A survey of deep learning models for camera based 3D object detection in autonomous driving

Anonymous Anonymous
Anonymous

Abstract

The majority of road accidents occur because of human errors, including distraction, recklessness, and drunken driving. One of the effective ways to overcome this dangerous situation is by implementing self-driving technologies in vehicles. Generally, autonomous vehicles are equipped with various sensors such as LiDARs, Cameras, and Radars to collect the data that assist the vehicle in terms of perception, localization, path planning, and motion control. 3D object detection is one of the key elements in the autonomous driving pipeline for scene perception. On the basis of input data format, 3D object detection methods can be classified into Camera-based, LiDAR-based, and Multi-modal based. In all these methods, deep learning technologies are employed for predicting the objects, their sizes, and locations in the driving scenes. LiDAR and Multi-modal based methods are continuously outperforming the camera-based methods, but LiDARs are more expensive. They not only increase the cost of the vehicle but also the computational cost of processing point clouds during object detection. Due to this reason, research on 3D object detection using 2D images from cameras is in place, and many papers have been published over the last decade. To give a proper direction for future research work, a comprehensive survey of the deep learning methods related to camera-based 3D object detection is the need of the hour. Though several survey papers on 3D object detection were already published, they did not focus on an in-depth analysis of the camera-based methods. In addition to this, several state-of-the-art methods were published in the last two years but were not discussed in any of the previous survey papers. This paper aims to present a comprehensive survey of deep learning methods employed for 3D object detection using camera images with a main focus on recently published works. It also presents the research gaps and will serve as starting point for future research in this area.

1. Introduction

3D object detection is one of the important components in autonomous driving pipeline to understand the driving surroundings of an ego car. Several methods are available to perform this activity using deep learning techniques as mentioned in Figure 1. We know that detecting the 3D objects using just images from the cameras is less expensive, faster in terms of computations and perform well in harsh weather conditions such as snow and rain. However, camera based methods are still in infancy in terms of performance when compared to the LiDAR and fusion-based methods. Due to this reason, a lot of research is currently in place to improve the performance. Several researchers contributed to this by proposing novel methods. But there is no comprehensive survey detailing these methods. Though several survey papers on deep learning techniques for autonomous driving were proposed, they did not provide an in-depth analysis of the camera based 3D object detection methods.

Tho high-level objectives of this paper are as follows. Firstly, we analyse the recently proposed sate-of-the-art (SOTA) deep learning models in terms of their performance on KITTI [32], NuScenes [7], and Waymo [122] datasets and compare with the old methods. Secondly, we discuss about the research gaps to be addressed to improve the performance on par with the LiDAR and the fusion-based methods.

The remainder of the paper is structured as follows. In section **??** we propose new taxonomy for camera-based 3D object detection methods. In section 3, we provide the previous survey papers on deep learning techniques for autonomous driving. In section 4, we provide architecture for various 3D object detection methods. In section 5, we propose our taxonomy for 3D object detection methods, discuss about the Monocular, Stereo, Multiple and RGB-D camera based methods. In section 6, we provide a list of publicly available datsets for 3D object detection. In section 7, we discuss about the performance analysis of various methods and evaluate them on key scores on the popular datasets. In section 8, we provide opportunities for future work. Finally, concluding remarks are mentioned in sec-

tion 9.

2. Taxonomy

We propose the taxonomy for the camera-based 3D object detection methods as shown in Figure. 1. In this paper, we focus only on camera-based methods due to their cost benefit and also camera sensors are robust in harsh weather conditions unlike LiDARs. Based on the number and type and the number of cameras, we have classified them into Single camera-based, Stereo camera-based, Multicamera-based and RGB-D camera-based methods. In all these four methods, several approaches are followed while applying deep learning techniques for 3D object detection. They are: Keypoint-based, Anchor-based, Anchor-free-based, Geometry-based, Bird's-Eye-View(BEV) based, Transformer-based, Pseudo-LiDAR-based, 3D representation, and 2D representation approaches. This type of taxonomy was not proposed in any of the earlier survey papers. We will discuss various unique and recent contributions to 3D object detection from this perspective.

3. Previous survey papers published on deep learning methods of autonomous driving

Grigorescu et al. [35] proposed a paper to survey the state-of-the-art deep learning technologies as on 2020 and the challenges encountered in designing AI architectures for autonomous driving. Li et al. [64] proposed a systematic review of existing compelling deep learning architectures applied in LiDAR point clouds, detailing for specific tasks in autonomous driving such as segmentation, detection, and classification. Yu Huang and Yue Chen [47] proposed a survey of autonomous driving technologies with deep learning methods in several key areas such as 2D/3D object detection in perception, depth estimation from cameras, multiple sensor fusion on the data, feature and task level respectively, behaviour modelling and prediction of vehicle driving and pedestrian trajectories. Papadeas et al. [95] proposed a comprehensive overview of the state-of-the-art semantic image segmentation methods using deep learning techniques aiming to operate in real time so that can efficiently support an autonomous driving scenario. Cui et al. [21] proposed to review recent deep-learning-based data fusion approaches that leverage both image and point cloud and compared these methods on publicly available datasets. Wen et al. [136] proposed a survey of deep learning methods for object detection tasks in autonomous driving using the data from LiDAR and cameras. Zamanakos et al. [154] proposed a comprehensive survey of LIDAR-based 3D object detection methods wherein an analysis of existing methods was addressed by taking into account a new categorisation that relies upon a common operational pipeline which describes the end-to-end functionality of each method. Guo et al. [37] proposed a comprehensive survey of deep learning-based approaches for scene understanding in autonomous driving across four work streams, including object detection, full scene semantic segmentation, instance segmentation, and lane line segmentation. Alaba et al. [1] proposed a survey to present the LiDAR based 3D object detection and feature-extraction techniques for LiDAR data and then reviewed state-of-the-art methods with a selected comparison. Liu et al. [74] proposed a survey to review and categorize existing learning-based trajectory forecasting methods from perspectives of representation, modeling, and learning. Feng et al. [28] proposed a survey to review and compare existing probabilistic object detection methods for autonomous driving applications based on an image detector and public autonomous driving datasets. Ni et al. [91] proposed a review to present a review of research on theories and applications of deep learning for self-driving cars and provided a detailed explanation of the developments and summarized the applications of deep learning methods.

Mozaffari et al. [90] proposed a comprehensive review of the state-of-the-art of deep learning based approaches for vehicle behaviour prediction based on three criteria: input representation, output type, and prediction method. Kuutti et al. [57] proposed a paper to survey a wide range of research works reported in the literature which aim to control a vehicle through deep learning methods and identified the strength and limitations of each method. Paravarzar et al. [96] proposed a review of the recent deep learning and reinforcement learning methods adopted to predict the behaviour of the self-driving vehicles and made a comparison between these two types of methods.

Kiran et al. [53] proposed a paper to summarise deep reinforcement learning (DRL) algorithms and to provides a taxonomy of automated driving tasks where (D)RL methods have been employed, while addressing key computational challenges in real world deployment of autonomous driving agents.

Deng et al. [23] proposed a thorough analysis of different attacks that may jeopardize Autonomnous driving systems(ADSs) designed based on deep learning technologies, as well as the corresponding state-of-the-art defense mechanisms. The analysis presented an in-depth overview of each step in the ADS workflow, covering adversarial attacks for various deep learning models and attacks in both physical and cyber context. Cao et al. [9] proposed a summary of the concepts, developments and recent research in deep learning security technologies in autonomous driving. They focused on the potential security threats of the deep learning based autonomous driving system.

Wang et al. [127] proposed a survey of multi-modal fusion methods for 3D object detection based on the categories such as feature representation, alignment, and fusion.

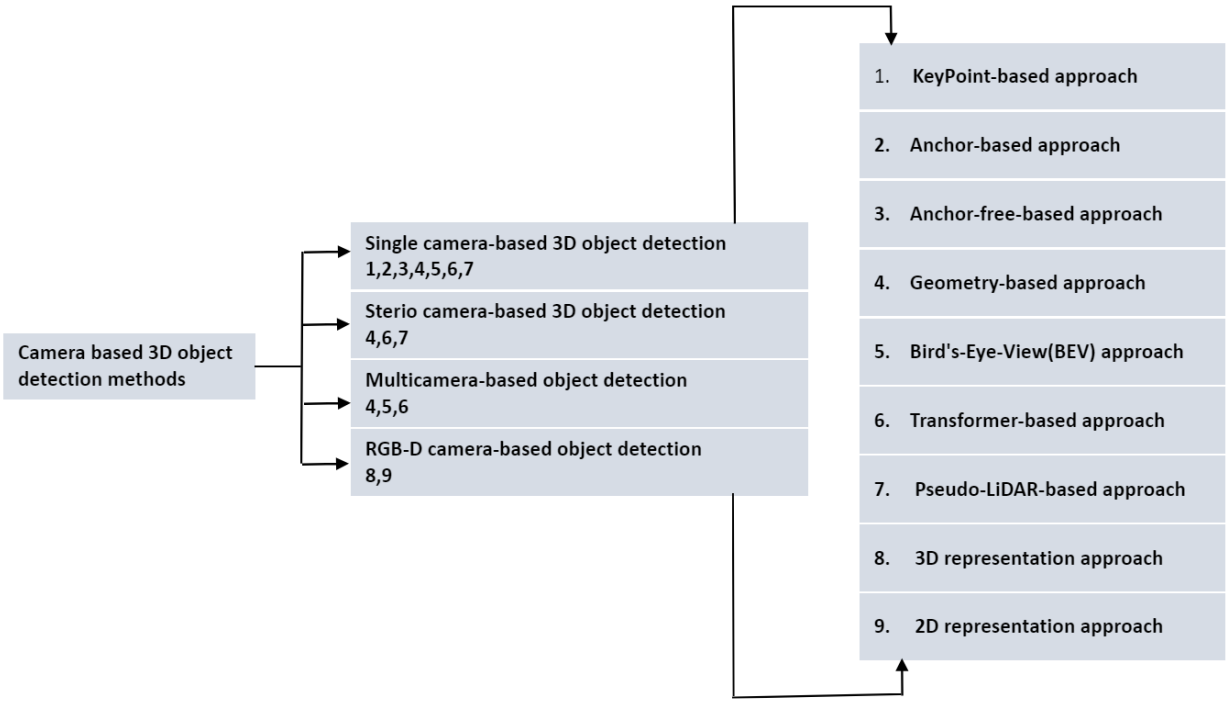

Figure 1. Taxonomy of camera-based 3D object detection for autonomous driving

Liang et al. [70] proposed a survey of 3D object detection methods based on the input data format such as LiDAR point cloud-based, Camera RGB image-based, and LiDAR point cloud-camera image fusion-based. Yao et al. [146] proposed a survey of Radar-Camera fusion methods based on categories such as "why to fuse", "what to fuse", "where to fuse", "when to fuse", and "how to fuse".

The above survey papers provided a comprehensive survey of deep learning methods applied for autonomous driving as a whole in general and for 3D object detection in particular. They gave more importance to LiDAR-based and Fusion-based methods as these methods are outperforming the Camera-based methods. But the major disadvantages of using LiDARs are very expensive and require high computational power to process point clouds. In this paper, we focus on camera-based methods for 3D object detection as there is a need to improve their performance because of the cost advantage associated with using cameras. For this reason, a continuous research is taking place but there is no comprehensive survey of the recently proposed SOTA models.

4. Methods

In this section, we discuss about the unique architectures of recently published papers in each of the four camera-based 3D object detection methods.

The base architecture [144] of a keypoint-based monocular 3D object detector is shown in the Figure 2. The whole framework is divided into four parts: (1) backbone, (2) detection head, (3) post process, (4) loss function. The backbone network is composed of an encoder and a decoder. The encoder extracts high dimensional features from a RGB image with residual networks (ResNet) [39] or deep layer aggregation (DLA-34) [152]. The decoder upsamples the bottleneck features to 1/4 times with respect to the input image by three deconvolutional layers. The detection head is formed by fully convolutional layers. The post process contains two sequence processes: a top-K process and a decode process to obtain 3D bounding boxes. The total loss function is made up of two parts: a keypoint classification loss and a regression loss.

The entire pipeline for monocular 3D object detection is shown in Figure. 3. Given a single RGB image, a depth estimation map and a semantic segmentation map are respectively generated by two separate modules. Then the output maps are processed by the depth correction module for pseudo-LiDAR point cloud. A 3D object detector is finally used to provide 3D estimation.

The network architecture of a Stereo CenterNet [118] that outputs 10 sub-branches for two tasks and the estimated 3D bounding box is shown in the Figure. 4. The overall network, which is built on CenterNet [163], uses a weight-share backbone network to extract consistent features on left and right images architecture. The network out-

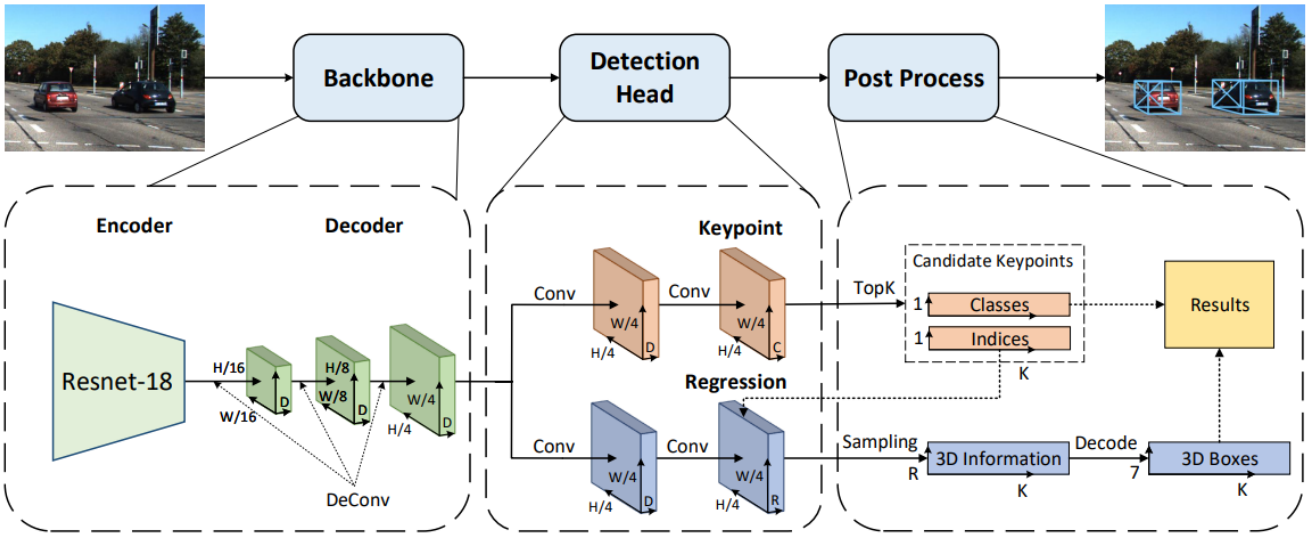

Figure 2. The base architecture of a keypoint-based monocular 3D object detector [144]

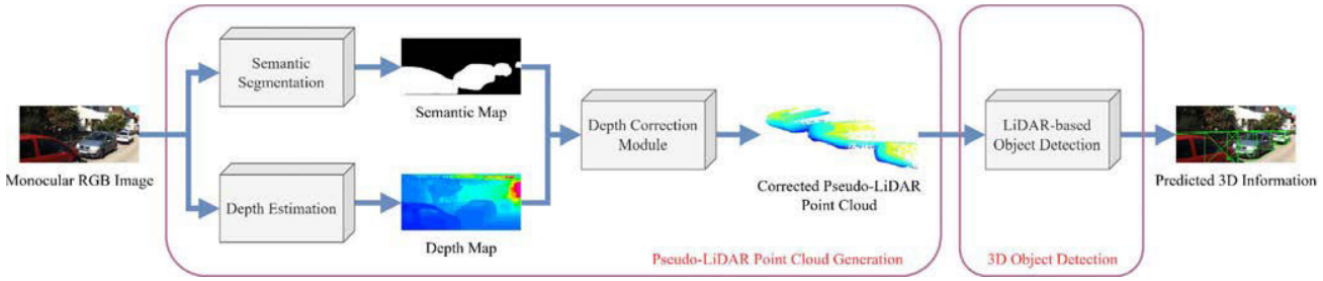

Figure 3. Pipeline for monocular 3D object detection using Pseudo-LiDAR point cloud method [130]

puts 10 sub-branches behind the backbone network to complete two tasks: (A) stereo 2D detection and (B) stereo 3D Components. In task A, we used five sub-branches to estimate the left objects center, left objects center offset, left objects bounding box size, left and right center distances, and the width of the right objects box. The five sub-branches in task B estimated the orientation, dimension, bottom vertices, bottom vertices offset, vertices, and left objects center distance of the 3D bounding box.

A proposal of Pseudo-LiDAR pipeline for stereo-based 3D object detection is shown in Figure 5. The left image is used to generate a semantic map and optional bounding box suggestions, together with the right image disparities are calculated. These are clustered and projected into a grid map with elevation information, which is then used to estimate the 3D bounding boxes.

The overall framework of a Simple baseline for Multi-camera 3D object detection(SimMOD) [159] is shown in Figure. 6. With the surrounding images as input, SimMOD first extracts multi-scale feature maps with the image encoder, which consists of the backbone and the feature pyramid network. Next, the proposal head processes each feature map to generate the object proposals, including the features, positions, and encodings. Finally, the multi-view and multi-scale proposals are collected in the ego-car coordinates and iteratively refined. The set-based detection loss is applied for end-to-end predictions.

Overall architecture with a RGB-D (a pair of image and point cloud) input is shown in Figure 7. (1) Frustum point clouds X are extracted from the input point cloud and 2D object detection boxes on the image. (2) fseg takes X as input and outputs class-agnostic segmentations used to mask X. (3) fbox predicts an initial 3D box Bˆ0 with the masked point cloud. (4) The pretrained fboxpc model refines Bˆ0 to Bˆ according to X, and predicts the BoxPC fit probability used to supervise fbox.

5. Camera based 3D object detection

5.1. Single camera-based or Monocular 3D object detection

Monocular 3D object detection refers to estimating 3D information such as location, direction, and size of an objects around an AV using a single image as an input. Monocular methods follow various approaches for this task

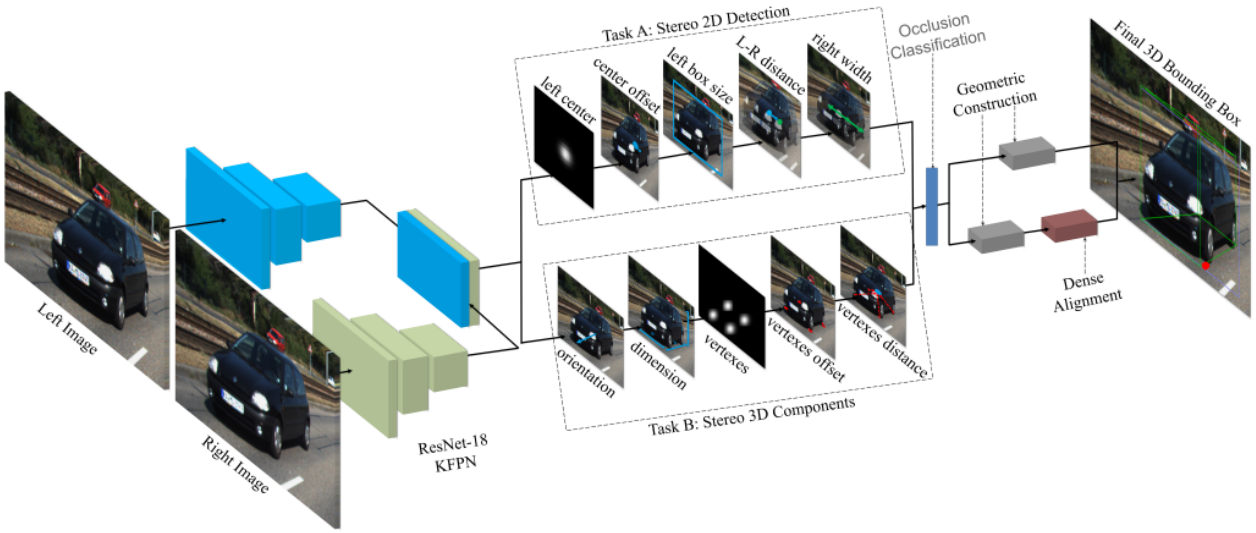

Figure 4. The network architecture of a Stereo CenterNet estimating the 3D bounding box [118]

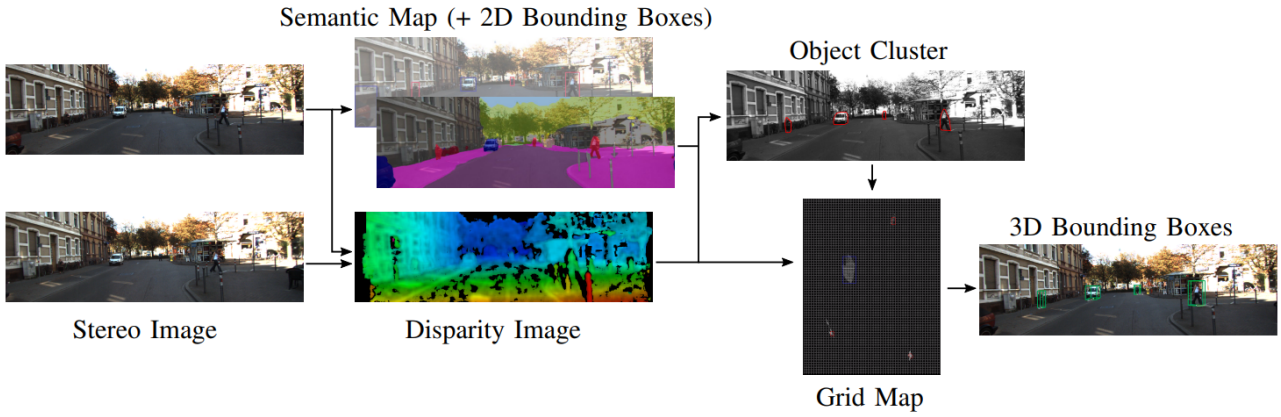

Figure 5. A Pseudo-LiDAR pipeline for stereo-based 3D object detection [54]

as shown in the taxonomy.

5.1.1 Keypoint-based approach

Li et al. [67] proposed an anchor-free and keypoint-based 3D object detector with monocular vision, named Keypoint3D. They leveraged 2D projected points from 3D objects' geometric centers as keypoints for object modeling. Additionally, for precise keypoints positioning, they utilized a novel self-adapting ellipse Gaussian filter (saEGF) on heatmaps, considering different objects' shapes. They tried different variations of DLA-34 backbone and proposed a semi-aggregation DLA-34 (SADLA-34) network, which pruned the redundant aggregation branch but achieved better performance. Yang et al. [144] proposed a lightweight feature pyramid network called Lite-FPN for keypoint-based monocular 3D object detectors that perform multi-scale feature fusion only at sparsely distributed keypoint locations. Besides, to alleviate the misalignment between classification score and localization precision, they proposed an effective regression loss named attention loss, which assigns predictions with misaligned classification score and localization precision larger weights in the training stage. Haq et al. [38] proposed a single stage monocular 3D object detection method that utilizes the discrete depth and orientation representation. Their proposed method predicted object locations on 3D space utilizing keypoint detection on the object's center point. To improve the point detection, they employed center regression on the objects segmentation mask, reducing the detection offset significantly. Ji et al. [48] proposed a method for formulating the 3D object localization as a paired keypoints regression problem. They exploited 2D bounding box priors to predict the projection of paired 3D keypoints on the image plane for each object, and the object localization was recovered via an inverse projection. A fast keypoint regression network was

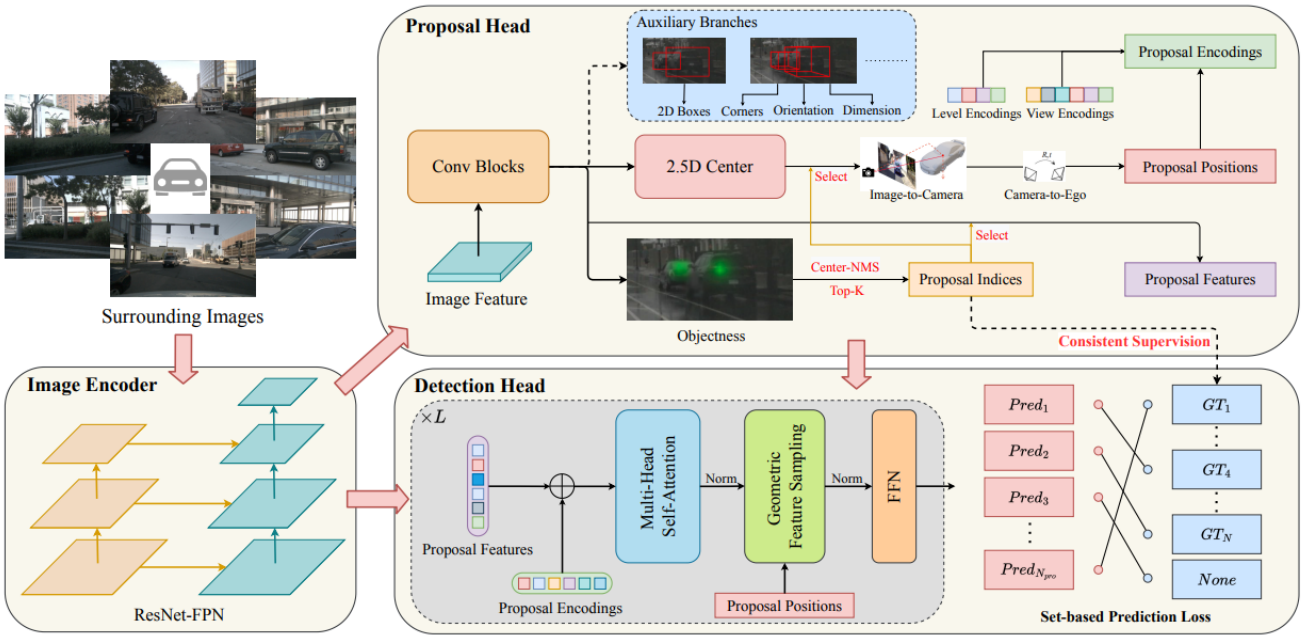

Figure 6. The overall framework of Simple baseline for Multi-camera 3D object detection(SimMOD) for 3D object detection [159]

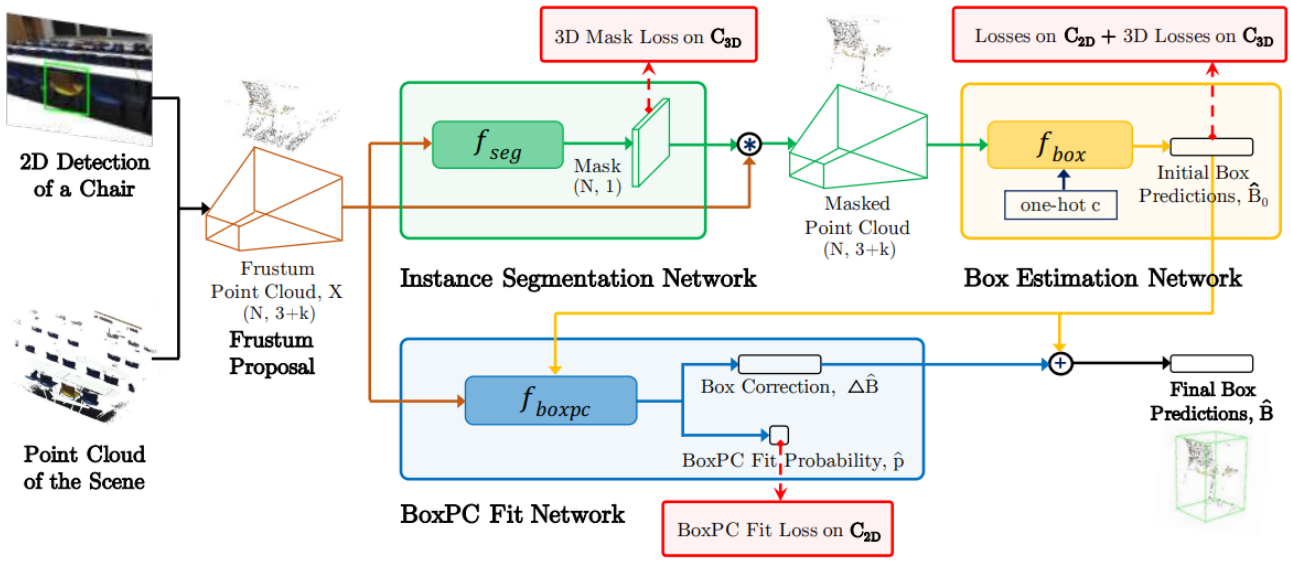

Figure 7. The Overall architecture with a RGB-D 3D representation input for 3D object detection [123]

proposed to predict the projection of keypoints and to generate the initial 3D bounding box.

5.1.2 Anchor-based approach

Anchors are pre-defined cuboids with fixed shapes that can be placed in the 3D space. 3D objects can be predicted based on the positive anchors that have a high intersection over union (IoU) with ground truth.

Liu et al. [76] proposed MonoXiver, a stage-wise approach, which combines the information flow from 2D-to-3D (3D bounding box proposal generation with a single 2D image) and 3D-to-2D (proposal verification by denoising with 3D-to-2D contexts) in a top-down manner. Qu et al. [107] proposed MonoDCN, monocular 3D object detection based on the dynamic convolution network. To make full use of the information in the depth map and RGB image, the backbone was represented as a two-branch network: the first branch is the feature extraction network of RGB images, and the other branch is the filter generation network, generated for dynamic convolutional layers con-

volution kernel. These two networks take RGB image and depth map as input, respectively, and then use the feature map of a feature extraction network as the input of 2D and 3D detection heads to obtain the position information of the detection object, and finally adopt non-maximum suppression and data conversion for visualization.

5.1.3 Anchor-free-based approach

Xie et al. [141] proposed a one-stage monocular 3D object detection network (MDS Net), which uses the anchor-free method to detect 3D objects in a per-pixel prediction. Firstly, a novel depth-based stratification structure was developed to improve the network's ability of depth prediction, which exploits the mathematical relationship between the size and the depth in the image of an object based on the pinhole model. Secondly, a new angle loss function was developed to further improve both the accuracy of the angle prediction and the convergence speed of training. An optimized Soft-NMS is finally applied in the post-processing stage to adjust the confidence score of the candidate boxes. Chen et al. [13] proposed a Geometric Appearance Awareness (GAA) module to improve the estimation of orientation of the 3D object. Specifically, a GAA module was proposed to obtain the geometry-guided appearance feature, which can be used to estimate reliable orientation. Furthermore, they designed a Sample-aware Feature Fusion (SFF) head in the 3D dimension regression branch. This head dynamically deals with the uniqueness of different samples for learning 3D dimension.

5.1.4 Geometry-based approach

Shi et al. [116] proposed multivariate probabilistic modeling framework by explicitly modeling the joint probability distribution of the physical height and visual height. This was achieved by learning a full covariance matrix of the physical height and visual height during training, with the guide of a multivariate likelihood. Zhu et al. [166] proposed a new regression target named keyedge-ratios as the parameterization of the local shape distortion to account for the local perspective, and derive the object depth and yaw angle from it. Theoretically, this approach does not rely on the absolute size or position of the objects in the image, therefore independent of the camera intrinsic parameters.

5.1.5 Bird's-Eye-View(BEV) based approach

Zhang et al. [155] proposed DA-BEV, an implicit depth learning method for Transformer-based camera-only 3D object detection in bird's eye view (BEV). First, a Depth-Aware Spatial Cross-Attention (DA-SCA) module wass proposed to take depth into consideration when querying image features to construct BEV features. Then, to make the BEV feature more depth-aware, they introduced an auxiliary learning task, called Depth-wise Contrastive Learning (DCL), by sampling positive and negative BEV features along each ray that connects an object and a camera. Reading [110] proposed CaDDN, which predicts a categorical depth distribution for each pixel to project feature information in 3D space and then used the bird's-eye-view projection and single-stage detector to produce the final output detections.

5.1.6 Transformer-based approach

Wang et al. [132] proposed a DETR monocular 3D object detection algorithm combining depth and salient information is proposed. A lightweight unsupervised depth module is constructed to extract object depth feature information, and Transformer model was introduced to obtain the global relationship of features. Zhang et al. [156] introduced a novel framework for Monocular DEtection with a depth-guided TRansformer, named MonoDETR. They modified the vanilla transformer to be depth-aware and guided the whole detection process by contextual depth cues. Zhou et al. [165] proposed an online Mono3D framework, called MonoATT, which leverages a novel vision transformer with heterogeneous tokens of varying shapes and sizes to facilitate mobile Mono3D. The core idea of MonoATT is to adaptively assign finer tokens to areas of more significance before utilizing a transformer to enhance Mono3D. Wu et al.[138] proposed MonoPGC, a novel end-to-end Monocular 3D object detection framework with rich Pixel Geometry Contexts. They introduce the pixel depth estimation as our auxiliary task and design depth cross-attention pyramid module (DCPM) to inject local and global depth geometry knowledge into visual features. In addition, they presented the depth-space-aware transformer (DSAT) to integrate 3D space position and depth-aware features efficiently. Huang et al. [44] proposed MonoDTR, a novel end-to-end depth-aware transformer network for monocular 3D object detection. It mainly consists of two components: (1) the Depth-Aware Feature Enhancement (DFE) module that implicitly learns depth-aware features with auxiliary supervision without requiring extra computation, and (2) the Depth-Aware Transformer (DTR) module that globally integrates context- and depth-aware features.

5.1.7 Pseudo-LiDAR based approach

Depth estimation from monocular images in inaccurate but denser pseudo-LiDAR point clouds can be generated. Wang et al. [130] proposed an approach to eliminate the performance degradation caused by deviation of depth estimation and realized 3D object detection based on pseudo-LiDAR point cloud. When tested this model on KITTI 3D benchmark dataset, it achieved more reliable performance on both

loclization and shape estimation. Kim et al. [52] proposed a self-supervised pseudo-LiDAR method for predicting absolute depth and detecting 3D objects using only monocular image sequences by enabling end-to-end learning of detection networks and depth prediction networks.

5.2. Stereo-based 3D object detection

In this method, 3D information is estimated using the left and right images generated by the stereo camera. It provides better depth information than of monocular camera.

5.2.1 Geometry-based approach

Chen et al. [16] proposed Deep Stereo Geometry Network (DSGN) that detects 3D objects on a differentiable volumetric representation – 3D geometric volume, which effectively encodes 3D geometric structure for 3D regular space. With this representation, they learnt depth information and semantic cues simultaneously. Guo et al. [36] proposed LIGAStereo (LiDAR Geometry Aware Stereo Detector) to learn stereo-based 3D detectors under the guidance of high-level geometry-aware representations of LiDAR-based detection models. Wand et al. [129] proposed a method to directly construct a pseudo-LiDAR feature volume (PLUME) in 3D space, which is then used to solve both depth estimation and object detection tasks.

5.2.2 Transformer-based approach

Tao et al. [124] proposed a novel pseudo-monocular 3D object detection framework called Pseudo-Mono. Firstly, stereo images are taken as input, then a lightweight depth predictor is used to generate the depth map of input images. Secondly, the left input images obtained from stereo camera are used as subjects, which generate enhanced visual feature and multi-scale depth feature by depth indexing and feature matching probabilities, respectively. Finally, sparse anchors set by the foreground probability maps and the multi-scale feature maps are used as reference points to find the suitable initialization approach of object query. Shi et al. [118] proposed Stereo CenterNet (SC), using geometric information in stereo imagery. SC predicts the four semantic key points of the 3D bounding box of the object in space and utilizes 2D left and right boxes, 3D dimension, orientation, and key points to restore the bounding box of the object in the 3D space.

5.2.3 Pseudo-LiDAR based approach

Pseudo-LiDAR point cloud are generated from depth image, which is the result of the disparity between two images from the stereo camera. Those methods try to improve the disparity estimation in stereo matching for more accurate depth prediction.

You et al. [151] proposed substantial advances to the pseudo-LiDAR framework through improvements in stereo depth estimation. They adapted the stereo network architecture and loss function to be more aligned with accurate depth estimation of faraway objects — currently the primary weakness of pseudo-LiDAR. They proposed a depth propagation algorithm, guided by the initial depth estimates, to diffuse these few exact measurements across the entire depth map. Qian et al. [104] proposed a new framework based on differentiable Change of Representation (CoR) modules that allow the entire PL pipeline to be trained end-to-end. Königshof et al. [54] proposed a 3D object detection and pose estimation method for automated driving using stereo images. Semantic information was provided by a deep convolutional neural network and used together with disparity and geometric constraints to recover accurate 3D bounding boxes. Li et al. [60] proposed CG-Stereo, a confidence-guided stereo 3D object detection pipeline that uses separate decoders for foreground and background pixels during depth estimation, and leverages the confidence estimation from the depth estimation network as a soft attention mechanism in the 3D object detector. Garg et al. [31] proposed a neural network architecture that is capable of outputting arbitrary depth values, and a new loss function that is derived from the Wasserstein distance between the true and the predicted distributions. We validate our approach on a variety of tasks, including stereo disparity and depth estimation, and the downstream 3D object detection. Le and Nguyen [58] proposed to exploit the simple linear iterative clustering algorithm to segment stereo images into superpixel feature maps. The segmented superpixel maps were then used to estimate a depth map. By utilizing the depth map and stereo images, a 3D point cloud can be generated.

5.3. Multicamera-based 3D object detection

5.3.1 Geometry-based approach

Pham et al. [101] proposed an end-to-end surround camera perception system for self-driving. It is a novel multi-task, multi-camera network which takes a variable set of time-synced camera images as input and produces a rich collection of 3D signals such as sizes, orientations, locations of obstacles, parking spaces and free-spaces, etc.

5.3.2 Bird's-Eye-View(BEV) based approach

Philion and Fidler [102] proposed an end-to-end architecture that directly extracts a bird's-eye-view representation of a scene given image data from an arbitrary number of cameras. The core idea behind our approach is to "lift" each image individually into a frustum of features for each camera, then "splat" all frustums into a rasterized bird's-eyeview grid. This feature map was used to perform downstream 3D

object detection. Huang et al. [42] proposed BEVDet4D to lift the scalable BEVDet paradigm [43] from the spatial-only 3D working space into the spatial-temporal 4D working space. They upgraded the naive BEVDet framework with a few modifications just for fusing the feature from the previous frame with the corresponding one in the current frame. In this way, with negligible additional computing budget, they enabled BEVDet4D to access the temporal cues by querying and comparing the two candidate features. Li et al. [66] proposed a new 3D object detector with a trustworthy depth estimation, dubbed BEVDepth, for camera-based Bird's-Eye-View (BEV) 3D object detection. They leveraged depth supervision and a camera-awareness depth estimation module was also introduced to facilitate the depth predicting capability. Besides, they designed a novel Depth Refinement Module to counter the side effects carried by imprecise feature unprojection. Wnag et al. [135] proposed a Surround-view Temporal Stereo (STS) technique that leverages the geometry correspondence between frames across time to facilitate accurate depth learning. Specifically, they considered the field of views from all cameras around the ego vehicle as a unified view, namely surroundview, and conduct temporal stereo matching on it. The resulting geometrical correspondence between different frames from STS was utilized and combined with the monocular depthto yield final depth prediction. Li et al. [65] proposed BEVSterio, an effective temporal stereo method to dynamically select the scale of matching candidates, enable to significantly reduce computation overhead. Going one step further, they designed an iterative algorithm to update more valuable candidates, making it adaptive to moving candidates. Park et al. [98] proposed to generate a cost volume from a long history of image observations, compensating for the coarse but efficient matching resolution with a more optimal multi-view matching setup. Further, they augmented the per-frame monocular depth predictions used for long-term, coarse matching with short-term, fine-grained matching and found that long and short term temporal fusion were highly complementary. Xie et al. [140] proposed M2BEV, a unified framework that jointly performed 3D object detection and map segmentation in the Bird's Eye View (BEV) space with multi-camera image inputs. Unlike the majority of previous works which separately process detection and segmentation, M2BEV infered both tasks with a unified model and improved efficiency. M2BEV efficiently transformed multi-view 2D image features into the 3D BEV feature in ego-car coordinates. Such BEV representation is important as it enables different tasks to share a single encoder. Huang et al. [41] proposed a simple yet effective framework, termed Fast-BEV, which is capable of performing real-time BEV perception on the on-vehicle chips. Towards this goal, they first empirically found that the BEV representation could be sufficiently powerful without expensive view transformation or depth representation. They further introduced (1) a strong data augmentation strategy for both image and BEV space to avoid over-fitting (2) a multi-frame feature fusion mechanism to leverage the temporal information (3) an optimized deployment-friendly view transformation to speed up the inference. Chu et al. [19] proposed OA-BEV, a network that can be plugged into the Birds-Eye-View(BEV) based 3D object detection framework to bring out the objects by incorporating object-aware pseudo-3D features and depth features. First, they explicitly guided the network to learn the depth distribution by object-level supervision from each 3D object's center. Then, they selected the foreground pixels by a 2D object detector and projected them into 3D space for pseudo-voxel feature encoding.

5.3.3 Transformer-based approach

Carion et al. [10] proposed DEtection TRansformer or DETR with many ingradients as a set-based global loss that forces unique predictions via bipartite matching, and a transformer encoder-decoder architecture. Given a fixed small set of learned object queries, DETR reasons about the relations of the objects and the global image context to directly output the final set of predictions in parallel. Inspired by DETR, Wang et al. [133] proposed DETR3D, that extracts 2D features from multiple camera images and then uses a sparse set of 3D object queries to index into these 2D features,linking 3D positions to multi-view images using camera transformation matrices. This model makes a bounding box prediction per object query, using a set-to-set loss to measure the discrepancy between the ground-truth and the prediction. Doll et al. [26] proposed SpatialDETR, infers the classification and bounding box estimates based on attention both spatially within each image and across the different views. After image feature extraction a decoder-only transformer architecture is trained on a set-based loss. To fuse the multi-view information in the attention block they introduced a novel geometric positional encoding that incorporates the view ray geometry to explicitly consider the extrinsic and intrinsic camera setup. Zhang et al. [159] proposed SimMOD, a Simple baseline for Multi-camera Object Detection,to incorporate multiview information as well as build upon previous efforts on monocular 3D object detection, the framework is built on sample-wise object proposals and designed to work in a two-stage manner. First, they extracted multi-scale features and generate the perspective object proposals on each monocular image. Second, the multi-view proposals were aggregated and then iteratively refined with multi-view and multi-scale visual features in the DETR3D-style. The refined proposals were end-to-end decoded into the detection results. Liu et al. [77] proposed position embedding transformation (PETR) for

multi-view 3D object detection. PETR encodes the position information of 3D coordinates into image features, producing the 3D position-aware features. Object query could perceive the 3D position aware features and perform end-to-end object detection. Li et al. [68] proposed BEVFormer that exploits both spatial and temporal information by interacting with spatial and temporal space through predefined grid-shaped BEV queries. To aggregate spatial information, they designed spatial cross-attention that each BEV query extracts the spatial features from the regions of interest across camera views. For temporal information, they proposed temporal self attention to recurrently fuse the history BEV information. Jiang et al. [49] proposed a new Polar Transformer (PolarFormer) by exploiting the Polar coordinate system for more accurate 3D object detection in the bird's-eye-view (BEV) taking as input only multi-camera 2D images. This model made best use of the Polar representation rasterized via attending to the corresponding image observation in a sequence-to-sequence fashion subject to the geometric constraints.

5.4. RGB-D camera-based 3D object detection

RGB-D images consist of an RGB image with an additional depth map [27]. An RGB image and a depth image ideally have a one-to-one correspondence between pixels [131]. RGB images include textures and contours information, and the features can be extracted by 2D object detection techniques. In addition, they contain depth images that represent the geometric structure of space, providing the ability to understand the physical world. RGB-D data devices usually are more suitable for indoor scenes. RGB-D images can be created with stereo or TOF cameras that provide depth information in addition to color information. RGB-D sensors such as Apple Depth Camera, Microsoft Kinect [161], and RealSense [50] are reliable and affordable. RGB-D images are convenient to use with the majority of 2D object detection methods, treating depth information similarly to the three RGB channels [34].

5.4.1 RGB-D 3D representation

Tang and Lee [123] proposed a transferable semi-supervised 3D object detection model that learns a 3D object detector network from training data with two disjoint sets of object classes - a set of strong classes with both 2D and 3D box labels, and another set of weak classes with only 2D box labels. In particular, we suggest a relaxed re-projection loss, box prior loss and a Box-to-Point Cloud Fit network that allow us to effectively transfer useful 3D information from the strong classes to the weak classes during training, and consequently, enable the network to detect 3D objects in the weak classes during inference.

5.4.2 RGB-D 2D representation

Rahman et al. [108] proposed to leverage 2D regions for this task. Given a pair of color and depth image as input, they first predicted 2D region proposals from the designed multimodal fusion region proposal networks and then they proposed an efficient method to generate 3D bounding boxes from those region proposals by scaling down the 2D bounding boxes with a scale factor and projected it to 3D space.

In addition to the above two papers, other scholars proposed several research papers [15] [29] [14] [40] [108] [81] for 3D object detection using RGB-D cameras but were not applied to autonomous driving.

6. Datasets

Large scale, representative, labeled and real world data serves as the fuel for training deep learning networks, critical for improving self-driving perception algorithms. We listed the popular publicly available datasets for the 3D object detection in autonomous driving scenarios in the Table 1

KITTI [32] is one of the foremost publicly available dataset for 3D object detection. Their recording platform is equipped with four high resolution video cameras, a Velodyne laser scanner and a state-of-the-art localization system. Their benchmarks comprise 389 stereo and optical flow image pairs, stereo visual odometry sequences of 39.2 km length, and more than 200k 3D object annotations captured in cluttered scenarios (up to 15 cars and 30 pedestrians are visible per image). Later, the following datasets were made publicly available which are different in terms of size of the data, diversity of the data, number of annotated categories, providing minor object classes, data from different modalities.

Choi et al. [18] proposed the KAIST multi-spectral data set, which covers a great range of drivable regions, from urban to residential, for autonomous systems. This datset provides the different perspectives of the world captured in coarse time slots (day and night), in addition to fine time slots (sunrise, morning, afternoon, sunset, night, and dawn). For all-day perception of autonomous systems, they proposed the use of a different spectral sensor, i.e., a thermal imaging camera. Towards this goal, they developed a multi-sensor platform, which supports the use of a co-aligned RGB/Thermal camera, RGB stereo, 3-D LiDAR, and inertial sensors (GPS/IMU) and a related calibration technique.

Huang et al. [46] proposed ApolloScape open dataset that contains much large and richer labelling including holistic semantic dense point cloud for each site, stereo, per-pixel semantic labelling, lanemark labelling, instance segmentation, 3D car instance, high accurate location for every frame in various driving videos from multiple sites, cities and daytimes. For each task, it contains at lease 15x

larger amount of images than KITTI dataset [32].

Patil et al. [99] proposed the Honda Research Institute 3D Dataset (H3D), a large-scale full-surround 3D multi-object detection and tracking dataset collected using a 3D LiDAR scanner. H3D comprises of 160 crowded and highly interactive traffic scenes with a total of 1 million labeled instances in 27,721 frames. With unique dataset size, rich annotations, and complex scenes, H3D is gathered to stimulate research on full-surround 3D multi-object detection and tracking. It provides sufficient data and labels to tackle challenging scenes where highly interactive and occluded traffic participants are present.

Kensten et al. [51] proposed Lyft L5, a self-driving dataset for motion prediction, containing over 1,000 hours of data. This was collected by a fleet of 20 autonomous vehicles along a fixed route in Palo Alto, California, over a four-month period. It consists of 170,000 scenes, where each scene is 25 seconds long and captures the perception output of the self-driving system, which encodes the precise positions and motions of nearby vehicles, cyclists, and pedestrians over time.

Chang et al. [11] proposed Argoverse, a dataset designed to support autonomous vehicle perception tasks including 3D tracking and motion forecasting. Argoverse includes sensor data collected by a fleet of autonomous vehicles in Pittsburgh and Miami as well as 3D tracking annotations, 300k extracted interesting vehicle trajectories, and rich semantic maps. The sensor data consists of 360 degree images from 7 cameras with overlapping fields of view, forward-facing stereo imagery, 3D point clouds from long range LiDAR, and 6-DOF pose. Their 290km of mapped lanes contain rich geometric and semantic metadata which are not currently available in any public dataset.

Yogamani et al. [149] proposed WoodScape, the first extensive fisheye automotive dataset that comprises of four surround view cameras and nine tasks including segmentation, depth estimation, 3D bounding box detection and soiling detection. Semantic annotation of 40 classes at the instance level is provided for over 10,000 images and annotation for other tasks are provided for over 100,000 images.

Weng et al. [137] proposed AIODrive dataset, a synthetic large-scale dataset that provides comprehensive sensors, annotations and environmental variations. It forms a union of various strengths of existing datasets such as Semantic KITTI, nuScenes and Waymo. They provided (1) 8 sensor modalities, annotations for all mainstream perception tasks, (2) rare driving scenarios such as adverse weather and lighting, crowded scenes, (3) high-speed driving, violation of traffic rules, and accidents.

Pham et al. [100] proposed a new challenging A*3D dataset which consists of RGB images and LiDAR data with a significant diversity of scene, time, and weather. The dataset consists of high-density images (approximately 10 times more than the pioneering KITTI dataset), heavy occlusions, a large number of nighttime frames (approximately 3 times the nuScenes dataset), addressing the gaps in the existing datasets to push the boundaries of tasks in autonomous driving research to more challenging highly diverse environments. The dataset contains 39K frames, 7 classes, and 230K 3D object annotations.

Geyer et al. [33] proposed Audi Autonomous Driving Dataset (A2D2), that consists of simultaneously recorded images and 3D point clouds, together with 3D bounding boxes, semantic segmentation, instance segmentation, and data extracted from the automotive bus. Their sensor suite consists of six cameras and five LiDAR units, providing full 360 degree coverage. The recorded data is time synchronized and mutually registered. Annotations are for non-sequential frames: 41,277 frames with semantic segmentation image and point cloud labels, of which 12,497 frames also have 3D bounding box annotations for objects within the field of view of the front camera. In addition, we provide 392,556 sequential frames of unannotated sensor data for recordings in three cities in the south of Germany.

Gählert et al. [30] proposed Cityscapes 3D, extending the original Cityscapes dataset [20] with 3D bounding box annotations for all types of vehicles. In contrast to existing datasets, our 3D annotations were labeled using stereo RGB images only and capture all nine degrees of freedom. This leads to a pixel-accurate reprojection in the RGB image and a higher range of annotations compared to lidar-based approaches. In order to ease multitask learning, they provided a pairing of 2D instance segments with 3D bounding boxes.

Caeser et al. [7] proposed nuTonomy scenes (nuScenes), the first dataset to carry the full autonomous vehicle sensor suite: 6 cameras, 5 radars and 1 lidar, all with full 360 degree field of view. nuScenes comprises 1000 scenes, each 20s long and fully annotated with 3D bounding boxes for 23 classes and 8 attributes. It has 7x as many annotations and 100x as many images as the pioneering KITTI dataset.

Sun et al. [122] proposed Waymo Open, a large scale, high quality, and a diverse dataset that consists of 1150 scenes that each span 20 seconds, consisting of well synchronized and calibrated high quality LiDAR and camera data captured across a range of urban and suburban geographies. It is 15x more diverse than the largest camera+LiDAR dataset available based on our proposed diversity metric. We exhaustively annotated this data with 2D (camera image) and 3D (LiDAR) bounding boxes, with consistent identifiers across frames.

Wang et al. [134] proposed Cirrus, a new long-range bi-pattern LiDAR public dataset for autonomous driving tasks such as 3D object detection, critical to highway driving and timely decision making. Our platform is equipped with a high-resolution video camera and a pair of LiDAR sensors with a 250-meter effective range, which is significantly

longer than existing public datasets. They recorded paired point clouds simultaneously using both Gaussian and uniform scanning patterns. Point density varies significantly across such a long range, and different scanning patterns further diversify object representation in LiDAR. In Cirrus, eight categories of objects are exhaustively annotated in the LiDAR point clouds for the entire effective range.

Xiao et al. [139] proposed PandaSet, the first dataset produced by a complete, high-precision autonomous vehicle sensor kit with a no-cost commercial license. The dataset was collected using one 360° mechanical spinning LiDAR, one forward-facing, long-range LiDAR, and 6 cameras. The dataset contains more than 100 scenes, each of which is 8 seconds long, and provides 28 types of labels for object classification and 37 types of labels for semantic segmentation.

Liao et al. [73] proposed KITTI-360, a suburban driving dataset which comprises richer input modalities, comprehensive semantic instance annotations and accurate localization to facilitate research at the intersection of vision, graphics and robotics. For efficient annotation, they created a tool to label 3D scenes with bounding primitives and developed a model that transfers this information into the 2D image domain, resulting in over 150k images and 1B 3D points with coherent semantic instance annotations across 2D and 3D.

Mao et al. [87] ONCE (One millioN sCenEs) dataset that consists of 1 million LiDAR scenes and 7 million corresponding camera images. The data is selected from 144 driving hours, which is 20x longer than the largest 3D autonomous driving dataset available (e.g. nuScenes and Waymo), and it is collected across a range of different areas, periods and weather conditions.

Zhang et al. [157] proposed the Open Multi-modal Perception dataset (OpenMPD), a multi-modal perception benchmark objected at difficult examples. Compared with existing datasets, OpenMPD focuses more on those complex traffic scenes in urban areas with overexposure or darkness, crowded environment, unstructured roads and intersections. They acquired the multi-modal data through a vehicle with six cameras and four LiDAR for a 360-degree field of view and collected 180 clips of 20-second synchronized images at 20 Hz and point clouds at 10 Hz. Particularly, they applied a 128-beam LiDAR to provide Hi-Res point clouds to better understand the 3D environment and sensor fusion. They sampled 15 K keyframes at equal intervals from clips for annotations, including 2D/3D object detections, 3D object tracking, and 2D semantic segmentation.

Yu et al. [153] proposed DAIR-V2X Dataset, which is the first large-scale, multi-modality, multi-view dataset from real scenarios for r Vehicle-Infrastructure Cooperative Autonomous Driving. DAIR-V2X comprises 71254 LiDAR frames and 71254 Camera frames, and all frames are captured from real scenes with 3D annotations.

Most existing detectors are unable to detect uncommon objects and corner cases (e.g., a dog crossing a street), which may lead to severe accidents in some situations, making the timeline for the real-world application of reliable autonomous driving uncertain. Li at al. [61] proposed a challenging dataset named CODA that exposes this critical problem of vision based detectors. The dataset consists of 1500 carefully selected real-world driving scenes, each containing four object-level corner cases (on average), spanning more than 30 object categories.

Ye et al. [148] proposed a Roadside Perception 3D dataset- Rope3D from a novel view. This dataset consists of 50k images and over 1.5M 3D objects in various scenes, which are captured under different settings including various cameras with ambiguous mounting positions, camera specifications, viewpoints, and different environmental conditions. There are totally 13 object classes with their corresponding category, 2D properties (occlusion, truncation) and the 7-DOF(Degrees of Freedom) 3D bounding box.

Brazil et al. [5] proposed OMNI3D by combining existing datasets resulting in 234k images annotated with more than 3 million instances and 98 categories. 3D detection at such scale is challenging due to variations in camera intrinsics and the rich diversity of scene and object types.

Dokanian et al. [25] proposed IDD-3D dataset to facilitate better research toward accommodating unstructured and complex driving layout found in several developing countries such as India. It which consists of multimodal data from multiple cameras and LiDAR sensors with 12k annotated driving LiDAR frames across various traffic scenarios.

Burnett et al. [6] proposed Boreas dataset that was collected by driving a repeated route over the course of one year, resulting in stark seasonal variations and adverse weather conditions such as rain and falling snow. In total, the Boreas dataset includes over 350km of driving data featuring a 128-channel Velodyne Alpha Prime lidar, a 360○ Navtech CIR304-H scanning radar, a 5MP FLIR Blackfly S camera, and centimetre-accurate post-processed ground truth poses.

7. Discussion

Table 2 compares performance analysis of Monocular-based 3D object detection methods on KITTI dataset. We capture AP_{R40} (%) for 3D car detection in easy, moderate and hard weather conditions. Out of all the monocular methods, MonoXiver [76] achieved SOTA performance , which is based on anchor-based approach. However, AP_{R40} (%) for 3D car detection in ease, moderate and hard are just 25.24, 19.04, and 16.39 respectively.

Table 3 compares performance analysis of Stereo camera

Table 1. A summary of datasets for camera-based 3D object detection in driving snenarios.

Dataset	Year	Images	LiDAR scans	Classes
KITTI [32]	2012	15k	15k	8
KAIST [18]	2018	8.9k	8.9k	3
ApolloScape [46]	2019	144k	20k	6
H3D [99]	2018	83k	23k	8
Lyft L5 [51]	2019	323k	46k	9
Argoverse [11]	2019	490k	44k	15
WoodScape [149]	2019	10k	10k	3
AIODrive [137]	2020	250k	250k	-
A*3D [100]	2020	39k	39k	7
A2D2 [33]	2020	41.3k	12.5k	14
Cityscapes 3D [30]	2020	5k	-	8
nuScenes [7]	2020	1.4M	400k	23
Waymo Open [122]	2020	12M	1M	4
Cirrus [134]	2021	6.2k	6.2k	8
PandaSet [139]	2021	49k	8.2k	28
KITTI-360 [73]	2021	300k	80k	37
ONCE [87]	2021	7M	1M	5
OpenMD [157]	2021	15k	15k	6
DAIR-V2X [153]	2021	71k	71k	10
CODA [61]	2022	1.5k	-	30
Rope3D [148]	2022	50k	-	13
Omni3D [5]	2022	234k	-	98
IDD-3D [25]	2023	93k	15.5k	17
Boreas [6]	2023	7.1k	7.1 k	3

based 3D object detection methods on KITTI dataset. We capture AP_{R40} (%) for 3D car detection in easy, moderate and hard weather conditions. Out of all the stereo methods, PLUMENet [129] achieved SOTA performance in easy environment, and LIGA-Stereo [36] achieved SOTA performance in moderate and hard environments.

Table 4 compares performance analysis of Multiview camera based 3D object detection methods on nuScenes dataset. We capture mAP (%) and NDS scores. Out of all the multi-camera-based methods, SOLOFusion [98] achieved SOTA performance.

Table 5 compares performance analysis of state-of-the-art models of LiDAR and Fusion based 3D object detection methods on KITTI dataset.

Among all the camera-based methods, we observed that stereo-based methods outperformed the other methods but their performance is low compared to the LiDAR and fusion methods. As camera based methods are less expensive, we have to exploit this opportunity to improve the perfomance on par with the LiDAR and Fusion methods.

8. Future work

There is a large gap between the performance achieved by SOTA models of camera-based and non-camera (LiDAR and Fusion) based 3D object detection methods. More research needs to take place to improve the performance of camera-based methods due the cost and less computations involved in processing the input from cameras.

Most of the methods dealt with 3D object detection as a stand-alone task. There is a need to perform this along with the other related tasks such as prediction and planning for end-to-end control of Avs. The proposed 3D object detection methods are not mature enough to perform well in harsh weather conditions and also in identifying the objects at a farther distance. Hence, it is a prospective future research opportunity. More methods and more datsets to process the images from RGB-D camera are essential to evaluate the performance with this input type. Currently, this input performs well for indoor scenes. In this case, we have the opportunity to achieve the performance on par with LiDAR and Fusion based methods. One more promising opportunity to improve the performance of the 3D object detection would be the fusion of different camera-based methods.

Most of the popular autonomous driving datasets do not provide the images from RGB-D cameras. These are not diverse in terms of driving scenarios and they provide data collected in specific cities in specific countries. They contain annotations mostly related to cars, pedestrians and cyclists. When the model is trained with these datasets, it cannot perform well when unknown objects appear in the driving scenes. These datsest may not be completely sufficient for real world applications. There is a need to conduct research on how comprehensive and diverse datasets can be prepared to handle these issues.

Most of the methods evaluated their performance based on the detection of a single object 'car' but the driving scenes contain many other objects. We recommend future methods should be evaluated on detecting multiple objects.

Detecting the objects at a farther distance from the AV is essential for better path planning and vehicle control. But most of the methods did not consider this and so, we recommend the future research to consider this aspect as well.

9. Conclusion

We summarized, reviewed, and analyzed the performance of various deep learning methods across camera-based 3D object detection methods and compared their performances with respect to LiDAR and Fusion based methods. We identified the research opportunities to improve the performance of the former methods on par with the later methods. In addition, we listed various publicly available datasets and identified the need for diverse datasets. Finally,

Table 2. A comprehensive performance analysis of Monocular-based 3D object detection methods on KITTI dataset. We capture AP_{R40} (%) for 3D car detection in easy, moderate and hard weather conditions.

Model	Camera type	Year	Easy	Moderate	Hard
OFT-Net [111]	Monocular	2018	1.32	1.61	1.00
FQNet [75]	Monocular	2019	2.77	1.51	1.01
ROI-10D [86]	Monocular	2019	4.32	2.02	1.46
GS3D [59]	Monocular	2019	4.47	2.90	2.47
MonoFENet [2]	Monocular	2019	8.35	5.14	4.10
MonoGRNet [105]	Monocular	2019	9.61	5.74	4.25
MonoDIS [119]	Monocular	2019	10.37	7.94	6.40
MonoPSR [55]	Monocular	2019	10.76	7.25	5.85
M3D-RPN [3]	Monocular	2019	14.76	9.71	7.42
AM3D [84]	Monocular	2019	16.50	10.74	9.52
Cai et al. [8]	Monocular	2020	11.08	7.02	5.63
MonoPair [17]	Monocular	2020	13.04	9.99	8.65
SMOKE [79]	Monocular	2020	14.03	9.76	7.84
RTM3D [63]	Monocular	2020	14.41	10.34	8.77
MoVi-3D [120]	Monocular	2020	15.19	10.90	9.26
UR3D [115]	Monocular	2020	15.58	8.61	6.00
D^4LCN [24]	Monocular	2020	16.65	11.72	9.51
Ye et al. [147]	Monocular	2020	16.77	12.72	9.17
Kinematic3D [4]	Monocular	2020	19.07	12.72	9.17
CaDDN [109]	Monocular	2021	19.17	13.41	11.46
PatchNet [83]	Monocular	2021	15.68	11.12	10.17
MonoDLE [85]	Monocular	2021	17.23	12.26	10.29
M3DSSD [82]	Monocular	2021	17.51	11.46	8.98
Kumar et al. [56]	Monocular	2021	18.10	12.32	9.65
MonoRCNN [117]	Monocular	2021	18.36	12.65	10.03
MonoRUn [12]	Monocular	2021	19.65	12.30	10.58
DDMP [125]	Monocular	2021	19.71	12.78	9.80
MonoFlex [158]	Monocular	2021	19.94	13.89	12.07
GUP Net [80]	Monocular	2021	20.11	14.20	11.77
PCT [126]	Monocular	2021	21.00	13.37	11.31
MonoEF [164]	Monocular	2021	21.29	13.87	11.71
Liu et al. [78]	Monocular	2021	21.65	13.25	9.91
Lite-FPN [144]	Monocular	2021	15.32	10.64	8.59
DD3D [97]	Monocular	2021	23.22	16.34	14.20
MonoDTR [44]	Monocular	2022	21.99	15.39	12.73
Ji et al. [48]	Monocular	2022	12.46	7.86	6.30
MDS-Net [141]	Monocular	2022	24.30	14.46	11.12
M3DGAF [13]	Monocular	2022	19.48	12.66	10.99
MonoRCNN++ [116]	Monocular	2023	20.08	13.72	11.34
MonoEdge [166]	Monocular	2023	21.08	14.47	12.73
MonoPGC [138]	Monocular	2023	24.68	17.17	14.14
MonoATT [165]	Monocular	2023	24.72	17.37	15.00
MonoDETR. [156]	Monocular	2023	25.00	16.47	13.58
MonoXiver [76]	Monocular	2023	25.24	19.04	16.39

we mentioned the future research opportunities in the filed of camera-based 3D object detection.

References

[1] Simegnew Yihunie Alaba and John E Ball. A survey on deep-learning-based lidar 3d object detection for autonomous driving. *Sensors*, 22(24):9577, 2022. 2

Table 3. A comprehensive performance analysis of Stereo camera based 3D object detection methods on KITTI dataset. We capture AP_{R40} (%) for 3D car detection in easy, moderate and hard weather conditions.

Model	Camera type	Year	Easy	Moderate	Hard
TLNet [106]	Stereo	2019	7.64	4.37	3.74
Stereo R-CNN [62]	Stereo	2019	47.58	30.23	23.72
Pseudo-LiDAR [128]	Stereo	2019	54.53	34.05	28.25
OC-Stereo [103]	Stereo	2020	55.15	37.60	30.25
ZoomNet [142]	Stereo	2020	55.98	38.64	30.97
Disp R-CNN [121]	Stereo	2020	58.53	37.91	31.93
P-LiDAR++ [151]	Stereo	2020	61.11	42.43	36.99
Qian et al. [104]	Stereo	2020	64.8	43.9	38.1
DSGN [16]	Stereo	2020	73.50	52.18	45.14
CG-Stereo [60]	Stereo	2020	74.39	53.58	46.50
CDN [31]	Stereo	2020	74.52	54.22	46.36
PLUMENet [129]	Stereo	2021	82.97	66.27	56.70
LIGA-Stereo [36]	Stereo	2021	81.39	64.66	57.22
Shi et al. [118]	Stereo	2022	49.94	31.3	25.62
Pseudo-Mono [124]	Stereo	2023	27.41	18.57	16.16

Table 4. Performance analysis of Multiview camera based 3D object detection methods on nuScenes dataset. We capture mAP (%) and NDS scores.

Model	Camera type	Year	mAP	NDS
DETR3D [133]	Multiview	2021	41.2	47.9
BEVDet [43]	Multiview	2021	42.2	52.9
PETR [77]	Multiview	2022	44.5	50.4
BEVerse [160]	Multiview	2022	39.3	53.1
BEVFormer [68]	Multiview	2022	48.1	56.9
BEVDepth [66]	Multiview	2022	52.0	60.9
BEVStereo [65]	Multiview	2022	52.5	61.0
SOLOFusion [98]	Multiview	2022	54	61.9

[2] Wentao Bao, Bin Xu, and Zhenzhong Chen. Monofenet: Monocular 3d object detection with feature enhancement networks. *IEEE Transactions on Image Processing*, 29:2753–2765, 2019. 14

[3] Garrick Brazil and Xiaoming Liu. M3d-rpn: Monocular 3d region proposal network for object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9287–9296, 2019. 14

[4] Garrick Brazil, Gerard Pons-Moll, Xiaoming Liu, and Bernt Schiele. Kinematic 3d object detection in monocular video. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXIII 16*, pages 135–152. Springer, 2020. 14

[5] Garrick Brazil, Julian Straub, Nikhila Ravi, Justin Johnson, and Georgia Gkioxari. Omni3d: A large benchmark and model for 3d object detection in the wild. *arXiv preprint arXiv:2207.10660*, 2022. 12, 13

[6] Keenan Burnett, David J Yoon, Yuchen Wu, Andrew Zou Li, Haowei Zhang, Shichen Lu, Jingxing Qian, Wei-Kang Tseng, Andrew Lambert, Keith YK Leung, et al. Boreas: A multi-season autonomous driving dataset. *arXiv preprint arXiv:2203.10168*, 2022. 12, 13

[7] Holger Caesar, Varun Bankiti, Alex H Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuscenes: A multimodal dataset for autonomous driving. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11621–11631, 2020. 1, 11, 13

[8] Yingjie Cai, Buyu Li, Zeyu Jiao, Hongsheng Li, Xingyu Zeng, and Xiaogang Wang. Monocular 3d object detection with decoupled structured polygon estimation and height-guided depth estimation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 10478–10485, 2020. 14

[9] Hui Cao, Wenlong Zou, Yinkun Wang, Ting Song, and Mengjun Liu. Emerging threats in deep learning-based autonomous driving: A comprehensive survey. *arXiv preprint arXiv:2210.11237*, 2022. 2

[10] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *Com-*

Table 5. Performance analysis of state-of-the-art models of LiDAR and Fusion based 3D object detection methods on KITTI dataset.

Model	Sensor type	Representation	Year	Easy	Moderate	Hard
3DSSD [145]	LiDAR	Point	2020	88.36	79.57	74.55
Point-GNN [114]	LiDAR	Point	2020	88.33	79.47	72.29
Pointformer [92]	LiDAR	Point	2021	87.13	77.06	69.25
CIA-SSD [162]	LiDAR	Voxel	2021	89.59	80.28	72.87
Voxel R-CNN [22]	LiDAR	Voxel	2021	90.90	81.62	77.06
Voxel Transformer [88]	LiDAR	Voxel	2021	89.90	82.09	79.14
HDNet [143]	LiDAR	BEV Image	2018	89.14	86.57	78.32
PV-RCNN [113]	LiDAR	Point-Voxel	2020	90.25	81.43	76.82
PVGNet [89]	LiDAR	Point-Voxel	2021	89.94	81.81	77.09
PV-RCNN++ [112]	LiDAR	Point-Voxel	2021	90.14	81.88	77.15
RangeRCNN [71]	LiDAR	Range Image	2020	88.47	81.33	77.09
RangeIoUDet [72]	LiDAR	Range Image	2020	88.60	79.80	76.76
MMF [69]	Fusion	Intermediate	2019	86.81	76.75	64.81
3D-CVF [150]	Fusion	Intermediate	2020	89.20	80.05	73.11
EPNet [45]	Fusion	Intermediate	2020	89.81	79.28	74.59
CLOCs [93]	Fusion	Late	2020	88.94	80.67	77.15
Fast-CLOCs [94]	Fusion	Late	2022	89.11	80.34	76.98

puter Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part I 16, pages 213–229. Springer, 2020. 9

[11] Ming-Fang Chang, John Lambert, Patsorn Sangkloy, Jagjeet Singh, Slawomir Bak, Andrew Hartnett, De Wang, Peter Carr, Simon Lucey, Deva Ramanan, et al. Argoverse: 3d tracking and forecasting with rich maps. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8748–8757, 2019. 11, 13

[12] Hansheng Chen, Yuyao Huang, Wei Tian, Zhong Gao, and Lu Xiong. Monorun: Monocular 3d object detection by reconstruction and uncertainty propagation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10379–10388, 2021. 14

[13] Mu Chen, Pengfei Liu, and Huaici Zhao. M3dgaf: Monocular 3d object detection with geometric appearance awareness and feature fusion. *IEEE Sensors Journal*, 2022. 7, 14

[14] Xiaozhi Chen, Kaustav Kundu, Yukun Zhu, Andrew G Berneshawi, Huimin Ma, Sanja Fidler, and Raquel Urtasun. 3d object proposals for accurate object class detection. *Advances in neural information processing systems*, 28, 2015. 10

[15] Xiaozhi Chen, Kaustav Kundu, Yukun Zhu, Huimin Ma, Sanja Fidler, and Raquel Urtasun. 3d object proposals using stereo imagery for accurate object class detection. *IEEE transactions on pattern analysis and machine intelligence*, 40(5):1259–1272, 2017. 10

[16] Yilun Chen, Shu Liu, Xiaoyong Shen, and Jiaya Jia. Dsgn: Deep stereo geometry network for 3d object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12536–12545, 2020. 8, 15

[17] Yongjian Chen, Lei Tai, Kai Sun, and Mingyang Li. Monopair: Monocular 3d object detection using pairwise spatial relationships. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12093–12102, 2020. 14

[18] Yukyung Choi, Namil Kim, Soonmin Hwang, Kibaek Park, Jae Shin Yoon, Kyounghwan An, and In So Kweon. Kaist multi-spectral day/night data set for autonomous and assisted driving. *IEEE Transactions on Intelligent Transportation Systems*, 19(3):934–948, 2018. 10, 13

[19] Xiaomeng Chu, Jiajun Deng, Yuan Zhao, Jianmin Ji, Yu Zhang, Houqiang Li, and Yanyong Zhang. Oa-bev: Bringing object awareness to bird's-eye-view representation for multi-camera 3d object detection. *arXiv preprint arXiv:2301.05711*, 2023. 9

[20] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3213–3223, 2016. 11

[21] Yaodong Cui, Ren Chen, Wenbo Chu, Long Chen, Daxin Tian, Ying Li, and Dongpu Cao. Deep learning for image and point cloud fusion in autonomous driving: A review. *IEEE Transactions on Intelligent Transportation Systems*, 23(2):722–739, 2021. 2

[22] Jiajun Deng, Shaoshuai Shi, Peiwei Li, Wengang Zhou, Yanyong Zhang, and Houqiang Li. Voxel r-cnn: Towards high performance voxel-based 3d object detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 1201–1209, 2021. 16

[23] Yao Deng, Tiehua Zhang, Guannan Lou, Xi Zheng, Jiong Jin, and Qing-Long Han. Deep learning-based autonomous

driving systems: A survey of attacks and defenses. *IEEE Transactions on Industrial Informatics*, 17(12):7897–7912, 2021. 2

[24] Mingyu Ding, Yuqi Huo, Hongwei Yi, Zhe Wang, Jianping Shi, Zhiwu Lu, and Ping Luo. Learning depth-guided convolutions for monocular 3d object detection. In *Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition workshops*, pages 1000–1001, 2020. 14

[25] Shubham Dokania, AH Hafez, Anbumani Subramanian, Manmohan Chandraker, and CV Jawahar. Idd-3d: Indian driving dataset for 3d unstructured road scenes. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 4482–4491, 2023. 12, 13

[26] Simon Doll, Richard Schulz, Lukas Schneider, Viviane Benzin, Markus Enzweiler, and Hendrik PA Lensch. Spatialdetr: Robust scalable transformer-based 3d object detection from multi-view camera images with global cross-sensor attention. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXIX*, pages 230–245. Springer, 2022. 9

[27] Xinxin Du, Marcelo H Ang, Sertac Karaman, and Daniela Rus. A general pipeline for 3d detection of vehicles. In *2018 IEEE International Conference on Robotics and Automation (ICRA)*, pages 3194–3200. IEEE, 2018. 10

[28] Di Feng, Ali Harakeh, Steven L Waslander, and Klaus Dietmayer. A review and comparative study on probabilistic object detection in autonomous driving. *IEEE Transactions on Intelligent Transportation Systems*, 23(8):9961–9980, 2021. 2

[29] Max Ferguson and Kincho Law. A 2d-3d object detection system for updating building information models with mobile robots. In *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1357–1365. IEEE, 2019. 10

[30] Nils Gählert, Nicolas Jourdan, Marius Cordts, Uwe Franke, and Joachim Denzler. Cityscapes 3d: Dataset and benchmark for 9 dof vehicle detection. *arXiv preprint arXiv:2006.07864*, 2020. 11, 13

[31] Divyansh Garg, Yan Wang, Bharath Hariharan, Mark Campbell, Kilian Q Weinberger, and Wei-Lun Chao. Wasserstein distances for stereo disparity estimation. *Advances in Neural Information Processing Systems*, 33:22517–22529, 2020. 8, 15

[32] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *2012 IEEE conference on computer vision and pattern recognition*, pages 3354–3361. IEEE, 2012. 1, 10, 11, 13

[33] J Geyer, Y Kassahun, M Mahmudi, X Ricou, R Durgesh, AS Chung, L Hauswald, VH Pham, M Mühlegg, S Dorn, et al. A2d2: Audi autonomous driving dataset. arxiv 2020. *arXiv preprint arXiv:2004.06320*, 2004. 11, 13

[34] Silvio Giancola, Matteo Valenti, and Remo Sala. *A survey on 3D cameras: Metrological comparison of time-of-flight, structured-light and active stereoscopy technologies*. Springer, 2018. 10

[35] Sorin Grigorescu, Bogdan Trasnea, Tiberiu Cocias, and Gigel Macesanu. A survey of deep learning techniques for autonomous driving. *Journal of Field Robotics*, 37(3):362–386, 2020. 2

[36] Xiaoyang Guo, Shaoshuai Shi, Xiaogang Wang, and Hongsheng Li. Liga-stereo: Learning lidar geometry aware representations for stereo-based 3d detector. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3153–3163, 2021. 8, 13, 15

[37] Zhiyang Guo, Yingping Huang, Xing Hu, Hongjian Wei, and Baigan Zhao. A survey on deep learning based approaches for scene understanding in autonomous driving. *Electronics*, 10(4):471, 2021. 2

[38] Muhamad Amirul Haq, Shanq-Jang Ruan, Mei-En Shao, Qazi Mazhar Ul Haq, Pei-Jung Liang, and De-Qin Gao. One stage monocular 3d object detection utilizing discrete depth and orientation representation. *IEEE Transactions on Intelligent Transportation Systems*, 23(11):21630–21640, 2022. 5

[39] 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 (CVPR)*, pages 770–778, 2016. 3

[40] Ruotao He, Juan Rojas, and Yisheng Guan. A 3d object detection and pose estimation pipeline using rgb-d images. In *2017 IEEE International Conference on Robotics and Biomimetics (ROBIO)*, pages 1527–1532. IEEE, 2017. 10

[41] Bin Huang, Yangguang Li, Enze Xie, Feng Liang, Luya Wang, Mingzhu Shen, Fenggang Liu, Tianqi Wang, Ping Luo, and Jing Shao. Fast-bev: Towards real-time on-vehicle bird's-eye view perception. *arXiv preprint arXiv:2301.07870*, 2023. 9

[42] Junjie Huang and Guan Huang. Bevdet4d: Exploit temporal cues in multi-camera 3d object detection. *arXiv preprint arXiv:2203.17054*, 2022. 9

[43] Junjie Huang, Guan Huang, Zheng Zhu, Yun Ye, and Dalong Du. Bevdet: High-performance multi-camera 3d object detection in bird-eye-view. *arXiv preprint arXiv:2112.11790*, 2021. 9, 15

[44] Kuan-Chih Huang, Tsung-Han Wu, Hung-Ting Su, and Winston H Hsu. Monodtr: Monocular 3d object detection with depth-aware transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4012–4021, 2022. 7, 14

[45] Tengteng Huang, Zhe Liu, Xiwu Chen, and Xiang Bai. Epnet: Enhancing point features with image semantics for 3d object detection. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XV 16*, pages 35–52. Springer, 2020. 16

[46] Xinyu Huang, Peng Wang, Xinjing Cheng, Dingfu Zhou, Qichuan Geng, and Ruigang Yang. The apolloscape open dataset for autonomous driving and its application. *IEEE transactions on pattern analysis and machine intelligence*, 42(10):2702–2719, 2019. 10, 13

[47] Yu Huang and Yue Chen. Autonomous driving with deep learning: A survey of state-of-art technologies. *arXiv preprint arXiv:2006.06091*, 2020. 2

[48] Chaofeng Ji, Guizhong Liu, and Dan Zhao. Monocular 3d object detection via estimation of paired keypoints for autonomous driving. *Multimedia Tools and Applications*, 81(4):5973–5988, 2022. 5, 14

[49] Yanqin Jiang, Li Zhang, Zhenwei Miao, Xiatian Zhu, Jin Gao, Weiming Hu, and Yu-Gang Jiang. Polarformer: Multi-camera 3d object detection with polar transformers. *arXiv preprint arXiv:2206.15398*, 2022. 10

[50] Leonid Keselman, John Iselin Woodfill, Anders Grunnet-Jepsen, and Achintya Bhowmik. Intel realsense stereoscopic depth cameras. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 1–10, 2017. 10

[51] 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 av dataset 2019. *urlhttps://level5. lyft. com/dataset*, 1:3, 2019. 11, 13

[52] Curie Kim, Ue-Hwan Kim, and Jong-Hwan Kim. Self-supervised 3d object detection from monocular pseudo-lidar. In *2022 IEEE International Conference on Multisensor Fusion and Integration for Intelligent Systems (MFI)*, pages 1–6. IEEE, 2022. 8

[53] B Ravi Kiran, Ibrahim Sobh, Victor Talpaert, Patrick Mannion, Ahmad A Al Sallab, Senthil Yogamani, and Patrick Pérez. Deep reinforcement learning for autonomous driving: A survey. *IEEE Transactions on Intelligent Transportation Systems*, 23(6):4909–4926, 2021. 2

[54] Hendrik Königshof, Niels Ole Salscheider, and Christoph Stiller. Realtime 3d object detection for automated driving using stereo vision and semantic information. In *2019 IEEE Intelligent Transportation Systems Conference (ITSC)*, pages 1405–1410. IEEE, 2019. 5, 8

[55] Jason Ku, Alex D Pon, and Steven 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*, pages 11867–11876, 2019. 14

[56] Abhinav Kumar, Garrick Brazil, and Xiaoming 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*, pages 8973–8983, 2021. 14

[57] Sampo Kuutti, Richard Bowden, Yaochu Jin, Phil Barber, and Saber Fallah. A survey of deep learning applications to autonomous vehicle control. *IEEE Transactions on Intelligent Transportation Systems*, 22(2):712–733, 2020. 2

[58] Duy Le and Linh Nguyen. Simple linear iterative clustering based low-cost pseudo-lidar for 3d object detection in autonomous driving. *Multimedia Tools and Applications*, pages 1–17, 2023. 8

[59] Buyu Li, Wanli Ouyang, Lu Sheng, Xingyu Zeng, and Xiaogang Wang. Gs3d: An efficient 3d object detection framework for autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1019–1028, 2019. 14

[60] Chengyao Li, Jason Ku, and Steven L Waslander. Confidence guided stereo 3d object detection with split depth estimation. In *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 5776–5783. IEEE, 2020. 8, 15

[61] Kaican Li, Kai Chen, Haoyu Wang, Lanqing Hong, Chaoqiang Ye, Jianhua Han, Yukuai Chen, Wei Zhang, Chunjing Xu, Dit-Yan Yeung, et al. Coda: A real-world road corner case dataset for object detection in autonomous driving. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXVIII*, pages 406–423. Springer, 2022. 12, 13

[62] Peiliang Li, Xiaozhi Chen, and Shaojie Shen. Stereo r-cnn based 3d object detection for autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7644–7652, 2019. 15

[63] Peixuan Li, Huaici Zhao, Pengfei Liu, and Feidao 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*, pages 644–660. Springer, 2020. 14

[64] Ying Li, Lingfei Ma, Zilong Zhong, Fei Liu, Michael A Chapman, Dongpu Cao, and Jonathan Li. Deep learning for lidar point clouds in autonomous driving: A review. *IEEE Transactions on Neural Networks and Learning Systems*, 32(8):3412–3432, 2020. 2

[65] Yinhao Li, Han Bao, Zheng Ge, Jinrong Yang, Jianjian Sun, and Zeming Li. Bevstereo: Enhancing depth estimation in multi-view 3d object detection with dynamic temporal stereo. *arXiv preprint arXiv:2209.10248*, 2022. 9, 15

[66] Yinhao Li, Zheng Ge, Guanyi Yu, Jinrong Yang, Zengran Wang, Yukang Shi, Jianjian Sun, and Zeming Li. Bevdepth: Acquisition of reliable depth for multi-view 3d object detection. *arXiv preprint arXiv:2206.10092*, 2022. 9, 15

[67] Zhen Li, Yuliang Gao, Qingqing Hong, Yuren Du, Seiichi Serikawa, and Lifeng Zhang. Keypoint3d: Keypoint-based and anchor-free 3d object detection for autonomous driving with monocular vision. *Remote Sensing*, 15(5):1210, 2023. 5

[68] Zhiqi Li, Wenhai Wang, Hongyang Li, Enze Xie, Chonghao Sima, Tong Lu, Yu Qiao, and Jifeng Dai. Bevformer: Learning bird's-eye-view representation from multi-camera images via spatiotemporal transformers. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part IX*, pages 1–18. Springer, 2022. 10, 15

[69] Ming Liang, Bin Yang, Yun Chen, Rui Hu, and Raquel Urtasun. Multi-task multi-sensor fusion for 3d object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7345–7353, 2019. 16

[70] Zhenming Liang and Yingping Huang. Survey on deep learning-based 3d object detection in autonomous driving. *Transactions of the Institute of Measurement and Control*, 45(4):761–776, 2023. 3

[71] Zhidong Liang, Ming Zhang, Zehan Zhang, Xian Zhao, and Shiliang Pu. Rangercnn: Towards fast and accurate 3d object detection with range image representation. *arXiv preprint arXiv:2009.00206*, 2020. 16

[72] Zhidong Liang, Zehan Zhang, Ming Zhang, Xian Zhao, and Shiliang Pu. Rangeioudet: Range image based real-time 3d object detector optimized by intersection over union. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7140–7149, 2021. 16

[73] Yiyi Liao, Jun Xie, and Andreas Geiger. Kitti-360: A novel dataset and benchmarks for urban scene understanding in 2d and 3d. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022. 12, 13

[74] Jianbang Liu, Xinyu Mao, Yuqi Fang, Delong Zhu, and Max Q-H Meng. A survey on deep-learning approaches for vehicle trajectory prediction in autonomous driving. In *2021 IEEE International Conference on Robotics and Biomimetics (ROBIO)*, pages 978–985. IEEE, 2021. 2

[75] Lijie Liu, Jiwen Lu, Chunjing Xu, Qi Tian, and Jie Zhou. Deep fitting degree scoring network for monocular 3d object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1057–1066, 2019. 14

[76] Xianpeng Liu, Ce Zheng, Kelvin Cheng, Nan Xue, Guo-Jun Qi, and Tianfu Wu. Monocular 3d object detection with bounding box denoising in 3d by perceiver. *arXiv preprint arXiv:2304.01289*, 2023. 6, 12, 14

[77] Yingfei Liu, Tiancai Wang, Xiangyu Zhang, and Jian Sun. Petr: Position embedding transformation for multi-view 3d object detection. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXVII*, pages 531–548. Springer, 2022. 9, 15

[78] Yuxuan Liu, Yuan Yixuan, and Ming Liu. Ground-aware monocular 3d object detection for autonomous driving. *IEEE Robotics and Automation Letters*, 6(2):919–926, 2021. 14

[79] Zechen Liu, Zizhang Wu, and Roland 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*, pages 996–997, 2020. 14

[80] Yan Lu, Xinzhu Ma, Lei Yang, Tianzhu Zhang, Yating Liu, Qi Chu, Junjie Yan, and Wanli Ouyang. Geometry uncertainty projection network for monocular 3d object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3111–3121, 2021. 14

[81] Qianhui Luo, Huifang Ma, Li Tang, Yue Wang, and Rong Xiong. 3d-ssd: Learning hierarchical features from rgb-d images for amodal 3d object detection. *Neurocomputing*, 378:364–374, 2020. 10

[82] Shujie Luo, Hang Dai, Ling Shao, and Yong Ding. M3dssd: Monocular 3d single stage object detector. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6145–6154, 2021. 14

[83] Xinzhu Ma, Shinan Liu, Zhiyi Xia, Hongwen Zhang, Xingyu Zeng, and Wanli Ouyang. Rethinking pseudo-lidar representation. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIII 16*, pages 311–327. Springer, 2020. 14

[84] Xinzhu Ma, Zhihui Wang, Haojie Li, Pengbo Zhang, Wanli Ouyang, and Xin 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*, pages 6851–6860, 2019. 14

[85] Xinzhu Ma, Yinmin Zhang, Dan Xu, Dongzhan Zhou, Shuai Yi, Haojie Li, and Wanli Ouyang. Delving into localization errors for monocular 3d object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4721–4730, 2021. 14

[86] Fabian Manhardt, Wadim Kehl, and Adrien 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*, pages 2069–2078, 2019. 14

[87] Jiageng Mao, Minzhe Niu, Chenhan Jiang, Hanxue Liang, Jingheng Chen, Xiaodan Liang, Yamin Li, Chaoqiang Ye, Wei Zhang, Zhenguo Li, et al. One million scenes for autonomous driving: Once dataset. *arXiv preprint arXiv:2106.11037*, 2021. 12, 13

[88] Jiageng Mao, Yujing Xue, Minzhe Niu, Haoyue Bai, Jiashi Feng, Xiaodan Liang, Hang Xu, and Chunjing Xu. Voxel transformer for 3d object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3164–3173, 2021. 16

[89] Zhenwei Miao, Jikai Chen, Hongyu Pan, Ruiwen Zhang, Kaixuan Liu, Peihan Hao, Jun Zhu, Yang Wang, and Xin 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*, pages 3279–3288, 2021. 16

[90] Sajjad Mozaffari, Omar Y Al-Jarrah, Mehrdad Dianati, Paul Jennings, and Alexandros Mouzakitis. Deep learning-based vehicle behavior prediction for autonomous driving applications: A review. *IEEE Transactions on Intelligent Transportation Systems*, 23(1):33–47, 2020. 2

[91] Jianjun Ni, Yinan Chen, Yan Chen, Jinxiu Zhu, Deena Ali, and Weidong Cao. A survey on theories and applications for self-driving cars based on deep learning methods. *Applied Sciences*, 10(8):2749, 2020. 2

[92] Xuran Pan, Zhuofan Xia, Shiji Song, Li Erran Li, and Gao Huang. 3d object detection with pointformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7463–7472, 2021. 16

[93] Su Pang, Daniel Morris, and Hayder Radha. Clocs: Camera-lidar object candidates fusion for 3d object detection. In *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 10386–10393. IEEE, 2020. 16

[94] Su Pang, Daniel Morris, and Hayder 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*, pages 187–196, 2022. 16

[95] Ilias Papadeas, Lazaros Tsochatzidis, Angelos Amanatiadis, and Ioannis Pratikakis. Real-time semantic image segmentation with deep learning for autonomous driving: A survey. *Applied Sciences*, 11(19):8802, 2021. 2

[96] Shahrokh Paravarzar and Belqes Mohammad. Motion prediction on self-driving cars: A review. *arXiv preprint arXiv:2011.03635*, 2020. 2

[97] Dennis Park, Rares Ambrus, Vitor Guizilini, Jie Li, and Adrien Gaidon. Is pseudo-lidar needed for monocular 3d object detection? In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3142–3152, 2021. 14

[98] Jinhyung Park, Chenfeng Xu, Shijia Yang, Kurt Keutzer, Kris Kitani, Masayoshi Tomizuka, and Wei Zhan. Time will tell: New outlooks and a baseline for temporal multi-view 3d object detection. *arXiv preprint arXiv:2210.02443*, 2022. 9, 13, 15

[99] Abhishek Patil, Srikanth Malla, Haiming Gang, and Yi-Ting Chen. The h3d dataset for full-surround 3d multi-object detection and tracking in crowded urban scenes. In *2019 International Conference on Robotics and Automation (ICRA)*, pages 9552–9557. IEEE, 2019. 11, 13

[100] Quang-Hieu Pham, Pierre Sevestre, Ramanpreet Singh Pahwa, Huijing Zhan, Chun Ho Pang, Yuda Chen, Armin Mustafa, Vijay Chandrasekhar, and Jie Lin. A 3d dataset: Towards autonomous driving in challenging environments. In *2020 IEEE International Conference on Robotics and Automation (ICRA)*, pages 2267–2273. IEEE, 2020. 11, 13

[101] Trung Pham, Mehran Maghoumi, Wanli Jiang, Bala Siva Sashank Jujjavarapu, Mehdi Sajjadi Xin Liu, Hsuan-Chu Lin, Bor-Jeng Chen, Giang Truong, Chao Fang, Junghyun Kwon, et al. Nvautonet: Fast and accurate 360 degree 3d visual perception for self driving. *arXiv e-prints*, pages arXiv–2303, 2023. 8

[102] Jonah Philion and Sanja Fidler. Lift, splat, shoot: Encoding images from arbitrary camera rigs by implicitly unprojecting to 3d. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIV 16*, pages 194–210. Springer, 2020. 8

[103] Alex D Pon, Jason Ku, Chengyao Li, and Steven L Waslander. Object-centric stereo matching for 3d object detection. In *2020 IEEE International Conference on Robotics and Automation (ICRA)*, pages 8383–8389. IEEE, 2020. 15

[104] Rui Qian, Divyansh Garg, Yan Wang, Yurong You, Serge Belongie, Bharath Hariharan, Mark Campbell, Kilian Q Weinberger, and Wei-Lun 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*, pages 5881–5890, 2020. 8, 15

[105] Zengyi Qin, Jinglu Wang, and Yan Lu. Monogrnet: A geometric reasoning network for monocular 3d object localization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 8851–8858, 2019. 14

[106] Zengyi Qin, Jinglu Wang, and Yan Lu. Triangulation learning network: from monocular to stereo 3d object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7615–7623, 2019. 15

[107] Shenming Qu, Xinyu Yang, Yiming Gao, and Shengbin Liang. Monodcn: Monocular 3d object detection based on dynamic convolution. *Plos one*, 17(10):e0275438, 2022. 6

[108] Mohammad Muntasir Rahman, Yanhao Tan, Jian Xue, Ling Shao, and Ke Lu. 3d object detection: Learning 3d bounding boxes from scaled down 2d bounding boxes in rgb-d images. *Information Sciences*, 476:147–158, 2019. 10

[109] Cody Reading, Ali Harakeh, Julia Chae, and Steven L Waslander. Categorical depth distribution network for monocular 3d object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8555–8564, 2021. 14

[110] Cody Ariel Reading. *Monocular 3D Object Detection and 3D Multi-Object Tracking for Autonomous Vehicles*. PhD thesis, University of Toronto (Canada), 2022. 7

[111] Thomas Roddick, Alex Kendall, and Roberto Cipolla. Orthographic feature transform for monocular 3d object detection. *arXiv preprint arXiv:1811.08188*, 2018. 14

[112] S Shi, L Jiang, J Deng, Z Wang, C Guo, J Shi, X Wang, and H Li. Pv-rcnn++: Point-voxel feature set abstraction with local vector representation for 3d object detection. arxiv 2021. *arXiv preprint arXiv:2102.00463*. 16

[113] Shaoshuai Shi, Chaoxu Guo, Li Jiang, Zhe Wang, Jianping Shi, Xiaogang Wang, and Hongsheng 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*, pages 10529–10538, 2020. 16

[114] Weijing Shi and Raj Rajkumar. Point-gnn: Graph neural network for 3d object detection in a point cloud. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1711–1719, 2020. 16

[115] Xuepeng Shi, Zhixiang Chen, and Tae-Kyun Kim. Distance-normalized unified representation for monocular 3d object detection. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXIX 16*, pages 91–107. Springer, 2020. 14

[116] Xuepeng Shi, Zhixiang Chen, and Tae-Kyun Kim. Multivariate probabilistic monocular 3d object detection. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 4281–4290, 2023. 7, 14

[117] Xuepeng Shi, Qi Ye, Xiaozhi Chen, Chuangrong Chen, Zhixiang Chen, and Tae-Kyun Kim. Geometry-based distance decomposition for monocular 3d object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15172–15181, 2021. 14

[118] Yuguang Shi, Yu Guo, Zhenqiang Mi, and Xinjie Li. Stereo centernet-based 3d object detection for autonomous driving. *Neurocomputing*, 471:219–229, 2022. 3, 5, 8, 15

[119] Andrea Simonelli, Samuel Rota Bulo, Lorenzo Porzi, Manuel López-Antequera, and Peter Kontschieder. Disentangling monocular 3d object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1991–1999, 2019. 14

[120] Andrea Simonelli, Samuel Rota Bulo, Lorenzo Porzi, Elisa Ricci, and Peter Kontschieder. Towards generalization across depth for monocular 3d object detection. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXII 16*, pages 767–782. Springer, 2020. 14

[121] Jiaming Sun, Linghao Chen, Yiming Xie, Siyu Zhang, Qinhong Jiang, Xiaowei Zhou, and Hujun 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*, pages 10548–10557, 2020. 15

[122] Pei Sun, Henrik Kretzschmar, Xerxes Dotiwalla, Aurelien Chouard, Vijaysai Patnaik, Paul Tsui, James Guo, Yin Zhou, Yuning Chai, Benjamin 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*, pages 2446–2454, 2020. 1, 11, 13

[123] Yew Siang Tang and Gim Hee Lee. Transferable semi-supervised 3d object detection from rgb-d data. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1931–1940, 2019. 6, 10

[124] Chongben Tao, JieCheng Cao, Chen Wang, Zufeng Zhang, and Zhen Gao. Pseudo-mono for monocular 3d object detection in autonomous driving. *IEEE Transactions on Circuits and Systems for Video Technology*, 2023. 8, 15

[125] Li Wang, Liang Du, Xiaoqing Ye, Yanwei Fu, Guodong Guo, Xiangyang Xue, Jianfeng Feng, and Li Zhang. Depth-conditioned dynamic message propagation for monocular 3d object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 454–463, 2021. 14

[126] Li Wang, Li Zhang, Yi Zhu, Zhi Zhang, Tong He, Mu Li, and Xiangyang Xue. Progressive coordinate transforms for monocular 3d object detection. *Advances in Neural Information Processing Systems*, 34:13364–13377, 2021. 14

[127] Li Wang, Xinyu Zhang, Ziying Song, Jiangfeng Bi, Guoxin Zhang, Haiyue Wei, Liyao Tang, Lei Yang, Jun Li, Caiyan Jia, et al. Multi-modal 3d object detection in autonomous driving: A survey and taxonomy. *IEEE Transactions on Intelligent Vehicles*, 2023. 2

[128] Yan Wang, Wei-Lun Chao, Divyansh Garg, Bharath Hariharan, Mark Campbell, and Kilian 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*, pages 8445–8453, 2019. 15

[129] Yan Wang, Bin Yang, Rui Hu, Ming Liang, and Raquel Urtasun. Plumenet: Efficient 3d object detection from stereo images. In *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 3383–3390. IEEE, 2021. 8, 13, 15

[130] Yijing Wang, Sheng Xu, Zhiqiang Zuo, and Zheng Li. Monocular 3d object detection based on pseudo-lidar point cloud for autonomous vehicles. In *2022 41st Chinese Control Conference (CCC)*, pages 5469–5474. IEEE, 2022. 4, 7

[131] Yilin Wang and Jiayi Ye. An overview of 3d object detection. *arXiv preprint arXiv:2010.15614*, 2020. 10

[132] Yonggui Wang, Jian Li, Zaicheng Zhang, and Bin He. Detr 3d object detection method based on fusion of depth and salient information. 7

[133] Yue Wang, Vitor Campagnolo Guizilini, Tianyuan Zhang, Yilun Wang, Hang Zhao, and Justin Solomon. Detr3d: 3d object detection from multi-view images via 3d-to-2d queries. In *Conference on Robot Learning*, pages 180–191. PMLR, 2022. 9, 15

[134] Ze Wang, Sihao Ding, Ying Li, Jonas Fenn, Sohini Roychowdhury, Andreas Wallin, Lane Martin, Scott Ryvola, Guillermo Sapiro, and Qiang Qiu. Cirrus: A long-range bi-pattern lidar dataset. In *2021 IEEE International Conference on Robotics and Automation (ICRA)*, pages 5744–5750. IEEE, 2021. 11, 13

[135] Zengran Wang, Chen Min, Zheng Ge, Yinhao Li, Zeming Li, Hongyu Yang, and Di Huang. Sts: Surround-view temporal stereo for multi-view 3d detection. *arXiv preprint arXiv:2208.10145*, 2022. 9

[136] Li-Hua Wen and Kang-Hyun Jo. Deep learning-based perception systems for autonomous driving: A comprehensive survey. *Neurocomputing*, 2022. 2

[137] Xinshuo Weng, Yunze Man, Dazhi Cheng, Jinhyung Park, Matthew O'Toole, Kris Kitani, Jianren Wang, and David Held. All-in-one drive: A large-scale comprehensive perception dataset with high-density long-range point clouds. *arXiv*, 2020. 11, 13

[138] Zizhang Wu, Yuanzhu Gan, Lei Wang, Guilian Chen, and Jian Pu. Monopgc: Monocular 3d object detection with pixel geometry contexts. *arXiv preprint arXiv:2302.10549*, 2023. 7, 14

[139] Pengchuan Xiao, Zhenlei Shao, Steven Hao, Zishuo Zhang, Xiaolin Chai, Judy Jiao, Zesong Li, Jian Wu, Kai Sun, Kun Jiang, et al. Pandaset: Advanced sensor suite dataset for autonomous driving. In *2021 IEEE International Intelligent Transportation Systems Conference (ITSC)*, pages 3095–3101. IEEE, 2021. 12, 13

[140] Enze Xie, Zhiding Yu, Daquan Zhou, Jonah Philion, Anima Anandkumar, Sanja Fidler, Ping Luo, and Jose M Alvarez. Mˆ 2bev: Multi-camera joint 3d detection and segmentation with unified birds-eye view representation. *arXiv preprint arXiv:2204.05088*, 2022. 9

[141] Zhouzhen Xie, Yuying Song, Jingxuan Wu, Zecheng Li, Chunyi Song, and Zhiwei Xu. Mds-net: Multi-scale depth stratification 3d object detection from monocular images. *Sensors*, 22(16):6197, 2022. 7, 14

[142] Zhenbo Xu, Wei Zhang, Xiaoqing Ye, Xiao Tan, Wei Yang, Shilei Wen, Errui Ding, Ajin Meng, and Liusheng Huang. Zoomnet: Part-aware adaptive zooming neural network for 3d object detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 12557–12564, 2020. 15

[143] Bin Yang, Ming Liang, and Raquel Urtasun. Hdnet: Exploiting hd maps for 3d object detection. In *Conference on Robot Learning*, pages 146–155. PMLR, 2018. 16

[144] Lei Yang, Xinyu Zhang, Jun Li, Li Wang, Minghan Zhu, and Lei Zhu. Lite-fpn for keypoint-based monocular 3d object detection. *Knowledge-Based Systems*, page 110517, 2023. 3, 4, 5, 14

[145] Zetong Yang, Yanan Sun, Shu Liu, and Jiaya Jia. 3dssd: Point-based 3d single stage object detector. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11040–11048, 2020. 16

[146] Shanliang Yao, Runwei Guan, Xiaoyu Huang, Zhuoxiao Li, Xiangyu Sha, Yong Yue, Eng Gee Lim, Hyungjoon Seo, Ka Lok Man, Xiaohui Zhu, et al. Radar-camera fusion for object detection and semantic segmentation in autonomous driving: A comprehensive review. *arXiv preprint arXiv:2304.10410*, 2023. 3

[147] Xiaoqing Ye, Liang Du, Yifeng Shi, Yingying Li, Xiao Tan, Jianfeng Feng, Errui Ding, and Shilei Wen. Monocular 3d object detection via feature domain adaptation. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IX 16*, pages 17–34. Springer, 2020. 14

[148] Xiaoqing Ye, Mao Shu, Hanyu Li, Yifeng Shi, Yingying Li, Guangjie Wang, Xiao Tan, and Errui Ding. Rope3d: Theroadside perception dataset for autonomous driving and monocular 3d object detection task. *arXiv preprint arXiv:2203.13608*, 2022. 12, 13

[149] Senthil Yogamani, Ciarán Hughes, Jonathan Horgan, Ganesh Sistu, Padraig Varley, Derek O'Dea, Michal Uricár, Stefan Milz, Martin Simon, Karl Amende, et al. Woodscape: A multi-task, multi-camera fisheye dataset for autonomous driving. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9308–9318, 2019. 11, 13

[150] Jin Hyeok Yoo, Yecheol Kim, Jisong Kim, and Jun Won Choi. 3d-cvf: Generating joint camera and lidar features using cross-view spatial feature fusion for 3d object detection. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXVII 16*, pages 720–736. Springer, 2020. 16

[151] Yurong You, Yan Wang, Wei-Lun Chao, Divyansh Garg, Geoff Pleiss, Bharath Hariharan, Mark Campbell, and Kilian Q Weinberger. Pseudo-lidar++: Accurate depth for 3d object detection in autonomous driving. *arXiv preprint arXiv:1906.06310*, 2019. 8, 15

[152] Fisher Yu, Dequan Wang, Evan Shelhamer, and Trevor Darrell. Deep layer aggregation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2403–2412, 2018. 3

[153] Haibao Yu, Yizhen Luo, Mao Shu, Yiyi Huo, Zebang Yang, Yifeng Shi, Zhenglong Guo, Hanyu Li, Xing Hu, Jirui Yuan, et al. Dair-v2x: A large-scale dataset for vehicle-infrastructure cooperative 3d object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21361–21370, 2022. 12, 13

[154] Georgios Zamanakos, Lazaros Tsochatzidis, Angelos Amanatiadis, and Ioannis Pratikakis. A comprehensive survey of lidar-based 3d object detection methods with deep learning for autonomous driving. *Computers & Graphics*, 99:153–181, 2021. 2

[155] Hao Zhang, Hongyang Li, Xingyu Liao, Feng Li, Shilong Liu, Lionel M Ni, and Lei Zhang. Da-bev: Depth aware bev transformer for 3d object detection. *arXiv preprint arXiv:2302.13002*, 2023. 7

[156] Renrui Zhang, Han Qiu, Tai Wang, Ziyu Guo, Ziteng Cui, Yu Qiao, Hao Dong, Peng Gao, and Hongsheng Li. Monodetr: Depth-guided transformer for monocular 3d object detection. 7, 14

[157] Xinyu Zhang, Zhiwei Li, Yan Gong, Dafeng Jin, Jun Li, Li Wang, Yanzhang Zhu, and Huaping Liu. Openmpd: An open multimodal perception dataset for autonomous driving. *IEEE Transactions on Vehicular Technology*, 71(3):2437–2447, 2022. 12, 13

[158] Yunpeng Zhang, Jiwen Lu, and Jie Zhou. Objects are different: Flexible monocular 3d object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3289–3298, 2021. 14

[159] Yunpeng Zhang, Wenzhao Zheng, Zheng Zhu, Guan Huang, Jie Zhou, and Jiwen Lu. A simple baseline for multi-camera 3d object detection. *arXiv preprint arXiv:2208.10035*, 2022. 4, 6, 9

[160] Yunpeng Zhang, Zheng Zhu, Wenzhao Zheng, Junjie Huang, Guan Huang, Jie Zhou, and Jiwen Lu. Beverse: Unified perception and prediction in birds-eye-view for vision-centric autonomous driving. *arXiv preprint arXiv:2205.09743*, 2022. 15

[161] Zhengyou Zhang. Microsoft kinect sensor and its effect. *IEEE multimedia*, 19(2):4–10, 2012. 10

[162] Wu Zheng, Weiliang Tang, Sijin Chen, Li Jiang, and Chi-Wing Fu. Cia-ssd: Confident iou-aware single-stage object detector from point cloud. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 3555–3562, 2021. 16

[163] Xingyi Zhou, Dequan Wang, and Philipp Krähenbühl. Objects as points. *arXiv preprint arXiv:1904.07850*, 2019. 3

[164] Yunsong Zhou, Yuan He, Hongzi Zhu, Cheng Wang, Hongyang Li, and Qinhong Jiang. Monocular 3d object detection: An extrinsic parameter free approach. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7556–7566, 2021. 14

[165] Yunsong Zhou, Hongzi Zhu, Quan Liu, Shan Chang, and Minyi Guo. Monoatt: Online monocular 3d object detection with adaptive token transformer. *arXiv preprint arXiv:2303.13018*, 2023. 7, 14

[166] Minghan Zhu, Lingting Ge, Panqu Wang, and Huei Peng. Monoedge: Monocular 3d object detection using local perspectives. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 643–652, 2023. 7, 14