandra, Z. Wu, L. Davis, D. Manocha, M3detr: Multi-representation, multi-scale, mutual-relation 3d object detection with transformers, in: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2022, pp. 772-782.

[60] X. Ma, Y. Zhang, D. Xu, D. Zhou, S. Yi, H. Li, W. Ouyang, Delving into Localization Errors for Monocular 3D Object Detection, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 4721-4730.

[61] Y. Zhang, X. Ma, S. Yi, J. Hou, Z. Wang, W. Ouyang, D. Xu, Learning Geometry-Guided Depth via Projective Modeling for Monocular 3D Object Detection, arXiv preprint arXiv:2107.13931, (2021).

[62] A. Sagar, Aa3dnet: Attention augmented real time 3d object detection, in: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2022, pp. 628-635.

[63] Y. Lu, X. Ma, L. Yang, T. Zhang, Y. Liu, Q. Chu, J. Yan, 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.

[64] J. Li, Y. Sun, S. Luo, Z. Zhu, H. Dai, A.S. Krylov, Y. Ding, L. Shao, P2V-RCNN: Point to Voxel Feature Learning for 3D Object Detection from Point Clouds, IEEE Access, 9 (2021) 98249-98260.

[65] G. Brazil, G. Pons-Moll, X. Liu, B. Schiele, Kinematic 3d object detection in monocular video, in: European Conference on Computer Vision, Springer, 2020, pp. 135-152.

[66] T. Jiang, N. Song, H. Liu, R. Yin, Y. Gong, J. Yao, VIC-Net: Voxelization Information Compensation Network for Point Cloud 3D Object Detection, in: 2021 IEEE International Conference on Robotics and Automation (ICRA), IEEE, 2021, pp. 13408-13414.

[67] Y. Shi, Y. Guo, Z. Mi, X. Li, Stereo CenterNet-based 3D object detection for autonomous driving, Neurocomputing, 471 (2022) 219-229.

[68] C. Reading, A. Harakeh, J. Chae, 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] Y. Zhou, Y. He, H. Zhu, C. Wang, H. Li, Q. Jiang, Monocular 3D Object Detection: An Extrinsic Parameter Free Approach, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 7556-7566.

[70] S. Pang, D. Morris, H. Radha, Fast-CLOCs: Fast Camera-LiDAR Object Candidates Fusion for 3D Object Detection, in: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2022, pp. 187-196.

[71] P. Sun, H. Kretzschmar, X. Dotiwalla, A. Chouard, V. Patnaik, P. Tsui, J. Guo, Y. Zhou, Y. Chai, B. Caine, 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.

[72] P. Sun, H. Kretzschmar, X. Dotiwalla, A. Chouard, V. Patnaik, P. Tsui, J. Guo, Y. Zhou, Y. Chai, B. Caine, 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.

[73] X. Chen, H. Ma, C. Zhu, X. Wang, Z. Zhao, Boundary-aware box refinement for object proposal generation, Neurocomputing, 219 (2017) 323-332.

[74] R. Girshick, J. Donahue, T. Darrell, J. Malik, Rich feature hierarchies for accurate object detection and semantic segmentation, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2014, pp. 580-587.

[75] X. Chen, K. Kundu, Z. Zhang, H. Ma, S. Fidler, 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] J. Hosang, R. Benenson, P. Dollár, B. Schiele, What makes for effective detection proposals?, IEEE transactions on pattern analysis and machine intelligence, 38 (2015) 814-830.

[77] F. Yu, D. Wang, E. Shelhamer, T. Darrell, Deep layer aggregation, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 2403-2412.

[78] R. Qian, X. Lai, X. Li, Boundary-Aware 3D Object Detection from Point Clouds, arXiv preprint arXiv:2104.10330, (2021).

[79] Z. Miao, J. Chen, H. Pan, R. Zhang, K. Liu, P. Hao, J. Zhu, Y. Wang, X. Zhan, PVGNet: A Bottom-Up One-Stage 3D Object Detector With Integrated Multi-Level Features, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 3279-3288.

[80] X. Liu, N. Xue, T. Wu, Learning Auxiliary Monocular Contexts Helps Monocular 3D Object Detection, arXiv preprint arXiv:2112.04628, (2021).

[81] G. Brazil, G. Pons-Moll, X. Liu, B. Schiele, Kinematic 3d object detection in monocular video, in: European Conference on Computer Vision, Springer, 2020, pp. 135-152.

[82] Q. Lian, B. Ye, R. Xu, W. Yao, T. Zhang, Geometry-aware data augmentation for monocular 3D object detection, arXiv preprint arXiv:2104.05858, (2021).

[83] H. Wang, Y. Cong, O. Litany, Y. Gao, L.J. Guibas, 3DIoUMatch: Leveraging iou prediction for semi-supervised 3d object detection, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 14615-14624.

[84] Z. Zou, X. Ye, L. Du, X. Cheng, X. Tan, L. Zhang, J. Feng, X. Xue, E. Ding, The devil is in the task: Exploiting reciprocal appearance-localization features for monocular 3d object detection, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 2713-2722.

[85] C.-H. Wang, H.-W. Chen, L.-C. Fu, VPFNet: Voxel-Pixel Fusion Network for Multi-class 3D Object Detection, arXiv preprint arXiv:2111.00966, (2021).

[86] Z. Liu, D. Zhou, F. Lu, J. Fang, L. Zhang, Autoshape: Real-time shape-aware monocular 3d object detection, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 15641-15650.

[87] S. Shi, X. Wang, H. Li, Pointrcnn: 3d object proposal generation and detection from point cloud, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2019, pp. 770-779.

[88] Y. Chen, S. Liu, X. Shen, J. Jia, Fast point r-cnn, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 9775-9784.

[89] Z. Wang, K. Jia, Frustum convnet: Sliding frustums to aggregate local point-wise features for amodal 3d object detection, in: 2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), IEEE, 2019, pp. 1742-1749.

[90] S. Shi, C. Guo, L. Jiang, Z. Wang, J. Shi, X. Wang, H. Li, Pv-rcnn: Point-voxel feature set abstraction for 3d object detection, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 10529-10538.

[91] R. Qian, X. Lai, X. Li, Boundary-Aware 3D Object Detection from Point Clouds, arXiv preprint arXiv:2104.10330, (2021).

[92] Y. Zhang, J. Lu, 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.

[93] Z. Miao, J. Chen, H. Pan, R. Zhang, K. Liu, P. Hao, J. Zhu, Y. Wang, X. Zhan, PVGNet: A Bottom-Up One-Stage 3D Object Detector With Integrated Multi-Level Features, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 3279-3288.

[94] X. Liu, N. Xue, T. Wu, Learning Auxiliary Monocular Contexts Helps Monocular 3D Object Detection, arXiv preprint arXiv:2112.04628, (2021).

[95] J. Lei, T. Guo, B. Peng, C. Yu, Depth-Assisted Joint Detection Network For Monocular 3d Object Detection, in: 2021 IEEE International Conference on Image Processing (ICIP), IEEE, 2021, pp. 2204-2208.

[96] Y. Li, S. Yang, Y. Zheng, H. Lu, Improved Point-Voxel Region Convolutional Neural Network: 3D Object Detectors for Autonomous Driving, IEEE Transactions on Intelligent Transportation Systems, (2021).

[97] S. Shi, L. Jiang, J. Deng, Z. Wang, C. Guo, J. Shi, X. Wang, H. Li, PV-RCNN++: Point-Voxel Feature Set Abstraction With Local Vector Representation for 3D Object Detection, arXiv preprint arXiv:2102.00463, (2021).

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

[99] Q. Lian, B. Ye, R. Xu, W. Yao, T. Zhang, Geometry-aware data augmentation for monocular 3D object detection, arXiv preprint arXiv:2104.05858, (2021).

[100] S.R. Bulo, L. Porzi, P. Kontschieder, In-place activated batchnorm for memory-optimized training of dnns, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 5639-5647.

[101] X. Chen, H. Ma, Learning a compact latent representation of the bag-of-parts model, in: 2014 IEEE International Conference on Image Processing (ICIP), IEEE, 2014, pp. 5926-5930.

[102] N.A.M. Mai, P. Duthon, L. Khoudour, A. Crouzil, S.A. Velastin, Sparse LiDAR and Stereo Fusion (SLS-Fusion) for Depth Estimationand 3D Object Detection, arXiv preprint arXiv:2103.03977, (2021).

[103] A. Geiger, P. Lenz, 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.

[104] Y. Xiang, W. Choi, Y. Lin, 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.

[105] F. Yu, D. Wang, E. Shelhamer, T. Darrell, Deep layer aggregation, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 2403-2412.

[106] P. Li, H. Zhao, P. Liu, F. Cao, Rtm3d: Real-time monocular 3d detection from object keypoints for autonomous driving, in: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part III 16, Springer, 2020, pp. 644-660.