IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Date of publication xxxx 00, 0000, date of current version V1.

Digital Object Identifier xx.xxxx/ACCESS.2021.DOI

A Comprehensive Review on 3D Object Detection and 6D Pose Estimation with Deep Learning

SABERA HOQUE[1], MD. YASIR ARAFAT, SHUXIANG XU[1], ANANDA MAITI[1], and YUCHEN WEI[1].

[1]School of ICT, University of Tasmania, Newnham, Australia (e-mail: sabera.hoque@utas.edu.au, yasir@arafat.info, shuxiang.xu@utas.edu.au, anandamaiti@live.com, yuchen.wei@utas.edu.au)

Corresponding author: Sabera Hoque (e-mail: sabera.hoque@utas.edu.au).

ABSTRACT Nowadays, computer vision with 3D (dimension) object detection and 6D (degree of freedom) pose assumptions are widely discussed and studied in the field. In the 3D object detection process, classifications are centered on the object's size, position, and direction. And in 6D pose assumptions, networks emphasize 3D translation and rotation vectors. Successful application of these strategies can have a huge impact on various machine learning-based applications, including the autonomous vehicles, the robotics industry, and the augmented reality sector. Although extensive work has been done on 3D object detection with a pose assumption from RGB images, the challenges have not been fully resolved. Our analysis provides a comprehensive review of the proposed contemporary techniques for complete 3D object detection and the recovery of 6D pose assumptions of an object. In this review research paper, we have discussed several proposed sophisticated methods in 3D object detection and 6D pose estimation, including some popular data sets, evaluation matrix, and proposed method challenges. Most importantly, this study makes an effort to offer some possible future directions in 3D object detection and 6D pose estimation. We accept the autonomous vehicle as the sample case for this detailed review. Finally, this review provides a complete overview of the latest in-depth learning-based research studies related to 3D object detection and 6D pose estimation systems and also points out a comparison between some popular frameworks. To be more concise, we propose a detailed summary of the state-of-the-art techniques of modern deep learning-based object detection and pose estimation models.

INDEX TERMS Machine Learning, Deep Neural Network, Computer Vision, Object Detection, Pose Estimation, Image Processing, CNN, 3D Object Detection, 6D Object Detection

I. INTRODUCTION

Recently with the advancement of three-dimensional (3D) technology, the reconstruction of 3D models with pose assumptions has become a popular research topic. The main purpose of 3D model identification is to extract powerful features from RGB or RGBD images that can automatically improve the transportation system. Advanced models can make the map smarter and reduce vehicle costs. There are many challenges to this research concept, such as differentiation of perspectives, scaling, posture determination, illumination change, partial inclusion, adaptation detection, and background clutter.

Although many approaches and algorithms have been proposed and implemented for 2D image detection, the challenges of retrieving 3D objects from 2D images are still being explored. Moreover, estimating poses from this model is also important for the robot industry. One of the core examples in the 3D object detection and pose estimation research sector is the autonomous vehicle, where image detection plays a vital role in recovering 3D objects from 2D images [108]. The modern world is automatically moving towards an intelligent transportation system that requires the successful implementation of autonomous vehicles. The most important issue for self-driving systems is how various modern technologies can be applied to enhance the efficiency of self-driving vehicles.

The great debate in smart car systems is which one works better for object detection, the LiDAR (Light Detection and Ranging) or camera. Also, it needs to be studied whether it is effective to use a combination of the LiDAR and camera systems. For example, both Waymo and Uber include LiDAR where Tesla only uses cameras in their smart car system. Yet no technology has been universally accepted as a final self-driving solution on the road [181].

LiDAR, a proven technique of measuring distance, applies light pulses to determine both the distance and range of the surrounding object to avoid a collision and reduce the vehicle's speed. The technology helps self-propelled vehicles create visual 3D maps using on-board software, sending millions of pulses per second based on readings from light pulses and providing the vehicle with information about its surroundings. LiDAR is used in conjunction with cameras that provide a 360-degree view of the surroundings in self-driving cars, so they are not a standalone solution in themselves.

The camera provides images in intelligent car software that can analyze with a high level of accuracy using AI (artificial intelligence). The autopilot system uses cameras to provide a 360-degree view of its surroundings. The system returns entirely visual data from the lens's optics to on-board software and does not rely on the range and detection like LiDAR for situation analysis. With the development of NNs (Neural Network) and CV (Computer Vision) algorithms, objects can be identified to provide surrounding information while driving. This helps the car avoid collisions, slow down or brake when there is traffic, change lanes safely, and read text from road or highway signs using OCR (Recognition of Optical Character).

Although LiDAR has been proven to see things even in dangerous or foggy weather, it is not always reliable, as it is affected by wavelength stability, temperature, and detective sensitivity. This difficulty makes LiDAR technology more expensive. Moreover, LiDAR requires more space to apply to cars, thus making self-driving cars look bulky and less attractive. On the other hand, cameras are better, easy to implement, and comparatively less expensive in visual recognition. The software requires more data processing to create images and identify objects for LiDAR data than visual data. Finally, the camera has been implemented with Tesla as a standalone system; however, other OEMs(original equipment manufacturer) believe that applying other sensors, including radar, to detect range and distance can improve the performance of self-driving.

The ultimate visual recognition system also required the accurate calculation of other vehicles pose (car, bus, bicycle). Without predicting the actual pose of other vehicles, an autonomous car cannot make accurate decisions on whether to slow, brake or change direction. Recent state-of-the-art RGB-based 6 DoF (Degree of Freedom) pose estimation frameworks can be divided into two stages [50, 115, 240], including the object detection with 3D rotation by applying a trained framework and the estimation of 3D translation and 3D orientation (6D pose estimation) via relative distance estimation. Basically, the camera pose estimation is related to object localization, coordinates, and orientation. It is a crucial task not only for the autonomous car but also for the robot and navigation technology, the medical sector, and AR (augmented reality) [268]. In this review, we will mainly focus on the papers that work on the autonomous car and predicting the position of on-road cars or obstacles.

The rest of the section is organized as follows: In section I-A we present the contributions of this review article of deep learning for 3D Object Detection and 6D Pose Estimation. Moreover, in section I-B we have shown the difference between our review and other existing review articles. In sections, I-C and I-D we have discussed the pervasiveness of both 3D object detection and 6D pose estimation. Finally, in section I-E, we have briefly discussed the paper collection process.

A. CONTRIBUTIONS OF THIS REVIEW TO DEEP LEARNING

The purpose of this thinking is to thoroughly review the advanced essays in the 3D learning object detection literature and the 6D pose assumptions from RGB and RGB-D images. It provides a brief overview of current research that is easily comprehensible, and anyone who is interested can grasp the basics of 3D object detection (3DOD) and a 6D pose aspiration (6DPE) system. Moreover, most importantly, this review provides explicit knowledge of 3DOD and 6DPE applications in the field of computer vision to encourage a whole new set of novel methods and ideas. This paper proposes a rich survey for academics interested in research, the autonomous industry and the 3DOD and 6DPE fields. The survey will provide rough guidelines and possible directions for 3D object detection and 6D pose estimation methods, where most of the paperwork relates to autonomous vehicles.

Altogether, the survey has several objectives, such as:

1) We have provided a comprehensive review for a 3D object detection and 6D pose estimation system based on deep learning.,
2) We have created an overview for advanced strategies,
3) We discussed the challenges, advantages, disadvantages of the various proposed strategies
4) We have identified and cited a significant number of innovative concepts and incoming directions in this research sector
5) We can detect vision and broaden the horizons of 3D object detection and research DL (Deep Learning) methods of 6D pose estimation research techniques,
6) In this review, we have tried to give a brief overview on some of the popular datasets available for computer vision.,
7) We have focused on a few popular assessment methods and created a shortlist.

B. DIFFERENCE WITH OTHER FORMER REVIEWS

To date, much work has been done on 3D Object detection (3DOD) and 6D Pose estimation (6DPE), where most of them are deep learning-based. Nevertheless, the progress of a comprehensive review on the subject is still insufficient. This review sought to create a broad abstraction of modern research with DNN (Deep Neural Network) based 3D object detection 6D pose estimation systems and showed future directions. We can keep an eye on the paper by Mukhtar et al. [172], where they reviewed 194 documents and worked on on-road-based vehicle detection and tracking systems for collision avoidance systems. This review is organized based on various vehicle detection processes, including car detection and tracking sensors.

Sahin et al. [208], presents a comprehensive and up-to-date review where authors discuss object detection, examine more than 200 documents, and pose recovery methods with some popular data sets. In addition, several evaluation methods, open issues, and future research directions have been discussed in the paper.

Sivaraman et al. [228], has also conducted a literary survey on the method of identifying, tracking and behaving on-road aspects of self-driving vehicles. This study focuses on the current literature related to vision and sensor-based vehicle detection techniques. It began with about 200 papers on environmental perception on the road from 2005. The review papers are mostly related to single vision, stereo vision, the combination of single and stereo vision and sensor-fusion methods for vehicle tracking, detailed image aircraft, 3D modelling, measurement and filtering. Finally, they have called for visionary vehicle identification, tracking, and behavioural analysis with future research directions.

Ioannidou et al. [104], discussed the various method of deep learning architecture on different types of 3D data and provided a classification of multiple approaches. Zhao et al. [294], provided a regular survey of DL-based object detection frameworks by reviewing a total of 194 research papers. This review begins with a brief history of deep learning with several DL type classifications. Generic Object Detection strategies are discussed here, along with some changes and improved detection performance concepts such as object detection, salient object detection, pedestrian detection, and face detection.

Zhou et al. [300], have conducted a review for aspect-based SFM (Structure Form Motion) method, VO (Visual Odometry), and SLAM (Simultaneous Localization and Mapping) based methods where the methods play an important role for support in autonomous driving systems. In their work, they focused on multiple sensor-based methods such as Internal Measurement Unit (IMU) sensors, LiDAR, GPS (global positioning system), monocular-based methods (depending on the height of the camera).

One of the latest online reviews of 3D object detection written by Liu [150], published in the science blog "Towards Data Science", has covered around 32 current state-of-the-art mono3DOD methods as of November 2019. This review did not focus on pose estimation and gave only a brief idea about it. This review is more organized (papers are grouped into several groups) than other previous surveys and gives a more accurate picture of the related article. Unfortunately, there are insufficient numbers of surveys on DL (deep learning) - which stem from the 6D pose estimation system, so researchers should focus on this.

Sahin et al. [209], wrote a review related to 6D pose hypotheses where they cover numerous research articles that analyze both object identification and pose hypotheses. Their review article mainly focuses on multiple dataset challenges such as occlusion, cluttered background, lighting conditions, symmetry, texture, illustration, and appearance. The reviewed datasets can be used to evaluate the effectiveness of methods that work in the RGB theme modality. According to the review, the 3D visual understanding is a challenge for complex interactions between objects in terms of perspective, fully or partially chaotic internal environments, and scale changes in different scenes.

Lateef et al., [131] and Minaee et al. [166], have provided a comprehensive review of the literature of pioneering works for semantics and example level image division using over one hundred deep learning-based segmentation methods proposed in 2019 and 2020, respectively. Naseer et al. [176], created a review of advanced technology based on visual concepts, including visual classification, object identification, pose estimation, semantic segmentation, 3D reconstruction, salinity detection, physics-based reasoning and internal visual skills.

In addition, a recent comprehensive review was presented by Rahman et al. [198], where they reviewed the latest 3 DODTs (3D object detection technology). This review maintains some common steps, including descriptions of some popular public datasets, several performance appraisal metrics, and 3D BB techniques. They focused on cutting-edge technology in the 3DOD sector with their significance, contributions and future directional flaws. Zaixing et al. [88], discussed several approaches for 6D pose estimation in their review, including the advantages and disadvantages. A further up to date survey for 3D object understanding, classification, identification, defining size and shape, and tracking with 3D visualization and segmentation is present by Guo et al. [79].

Additionally, when listing recent approaches, we ignore traditional solutions to offer up-to-date reviews. Our survey paper looks back at later high profile research publications from a variety of perspectives on object detection and pose estimation. At the end of our survey, we proposed some new insights. In short, as of June 2021, this survey summarized and discussed more than 300 high profile states of art techniques (most of them the most recent). We have tried to make this review paper exceptional and comprehensive than other existing reviews by presenting the graphical outlines of the currently relevant papers. Also, we mentioned the future directions given by multiple authors and aim to make a decision based on them. This survey will help researchers (from start to end) who want to work with 3D object detection or 6D pose assessment.

C. UNIVERSALITY AND UBIQUITY OF DEEP LEARNING IN 3D OBJECT DETECTION (3DOD) SYSTEMS

One of the critical and mandatory tasks for developing computer vision (CV) in the autonomous field is 3D object detection. Driving without a driver, for example, requires an authentic representation of 3D space around autonomous vehicles of various important categories (prediction, planning, detection and speed control). Although LiDAR point cloud has proven successful for accurate 3D object detection, it is weather sensitive and expensive. Although the concept of monocular 3D object detection (mono3DOD) from RGB or RGB-D image is not a fancy concept, it still differs immensely from the LiDAR-based approach.

In order to detect a 3D object and guess the pose, we need to fully understand an image, rather than just knowing the classification or image localization. 3DOD is a significant work that can be broken down into several subtasks to make important steps for accurate knowledge of images and videos, such as some other notable applications are classification [109, 122], human behavior analysis [25], pedestrian detection [52], skeleton detection [120], face recognition [298] and autonomous driving [32].

There are some significant hurdles in achieving the identification and object localization tasks such as occlusions, chaotic environment, lighting conditions, size differences and viewpoints. Due to the notable impact of accurate object detection in robotic and autonomous fields, more efforts are being made to identify a (3D / 2D) object more accurately with intense care and attention [75, 76], [201, 202]. 3D object identification can be divided into object localization (specific content located in a test image) and object classification (category by object). Conventional 3D object detection models can be divided into three main categories: informative zone selection, feature extraction, and classification.

Any given image can have multiple objects in different positions of the image with different aspect ratios or sizes; It is best to handle the whole image with a different image sliding window. Strategies attempt to identify all possible positions and orientations of objects. Due to a large number of test windows, the process is comparatively expensive and generates additional windows. Moreover, if some stable sliding window template is applied, it will create unsatisfactory areas.

Some important steps to detect object:

- **Feature extraction:** This step helps to identify diverse objects and reveal features with meaningful and strong representations about complex cells as neurons in the human brain [156] such as HOG [44], Haar-like [144] and SIFT [156]. However, due to the varied lighting conditions, it is challenging to accurately describe all kinds of things.
- **Classification**: A classifier has to differentiate the target object from different types to create recognition of more semantic, categorized, and informative ocular objects. The common classifier used for classifications is SVM [40], AdaBoost [67], DPM (Deformable Part-bas ed Model) [63] (more flexible for low level features).

A state of the art results has been achieved in the Pascal-VOC [61] object identification competition by applying the concept of describing local features. However, there were some issues with this model, such as inaccurate bounding boxes, inefficient and unwanted low-level descriptors, and improperly trained models. These earlier object detection problems were overcome with the emergence of the Deep Neural Network (DNN) [122]. Eventually, identifying and detecting 3D objects from 2D images is a difficult task. The task becomes even more challenging as the level of depth of the 2D image during formation. Nevertheless, it is possible to identify 3D objects from 2D images with some efficient proposed methods.

D. UNIVERSALITY AND UBIQUITY OF DEEP LEARNING IN 6D POSE ESTIMATION SYSTEMS

To detect 3D objects from monocular 2D RGB images, we need to create a 3D oriented BB (bounding box), while 3D reasoning from a single 2D input is a complex and difficult task. In the autonomous sector, other than object detection, pose estimation is a complex job that needs to be done. It is easier to predict the 6D pose in RGBD images than in RGB images because the 6D pose is a complex combination of 3D rotation of an object (raw, pitch, yaw) and 3D coordinates (X, Y, Z) at the camera focal point [151]. One significant step in identifying the 3D object and estimating the 6D pose of any object from the image can be divided into egocentric and allocentric positions [118]. In the context of autonomous driving, the orientation related to the camera is called egocentric, and the orientation related to an object is called allocentric. Also, full 6D pose estimation is required for successful implementation of AR (augmented reality) [162], robotics grasp [38], autopilot [32], and so on.

Recent improvements to visual depth sensors and the availability of low-cost depth data have significantly improved object pose estimation. In addition, successful implementation of 6D pose estimation method to solve some problems such as variability of viewpoint, similar objects, symmetrical property, occlusion and cluttered environment; All have been overcome due to the availability of RGB-D sensors and the recent improvement of the Convolutional Neural Networks (CNN).

Typically, the recovery of a 6D pose estimation depends on two factors, the familiar instances and the raw/unknown instance of an object. Moreover, some challenges such as shape mood, target domain, shift distribution between several sources, and classification of objects prevent calculating the pose accurately. These challenges have been widely studied in recent years because of their significance in augmented reality (AR) [162], robotics [250], and autonomous vehicles [73]. In the robotics and automated car industries, accurate object detection, the successful application of self-management of objects (robotic groups), and the assumption of 6D poses by robots play an important role in advancing the challenge of autonomous manipulation.

E. THE PAPER COLLECTION PROCESS

Google Scholar is one of the primary sources of our paper collection. Also, the well-known Database "Web of Science " is another notable source through which we have introduced and collected a number of related papers. In addition, we should mention "Wikipedia", which is an authentic source of information and documentation. YouTube plays a vital role in understanding any new concept in this case. UTAS (University of Tasmania) Open Access Repository [183], [182] is a great choice for collecting recent papers.

The keywords we have used for searching references include deep learning (DL), deep neural networks (DNN), Convolutional neural network (CNN), object localization, image processing, autonomous vehicle (AV), 2D/3D object detection, 2D/3D bounding box (BB), 2D/3D object proposal, 2D/3D object identification, and 6D pose estimation. In addition, ACM Digital Library, IEEE Explore, Scopus, ScienceDirect is a collection of the best research databases that make our survey resourceful. Last but not least, Researchgate is a legitimate source of information and paper. Most importantly, we have gone through some highly ranked conferences such as CVPR, ICCV, NIPS, AAI, ICLR, ECCV, ICRA, ICML, IV, IROS, ACM, ITSC, ICIP, TPAMI, IRS, WACV, ECCV, ACCV, and Sensors.

F. OUTLINE OF THE REVIEW:

This paper proposes a comprehensive study reviewing the current methods of object pose detection and recovery. Our contributions are as follows:

- Discussed computer vision and deep learning networks, autonomous car and its challenges in section II and III briefly.
- The datasets used for the 3D object detection and 6D pose estimation method were observed to identify its challenges, which are represented in Table 2.
- Discusses the range of state-of-the-art (SOTA) technology from 3D BB detectors to full 6D pose guessers in IV section.
- In Table 4 where some of the SOTA 3D object detection methods are compared, and this table is represented graphically in Figure 5. A similar type of comparison has been done in Table 6 for some SOTA 6D pose estimation methods where this table is represented figuratively in Figure 7.
- Open issues are discussed to identify potential future research directions in VI.,
- Finally, section VII sums up the present situation of the field and concludes the review work.

II. COMPUTER VISION AND DEEP LEARNING

A. COMPUTER VISION

Computer Vision (CV) in artificial intelligence trains computers to interpret and understand the visual world, working with technologies where computers can achieve a high-level understanding of any digital image or video. Therefore, the CV is responsible for generating a theoretical and algorithmic basis for achieving automatic visual comprehension. Artificial Intelligence theory (AI) is applied to computer vision to reveal information from images. Data from this image can be obtained in many forms, such as single or multiple images from cameras, videos, or multiple images from a treated OCR (optical character lesson).

The CV system is a method of taking, processing, exploring and mastering digital images and using models built with the help of geometry, statistics, physics and some teaching theories to generate numerical or symbolic data from those images [107, 119, 169]. The CV process is gradually seeing new revolutionary concepts related to object detection where the main challenges are image processing and machine vision [192]. This rapid improvement of the algorithm enables the successful implementation of CV technology in various sectors and leads to the advancement of autonomous technology.

Moreover, a self-driving vehicle is a notable example where ANN and CV have been widely used. However, it is a big challenge for autonomous vehicles to accurately estimate the position of a 3D object from a 2D image. Although much progress has been made in identifying 2D objects from an image or video, identifying a 3D object and determining the 3D properties of an object from a single image is still a challenging problem.

Typical tasks of computer vision: Content-based image retrieval [229], Pose estimation [268], Optical character recognition (OCR) [164], 2D code reading [205], Automatic face recognition, Recognition Features [66], Egomotion [15, 302], Optical flow [10]. Scene reconstruction [232], Image restoration [7], Image acquisition [45], Feature extraction [45], Detection/segmentation [154], High-level processing [45], Decision making [45].

B. ARTIFICIAL NEURAL NETWORK (ANN):

The function of ANN is almost the same as that of the human brain, as knowledge is acquired through the network through a learning process from near it and stored using some synaptic weight neurons. To achieve the final design goal and change the synaptic weight of the network, NN has implemented a learning process known as a learning algorithm. Nowadays, ANN has been applied to multiple jobs, including computer vision, image recognition, speech recognition, social network filtering, machine translation, diagnostics, and video games [71, 177].

C. DEEP NEURAL NETWORK (DNN):

Deep Neural Network (DNN), a section of a machine learning (ML) where the machine has to predict any output, can be supervised, semi-supervised or unsupervised [218]. Since traditional ML techniques cannot process natural data in their raw form, DL (Deep Learning), an advanced DNN technique, applies multiple layers to reveal high-level features from the raw data. For example, in image processing, where the lower layers of the DL model can recognize the edges only, the upper layers can detect a certain number of letters or objects or features of the object [238].

Eventually, DL processed unsorted/sorted, labelled/unlabelled data and construct a pattern to make a better prediction [119, 122, 175]. Though DL was popular since 1980-90s, offered the concept of the back-propagation classifier [206]; nonetheless, it soon lost its popularity due to over fitting, scarcity of big data, and poor computation capacity as compared to other ML tools.

The popularity of Deep learning algorithm has increased since 2006 [93] with the advancement in speech recognition [92] application. Convolutional Neural Networks (CNN), the popular DL framework, which is applied on multiple sectors such as Natural Language Processing (NLP), Computer Vision, Speech Recognition, Audio Recognition, Machine Translation, Social Network Filtering, Bioinformatics, Medical image analysis, and much more [170].

1) Convolutional Neural Network (CNN)

The Convolutional Neural Network (CNN) has a deep feed-forward architecture and remarkable ability to generalize to better networks with fully connected layers. CNN has largely applied to image analysis, especially pattern recognition, which can also be employed to solve other data analysis problems, such as classification problems. CNN is a deep learning network developed for image and video processing that has made significant progress since 2010 and is now widely used worldwide.

Usually, the CNN structure consists of multiple layers, including input, convolution, pooling, fully connected, softmax, and output layers. In the convolution level, the filter is used as the navigator above the image. This filter then navigates on top of an image and counts in pixels where the filter is located. Furthermore, CNN is the most interdisciplinary model of the deep learning method, where each layer is called a feature map. Multiple filters are used on a CNN network, and the most accessed feature map is sized according to the property of the filter. In contrast to the traditional theoretical approach, CNN's advantages can be described as a classified feature presentation.

The two most notable qualities as classified composition and the ability to extract powerful features from an image prove that CNN is one of the most powerful object detection classifieds. Several important CNN architectures have been proposed times for image processing such as ImageNet [47], AlexNet [122], ConvNet [123, 133] LeNet [242], VGGNet [227], ResNet [86], ZFNet [283], GoogLeNet [242], GPU (Graphics Processing Unit) processor large-scale distributed clusters [46], and OverFeat [217].

On top of that, CNN is a powerful algorithm that is widely used for image classification and object detection [122, 283]. Because of the notable advantages, CNN has been widely applied in many research fields including image super-resolution reconstruction [178, 284], image classification, image retrieval [109], face recognition [298], pedestrian detection [248, 271, 293] and video analysis [227, 269, 283], car detection [32], and pose estimation [295].

Most importantly, CNN can be adequately trained that does not suffer from over fitting and is easy to apply to large networks [122]. However, CNN cannot provide accurate results when the length of the output level is variable and the presence of objects of interest is not fixed. Therefore, more sophisticated algorithms such as R-CNN, Fast R-CNN and YOLO have been developed to solve advanced image processing problems.

2) Region Based Convolutional Neural Network (RCNN)

Girshick [75] proposed a method where a large number of regions were selected, and the Selective Search (SS) [252] method was applied to select only 2000 regions from an image, which he named the region proposals. First, in R-CNN, the image is divided into about 2000 regions, and then CNN (Convent) is applied to each region gradually. The size of the regions can be identified, and the exact region is inserted into the artificial neural network. From now on, instead of categorizing broad regions, we can use 2000 regions. These 2000 regions are then curved into quadrilaterals and processed into a CNN, resulting in a vector featuring 4096D. Since each region of the image is applied to CNN individually, the training time is about 84 hours, and the forecast time is about 47 seconds. As a result, the process becomes time-consuming because it has to classify 2000 region propositions for each image. Here, the CNN functions as a feature extractor and the revealed features are processed through an SVM [40] classifier to distribute the object inside the region proposal. Additionally, to anticipate the region proposals and increase the bounding box's precision quality, the algorithm creates four offset values. The main problem with this classifier is time.

3) Fast RCNN

The algorithm previously proposed to create a quick object recognition classification updated some of the errors of R-CNN and renamed as Fast R-CNN [75]. This proposed method is almost the same as the R-CNN classification. It feeds the region propositions and provides the processed image to CNN as an input, which creates a convoluted feature map. So the difference between the Fast R-CNN and R-CNN is that the former does not divide into official region recommendations but first applies CNN and then allocates it to region recommendations. Fast R-CNN completes significantly faster than R-CNN for both training and testing sessions. In fast R-CNN, the convolutional operation is performed once for each image, and then a feature map is created from it. Since it uses CNN once, there is a significant gain over time. The training time is about 8.75 hours, and the estimated time is about 2.3 seconds.

4) Faster RCNN

Both of the above algorithms (R-CNN and Fast R-CNN) use SS to determine region proposals. SS [252] is a slow

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3114399, IEEE Access

and time-consuming process that over-segmenting the image affects network performance. Therefore, Shaoqing Ren et al. [202] proposed an object identification algorithm that removes the SS algorithm and allows the network to learn region recommendations. After the predicted regions are resized using the ROI(Region of Interest) pooling layer, which is then used to classify the image in the proposed region and predict the IoU (Intersection-over-Union) ratio of the bounding boxes.

5) Single Shot MultiBox Detector (SSD)

Liu et al. [152] proposed SSD (Single Shot Multibox Detector), a single shot detector for multiple segments, applies an additional small conventional filter to maps that are faster and significantly more accurate than previous single shot detectors like YOLO.

6) Mask RCNN

He et al. [85] presents the concept of flexible structures called Mask R-CNN for object instance segmentation. This method effectively recognizes objects from an image while creates a high-quality segmentation mask for each instance at the same time. Mask R-CNN is a practical extension of Faster R-CNN, where an additional branch is added to predict an object mask parallel to an existing branch. Moreover, this method is a slightly improved version of R-CNN that runs at 5fps and can adapt quickly to predict human posture. Also, Mask R-CNN has won the COC 2016 Challenge by overcoming three key issues: Instant Segmentation, Bounding-Box Object Identification, and Individual Keypoint Identification.

7) YOLO

Redmon et al. [201] Proposed a novel object detection technique called YOLO (You Only Look Once), where the classifier does not process the whole image; Instead, it focuses partly on the image with a high probability of having the object in that part. This single convolutional network is faster than existing object detection algorithms. However, above all advantages, the YOLO algorithm struggles to detect small objects within the image. For example, the spatial limitations of algorithms can make it difficult to identify flocks of birds. Some other notable DL structures are: RefineDet [286], Retina-Net[146], Deformable convolutional networks [43], Cascade R-CNN [20], 3D-RCNN [127], Libra R-CNN [185].

8) Mesh R-CNN

Facebook introduced a novel RCNN method in artificial intelligence called Mesh R-CNN that can convert 2D objects to 3D shapes and mesh [266]. Facebook has highlighted its latest advances that can identify complex issues. This study has applied in-deep learning to understand the 3D shapes of complex objects and novel architectures such as Bounding Box, 3D Voxel Pattern, Point Cloud and Message for prediction and localization. Mesh R-CNN can effectively detect and classify objects in 3D form from chaotic 2D images and occluded objects and ultimately estimate their full 3D shape.

TABLE 1. Key-features of some state-of-the-art Deep Neural Network framework.

No.	Algorithms	Features
1	CNN [283]	Impressive classification performance.
2	R-CNN [75]	Robust architectural structure [134]
3	Fast R-CNN [76]	Has additional sub-network.
4	Faster R-CNN [202]	Improved outcome by applying the RPN.
5	R-FCN [42]	Manages to address the dilemma. [86].
6	Mesh R-CNN [266]	Novel architectures for 3D shape [28] .
7	YOLO [201]	Improved model for the fixed-grid regression.
8	3D-RCNN [127]	3D scene understanding to map image.
9	Libra-RCNN [127]	Re-balances by mixing : IoU, feature and L1 loss.
10	Cascade-RCNN [20]	It is extensively contextual for detection.

III. AUTONOMOUS VEHICLE (AV):

An autonomous vehicle (AV) is a combination of some actuators, sensors, complex analytical algorithms, machine learning methods, and high-speed processors that are needed to implement complex software. Self-driving vehicles create and maintain a map called Simultaneous Localization and Mapping (SLAM) of their surroundings by multiple sensors placed in different parts. Such as LiDAR or radar measures distances of other vehicles or obstacles, detect road edges. In addition, one or more cameras mounted on autonomous vehicles can detect traffic lights, road signs, lane signs, vehicles, obstacles and pedestrians [51].

During parking, ultrasonic sensors placed on wheels to detect obstructions and other vehicles. Advanced complex software [130] then processes all these sensory inputs, generates outputs and sends commands to the vehicle actuator responsible for steering, braking and control acceleration. AV can be identified as a complete package of hard-coded rules, a complex algorithm and efficient predictive models, helping sophisticated software to run on the road smoothly [72, 191, 244].

To date, autonomous vehicles are equipped with two types of sensor such as active sensor: LiDAR [137, 138, 278] , radar [276] and passive sensors: Single/Stereo cameras [31, 171], and their fused systems [32, 106], short-range sensor (Ultrasonic sensors) [121]. Veli et al. [103] made lots of progress in sensor technology and GNSS (Global Navigation Satellite Systems).

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3114399, IEEE Access

A research team from the Massachusetts Institute of Technology (MIT)[83] announced in May 2018 that they had successfully built a driverless car that could successfully navigate unmapped roads with a novel system known as MapLite [39]. This application enables the driverless car to drive on a completely new road without using pre-loaded 3D maps. The basic idea is to combine the vehicle's position with sensors that monitor the surrounding conditions, and OpenStreetMap (OSM) is used to detect the GPS of a vehicle [39].

Also, an AV has been divided into 5 levels such as level 1 - requires driver support, level 2 - partial automation phase, level 3 - limited driver support, level 4 - higher automation and level 5 - fully automated [224], [83] [189]. At present, level 3 autopilot is available on the road, as Level 4 and Level 5 autonomy require large-scale neural network training and visual recognition, including accurate pose estimations. Multiple companies produce intelligent vehicles and test them to drive autonomously in certain situations, such as Tesla Autopilot, Waymo, Uber, Volvo, Google, BMW, Mercedes Benz, Nissan and General Motors. However, they are still in the testing phase and unable to operate without assistance.

A. TECHNICAL AND SOCIAL CHALLENGES OF AUTONOMOUS VEHICLES :

Although the concept of autonomous or self-propelled vehicles has come a long way in recent years and numerous studies have been done in this sector, this technology is still not flawless. Lawmakers and consumers still feel confused and anxious about implementing self-driving cars and feel insecure and uncertain about the autonomous vehicle's ability to move freely. So self-propelled cars are still in the experimental stage, and more research is needed to perform them properly. One of the significant challenges of autonomous vehicles is accurately estimating the exact position and orientation of nearby vehicles. The five core reasons are classified as why the AV still are not on the roads are listed in below:

Sensors: An autonomous vehicle faces various challenges for smooth automation systems such as proper vehicle navigation system, GPS, environmental perception, LiDAR and radar, visual perception, speed and direct perception and a vehicle control system [191, 292]. Furthermore, to be qualified as perfect autonomous vehicles, these sensors need to work in all weather conditions anywhere on the earth. Undoubtedly, the critical issue for driverless vehicles is that there should be a control system capable of automatically analyzing sensor data and making accurate estimates of vehicle postures, obstacles, pedestrians and road signs [305].

Machine learning Algorithm: At the moment, there is no widely approved and authorised ML algorithm to ensure that they are 100 % error-free, safe and secure for use in any driverless vehicle. The most popular algorithm applied to current driverless vehicles is SLAM, which integrates data from various sensing components and uses offline maps [265]. WAYMO has improved the performance of the algorithm SLAM and named DATMO (Detection and Tracking of Moving Objects), which can handle any curbs, including vehicles and pedestrians. Zhang et al. [290], proposed a concept that collaborated with the existing Visual odometry (VO) system such as SLAM and ORB-SLAM2 (an updated version of the SLAM) [173].

The open road : When the AV drives on new roads, it should identify things that did not come before in the training process and may be subject to software updates. As a result, it would be not easy to ensure that the system is as secure as its former version.

Regulation To date, adequate standards and regulations for autonomous systems do not exist. There have been numerous high-profile accidents involving Tesla's current automobiles, as well as other automotive and autonomous vehicles [9].

Social acceptability: Applying a automatic car on the road is not only a problem for those who want to buy and use a driver-less vehicle, but also for others who share the road with them [9].

IV. LITERATURE REVIEWS

An essential part of computer vision is the identification of objects from images or videos. Object detection helps in pose estimation, vehicle detection, pedestrian and other curb detection. Previously, the image was processed and classified using traditional machine learning (ML) algorithms such as colour histogram [219], SVM (Support Vector Machine) [40], logistic regression [201]. However, there are some differences between the recent object detection algorithms (CNN, R-CNN, YOLO) and traditional ML classification algorithms (SVM, logistic regression).

The definition of object identification problem determines where objects are located in a given image is called object localization, and what class each object belongs to is called object classification. Thus, the traditional thematic object detection model's pipeline is divided into three stages described in FIGURE 1.

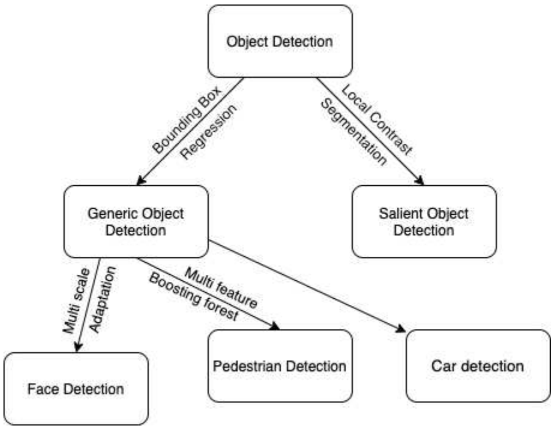

FIGURE 1. The application domain of Object detection

1) Informative region selection: Different objects can appear in any position of the image and have different aspect ratios or sizes, so scanning the entire image with a multi-scale sliding window is a natural choice.
2) Feature extraction: To identify different objects, we need to figure out visual features that can represent a semantic and robust.
3) Classification: A classifier needs to differentiate a target object from all other categories and further classify the presentation.

A. 3D OBJECT RECOGNITION

Object recognition is one of the primary pillars of a computer's vision and is sometimes confused with the problem of object classification/shape retrieval. 3D object recognition methods can be divided into two main categories such as voting methods, Hough transform [6], and geometric hashing [129] and the correspondence based method, spin images [111], local feature histograms [89], 3D shape and harmonic shape context [68].

David et al. [155] developed an object recognition system using local image features in cluttered real-world scenarios. Cordelia et al. Schmid et al. [213], has shown that recognition of successful objects can often be achieved by applying a sample local image descriptor to a large number of repetitive locations. Papazov et al. [186] proposed the recognition of a 3D object, especially for noisy and scattered data in cluttered and occluded environments. This proposed concept applies a combination of strong geometric descriptors, a hashing technique and a sampling technique - RANSAC [64].

B. 3D OBJECT DETECTION FROM RGB AND RGB-D IMAGE

3D object detection is a significant key part of the visual perception system of robotic and autonomous technologies.

1) Feature extraction, Segmentation and Matching

Rapid and accurate image segmentation with feature extraction is the primary task of the computer vision field. Lowe et al. [155] proposed SIFT (Scale Invariant Feature Transform), an object recognition system for image scaling, translation, matching and rotation, and a partial constant for illumination changes, including 3D projection. The images here have been converted to a wide collection of local feature vectors and can generate approximately 1000 SIFT keys in 1k ms during each image count by applying classification. Although occlusion may be present in the image, SIFT can provide a high level of accuracy.

Kang et al. , [113] Created a structure called DaSNet-V2 that matches identification, category, localization, and object instances. A method capable of achieving real-time performance by adopting PWP 3D (count per pixel) and applying the region-based simultaneous strategy of 2D partitioning using the NVIDIA CUDA framework is largely developing parallel algorithms [193]. Fu et al. [69] introduced DORN (Deep Ordinary Regression Network), a multi-scale network framework that achieves a spacing-increasing discretion (SID) strategy to rebuild depth and depth networks to reduce the complexity of existing feature maps.

2) Shape variation

Xiang and Dollar offered 3DVP (3D Voxel pattern), which uses ACF (Aggregate Channel Features) detectors to find out the basic features of each object such as shape, appearance, aspect and curbs [53, 270]. In addition, the 3D pose of a vehicle can be accurately localized from the context of this method and can detect other vehicles and guess the pose [73].

Novotny et al., [180] have created the C3DPO (Canonical 3D Pose) Network for non-rigid structure motion where no training images and messes are available. It has partially reconstructed a 3D object from a monochromatic RGB image to change perspectives and distort the object. It has also emphasized the mandatory presence of certain canonicalization functions of reconstituted size and shape. The input depth proposes objects pose according to the classification of convex hulls that align the clusters of convex sections drawn from the images. This is an example of a highly efficient size identification pipeline that uses the CHAL (convex hull alignment) algorithm for hypothesis generation and is used to identify objects in complex scenes with multiple objects [41].

Qian et al., [195] presented a method for evaluating individual 3D sizes, where there was a balance and robustness between the accuracy and efficiency of the conventional stage recovery method, significant measurement limits and high-frequency fringe patterns. Chabot [27] made a framework called Deep MANTA for 3D object detection based on a single-dimensional image in an end-to-end fashion network, determining the object class, 2D region proposal generation, 2D location, orientation, dimension and 3D position. This model has implemented a 3D vehicle dataset featuring 3D mesh with real size to match vehicle parts (wheels, headlights, mirrors) and defines a 3D shape for each 3D model. Zhou et al., [303] has built the CenterNet framework, which is simpler, faster, and more accurate than traditional BB detectors and poses estimators.

3) 3D projections of the 3D bounding box vertices

Chen et al., [33] proposed 3DOP (3D Object Proposal) for accurate object class identification in the context of autonomous driving. 3DOP produced several sets of 3D candidate boxes to identify almost every object in 3D space. This method has featured object size, ground plane, different depths, spaces, the density of points inside the box, visibility and soil distance.

Mono 3D (Monocular 3D Object Detection) [31] uses ground planes and some segmentation features to generate 3D proposals from monocular images in the context of autonomous driving. In addition, both 3DOP and Mono3D methods applied some common hand-crafted features. This technique applies several intuitive potentials to each candidate box expected in the image plane encoding synthetic

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3114399, IEEE Access

segmentation, relevant information, size and location prerequisites, and ideal object sizes. Also, the S-SVM [110], structured SVM [251], parallel cutting plane [227] and IoU has been implemented with a comprehensive search model.

The proposed DSS (Deep Sliding Shapes) [235] is a 3D convergent formulation that takes 3D volumetric views as input from an RGB-D image and then outputs a 3D object bounding boxes. In addition, this method proposes the first 3D Region Proposal Network (RPN) to learn objects from geometric shapes and the first Joint Object Recognition Network (ORN) to extract geometric features in colour properties in 2D.

Ding et al., [48] proposed a fancy wire-frame model called the CPO (Cross Projection Optimization Method) that can detect both 3D pose and shape estimation of a vehicle for an autonomous vehicle. The CPO method applies a simple wire-frame model combined with the Hierarchical Wire-frame Constant (HWC) method instead of bounding box annotation to shape detection for 3D pose and accurate 3D localization [32].

The solution provided is primarily based on local properties, especially for matching objects in a 2D image of a rigid 3D object [78]. This method creates an accurate 3D model of the object with the locations of its features and then places it in an image to identify new features. Finally, the position, orientation, and shape of the virtual object are defined concerning the object's coordinates.

Rad [197] has created a framework where a total of 8 corners of the bounding box are applied to the multiple-input image called BB8. This method is trained to predict their poses in the form of 2D projections of the corners of their 3D bounding boxes and calculates 3D poses from this 2D-3D correspondence with a PNP algorithm [135].

Another strategy called Mono3OD [226] where a single RGB image uniquely transformed to reduce object detection and increase the credit count for 3D BBs. Li [141] suggested RTM3D (Real-time Monocular 3D Detection), the first real-time 3D identification method for autonomous driving, predicting the nine point-of-view of the 3D BB in place of the image and using 3D and 2D perspective geometry to restore orientation, location and dimensions in 3D space.

Liu [149] has claimed a deep fitting degree scoring network for mono 3DOD, which focuses on the active fitting degree among proposals and objects. It is discrete from other existing monocular frameworks by attaining localization by computing the visible degree of calculation among the 3D project proposals and the object. A concept named FQNet, [149], can assume the 3D IoU (Intersection over Union) among the 3D proposals and the object.

Zhang et al., [290] proposed a framework for 3D object detection by determining object class, 2D region proposition production, 2D position, position, dimension and 3D positioning based on a single image in an end-to-end fashion network. Furthermore, Bao et al. [8], recently introduced Mono-Fenet, a compelling feature enhancement method for the 3D object detection, which includes the ROI Mean Pooling layer, the PointFE network, and feature enhancement networks using 3D-NMS and exclusive RGB imagery .

The full 3D poses and dimensions of an object from a 2D BB by applying some restrictions to calculate the orientation and volume of the object using DCNN, where the novel DCNN method known as MultiBin regression is used to estimate the orientation of the object [171]. SS3D, a single-phase monocular 3D object detector where the 3D representation is returned by a representative and uses for the geometric shapes of the 3D box with autonomous driving [112].

Hu et al., [101], introduced a complete 3D vehicle bounding box tracking information method from exclusive videos and a method for dealing with 3D vehicle detection guesswork. A new pipeline based on LSTM [253] is designed to collect large-sized 3D trajectories from real-world driving environments and track 3D vehicles within 30 meters. The method called M3D-RPN implemented exclusive 3D identification and 3D zone proposal networks and lifting the geometric relationship of 2D and 3D perspectives, including 3D boxes [14].

4) 3D Object Detection in Point Cloud

Scientists proposed a method for identifying free-form 3-dimensional objects in point clouds with global representations [55] . The basic idea of the model is to create a universal approach statement based on the point pair factor. Free-form objects in 3D datasets can be achieved by a number of sensors, such as a laser scan, a TOF (Time of Flight) camera, which has been widely disseminated from a computer perspective [24, 159].

Chen et al., [32] introduced accurate 3D object detection for individual behaviour, known as Multi-View 3D Network (MV3D), which works with multimodal datasets. MV3D framework creates efficient 3D candidate boxes from a 3D point cloud BEV (Bird's Eye View) [153] image, and the main goal of this method is to identify 3D objects using both LiDAR and image data. Current LiDAR-based methods set 3D windows in 3D Voxel Grid [57, 259] or apply convolutional networks [138] to front viewpoint maps.

On the other hand, a hybrid method has introduced that combined both LiDAR and camera data for 2D detection to get accurate results [58, 77]. Qi et al., [194] offered a fancy concept called "Frastum PointNets" based on RGB-D data in a point cloud and expects a semantic class for each point in that point cloud. A method named PV-RCNN provides accurate 3D object detection from point clouds that deeply integrates 3D visualization with point-to-point set-based abstraction with a 3D visual convoluted neural network and multiple receiving fields [220]. Finally, a novel method called SAANet (Special Adaptive Alignment) uses an "SAA" module that addresses fusion-based deep structures that combine clouds and images for 3D object detection with complements cloud properties and image properties [30].

TABLE 2. Data-sets used for multiple 3D object detection and pose estimation method.

No.	Benchmarks	Frame	Format	Categories	Features
1	KITTI [73]	14999	RGB	3	>100 gb data
2	COCO [147]	328000	RGB	11	2,500,000 labeled instances.
3	SUN RGB-D [54]	10335	RGB- D	800	Combination of NYU, B3DO and SUN3D.
4	NYU [225]	1449	RGB	19	795 trained data.
5	PASCAL-VOC[61]	14999	RGB	12	Extension of PASCAL3D+
6	ImageNet [47]	>14 million	RGB	21	7481 trained and 80256 labeled objects.
7	RGB-D Object Dataset [128]	250,000	RGB-D	51	300 physically distinct objects.
8	Parsing IKEA Objects [145]	>1k	RGB	NA	3D models of IKEA furniture.
9	CVonline [65]	All	NA	NA	A rich database for CV, ML, and IP.
10	Yu Xiang et al. [274]	2640	RGB	4	Ratio between test and training images is 50%
11	NYC3DCars [163]	2299	RGB	20	Over 2000 annotated photos from New York.
12	EPFL Cars [184]	2000	RGB	20	Acquired from a car show.
13	ApolloCar3D (Kaggle) [237]	5,277	RGB (Monocular)	3	5 GB Data..
14	LINEMOD [90]	30899	RGB	13	7481 trained and 80256 labeled objects.
15	MULT-I [246]	1,100	RGB	13	Both cluttered and occluded.
16	OCC [13]	10k	RGB-D	20	cluttered, textured and texture-less, rigid and non-rigid objects.
17	BIN-P [54]	NA	RGB-D	20	The first fully annotated bin picking dataset
18	T-LESS [95]	> 49K	3	50	39K training and 10K test images.
19	RU-APC [203]	nearly 10,000 images	RGB-D	25	Low illumination
20	BOP [97]	Nearly 340k images	RGB-D	89	Combination of RU-APC,TUD-L, and TYO-L Datasets.
21	YouTube-BoundingBoxes (YT-BB) [200]	380,000 video segments	videos	23	YTBB is the largest human-annotated detection data set
22	ShapeNet [28]	3,000,000	3D models	3,135	Autonomous robots and vehicles.
23	YCB-Video dataset [275]	133K images	video dataset	21	Live RGB-D camera.
24	Yale-CMU-Berkeley dataset [21, 22]	2K images	RGB + RGB-D	NA	BigBIRD Object .
25	JHUScene-50 [36]	22520 images	RGB-D images	50	>20K labeled poses.
26	NOCS-REAL275 [260]	(275K training, 25K testing) images	RGB-D images	6	18 different scenes and 42 unique instances.
27	Falling Things (FAT) [249]	60k images	RGB	21	2D/3D bounding box coordinates for all objects
29	ObjectNet3D [272]	90,127 i images	2D images	100 categories	201,888 objects and 44,147 3D shapes
30	nuScenes [19]	1,400,000 images	390,000 LiDAR sweeps	23	Contains 100 times images than KITTI dataset.
31	RobotP [280]	4200000	RGBD	NA	synthesized photo-realistic color-and-depth images

TABLE 3. Evaluation metrics: Different types of evaluation metrics to identify and measure the performance of proposed classifiers.

No.	Evaluation metrics	Estimation	Function
1	Intersection over Union (IoU) [60, 235]	Assesses the 2D space performance.	$\text{W}_{I_o}U_{2D} = \frac{B\cap\overline{B}}{B\cup\overline{B}}$
2	Average Precision (AP) [60]	Assesses the shape of the Precision-Recall (PR) curve.	AP = 1 $\frac{}{11}\sum_{r\in 0,0.1,..,1} pinterp(r)$
3	Average Orientation Similarity (AOS) [60, 73]	Applies the AP metric for object detection.	AOS = 1 $\frac{}{11}\sum_{r\in 0,0.1,..,1} max_{\widetilde{r}:\widetilde{r}\geq r} s(\widetilde{r})$
4	Average Viewpoint Precision (AVP) [273]	Generats a VPR (Viewpoint Precision-Recall) curve.	NA
5	Translational and Angular Error [55]	Measure the ground truth angel and translations and rotation of an object.	$\text{W}_{TE} = \Vert X - \overline{X} \Vert_2$
6	2D Projection. [273]	Estimated poses by projecting the model's vertices onto the image plane.	$\text{W}_{2Dproj} = \frac{1}{\vert V\vert}\sum_{v\in V} \Vert \text{C}\overline{R}_v - CR_v\Vert$ $W_{RE} =$across$(Tr(R\overline{R}^{-1} - 1)/2)$
7	Average Distance (AD). [90]	Remove obscurities from the symmetry property and occluded background.	$\text{W}_{AD} = avg_{s\in M}\Vert(\overline{R}s + \overline{T})(Rs + T)\Vert$ $\text{W}_{AD} = avg_{s_1\in M} min_{s_2\in M}\Vert(\overline{R}s + \overline{T})(Rs + T)\Vert$
8	Visible Surface Discrepancy (VSD) [96]	A method that remove ambiguities.	$\text{W}_{VSD} = \begin{cases} 0 & \text{if } p \in \overline{V}1 \\ \quad\text{otherwise} & \end{cases}$
9	Sym Pose Distance. [96]	Fixed the ambiguity due to the symmetrical shape of an object.	$\text{W}_{Sym} = min_{G_1,G_2\in G}\sqrt{\frac{1}{S}\int_s \Vert T_2\circ G_2(X) - T_1\circ G_1(X)\Vert^2 \times ds}$
10	Average 3D precision (A3DP) [236]	Inspired by the AVP metric (for both 3D shape and 3D pose)	NA

TABLE 4. Advantages and disadvantages of some state-of-the-art 3D Object Detection Techniques.

No.	Algorithms	Advantages	Disadvantages	DataSets
1	**3DVP [270]**	Has strong accuracy in occluded and complex scenes.	Its overlapping detection graph is often very complex.	KITTI [73]
2	**C3DPO** [180]	3D reconstruction and object deformations capacity.	Requires expensive hardware.	PASCAL3D+
3	MV3D [32]	High precision accuracy.	Computationally expensive.	KITTI
4	3DOP [33]	Highly accurate object proposal skills.	Average performance on cluttered environment.	KITTI
5	Mono3D [31]	High-performance in monocular imagery with better recognition rate.	Detection GRAPH is often very complex	KITTI
6	CPO [48]	High precision accuracy.	Overlapping in 3D shape and pose.	4 Real World Benchmark.
7	DSS [235]	Reveal powerful 3D and color features from the data.	NA	KITTI
8	SSD [152]	High-accuracy, high speed and very robust.	Confused with similar categories and worse performance on smaller objects.	SUN RGB-D andNYUv2
9	3D-SSD [30]	Significant accuracy and computation efficiency.	Receptive field is limited.	SUN RGB-D andNYUv2
10	SAANet [30]	Higher Precision and inference time in LiDAR + Img-based methods	Fails to localize and explore local orientation information.	KITTI
11	Deep MANTA [27]	Less time consuming and overcomes loss of information.	NA	KITTI
12	MonoDepth [290]	Perform very good especially for finding smaller objects	NA	KITTI
13	Mousavian et al [171]	Better performance for large dataset.	NA	KITTI and Pascal 3D+

5) Speed / Accuracy Trade-off

Huang et al., [102] introduced a process that helps determine the speed and accuracy of the calculation and also recommend which method is better suited for a specific application. Shrivastava et al., [223] proposed a TDM (top-down modulation) approach to include image quality for a ConvNet architecture such as VGGNet [227], ResNet [86], and Inception-Resnet [241]. Song [235] proposed Deep Sliding Shapes (DSS) that convert an RGB-D image into a point cloud and then slides a 3D detection window into 3D space. Luo [157] made a concept that identifies 3D objects and accurately predicts the position, size, orientation and division of objects in 3D space at very fast speeds. Li [139] has come up with an idea called GS3D, a 3DOD method based on an RGB (single) image in autonomous driving.

6) Object detection by key point estimation.

The most famous classifier that detected an object using key-point inference (identifies the object as a point to the key) is Cornernet [132], ExtremeNet [304], and CenterNet [303]. In CornerNet, the corners of 2D BB are used as semantic key points. ExtremeNet, on the other hand, highlights all points, including the top, left, bottom, right, and centre of the bounding box. Compared to these classifiers, the Centernet is much faster, which only chooses the object's centre.

V. LITERATURE REVIEW OF 6D/6DOF (DEGREE OF FREEDOM) POSE ESTIMATION

In the computer vision sector, guessing a 6D pose of an object is a significant problem that needs to detect both 3D orientation and 3D position of an object in the case of the camera centred coordinates [115]. In short, the three factors for the 6D pose estimation are the critical role of rotating left and right on the X-axis (roll) side as well as on the Y-axis (pitch) and tilting backwards on the Z-axis (Yaw). Thus, these features encourage concentration on the recovery of vehicle posture and size estimates to enhance the intelligence of the intelligent transport system and the robotic sector. Therefore, the conventional states of industrial techniques of 6D pose estimation are discussed here in the context of the autonomous car.

A. 6D POSE ESTIMATION DIRECTLY FROM RGB IMAGES.

Wu et al., [268] proposed an algorithm named 6D-VNet, and won the first place in the "Apolloscape Challenge 3D Car Instance" competition. It is an abstract structure for autonomous vehicles assuming 6 DOF object poses that can detect all aspects of traffic in a single RGB image while rotating vectors and 3D translation. The basic technique of this method is to control the 6D position of the vehicle using the outputs from the RPN (Region Proposal Network) [76] and 2D object detection network (Mask R-CNN) [85] that can learn both rotation and translation by outlining a loss function model.

Brachmann et al., [12] created a template-based model for calculating 6D pose for a specific object from a single RGB image. The algorithm optimizes the power following the RANSAC concept for a large and uninterrupted 6D pose space. The technical feasibility of classification is using a new composite dense 3D object coordinate form, including object class labelling. Kehl et al., [115] developed SSD-6D, a CNN method to detect the 3D object and accurately guess the 6D pose from an RGB image. It is a unique detector method for relevant training on synthetic model information, which applies to the collection of small objects and objects with many conceptual and practical advantages.

Inspired by BB8 [197] method Zhang et al., [288] re-imposed the coordinates of the image and applied the Perspective-n-Point (PNP) [135] algorithm without any post-refinement. Similar to recent work, the method uses 2D BB to calculate the coordinate regression of images based on their centres, focusing on the gap between image classification and pose estimates [247]. Deep-6DPose is an end-to-end deep learning solution, which finds objects and compresses them and retrieves instances of 6D objects from single RGB images [49]. It consists of two main components, such as RPN and a mask R-CNN, including Lie algebra.

Billings et al., [11] has developed a new proposal to predict 6D object poses from monocular RGB images by applying the CNN pipeline with the ROI proposal. It predicts the presence of intermediate outlines for 3D objects, 3D orientation and 3D translation vectors. For the 6-D category level pose estimation, two-level BB-based alternative methods have been developed that directly output the 6D pose without the use of any PNP but consist of ResNet (Residual Neural Network), RPN, and FCN (Fully Convolutional Network) [148].

B. POSE FROM DEPTH / POINT CLOUD METHOD

Mitash et al., [167] advocated a concept for efficient object 6D pose estimation in cluttered scenes, where the Cartesian product of the candidate's post for interactive objects is used to identify the best view and create an efficient search, and the candidate post clusters for each object. The MCTS (Monte Carlo Tree Search) technique is applied to conduct tradeoffs in fine-tuning and explore new instances.

Xiang et al., [275] has created a generic structure called POSNN that calculates the 3D translation of an object in the image and predicts its distance from the camera. Furthermore, this method reduces the ShapeMatch-Loss function and enables POSNN to handle symmetrical objects where the VGG16 backbone is used to extract features.

The PointPoseNet classifier for 6DoF objects gives the idea of inference of rigid objects using deep learning in point clouds. A point-to-point correspondence assignment is performed with a joint classification and segmentation within a point cloud system [82]. Capellen [26] suggested that ConvPoseCNN has evolved from the concept of PoseCNN but can avoid cutting individual objects. Instead, it offers accurate predictions for pixel-based translation of object

TABLE 5. 3D Object Detection Paper Collection

No.	Approach	References
1	Classification	SVM [40], S-SVM [110], [110], [251] , [99], [4],
1	Representation Transformation (Pseudo-LiDAR, BEV)	MLF [27] , Pseudo-LiDAR [263], Pseudo-LiDAR++ [279], Pseudo LiDAR-e2e [264], pseudo LiDAR color [158], ForeSeE [262], BEV-IPM [117], [16], SAANet [30], PIXOR [278], [81] .
2	Keypoints and Shape	Deep MANTA [27] , 3D-RCNN [127], Mono3D [31], ROI-10D [160], MonoGRNet [196], ApolloCar3D [236], RTM3D [141], [214], [273], [34], [299], [41], [258].
3	Distance via 2D/2.5D/3D constraint	[158] , 3D-RCNN [127], Mono3D [31], ROI-10D [160], MonoGRNet [196], ApolloCar3D [236], 6D-VNet [268], GPP [199], RTM3D [141], [48] , [78], Deep3dBox [171], Shift R-CNN [174], GS3D [139], [277], [168], [102].
4	Feature extraction and Matching	[27] , [78], [12], Deep Sliding Shape [235], [136], MVTec ITODD [56], [13], [155], [204], [113], [301].
5	3D object proposal methods	3DVP [270], 3DOP [33], Mono3D [31] , MV3D [32] , Deep Sliding Shapes [235], 3D voxel grids [57, 259, 270] , CPO [48], Joint mono3D [101], SS [252], SS3D [112], CasGeo[62], M3D-RPN [14], MonoDIS [226], [101], CenterNet[303], MLF-Mono [8], MonoDepth [290], [143].
7	3D Object Detection in Point Cloud.	[55], [161], [13], [159], [220], [137], [194], PV-RCNN++ [221], SE-SSD [297], SVGA-Net [87], CIA-SSD [296], PC-RGNN [291], [261].
8	3D Object Detection from RGB-D IMAGE.	DSS [235], SS [234].
9	Speed /Accuracy Trade-off	[102], [223], [243], SSD [152], 3D-SSD [157].

poses and orientation modules and has been replaced with a complete CNN prediction network. Also, [190] recently removed the ROI pooled orientation layer and introduced PVNet (Pixelwise Voting Network) to deny pixel-based vectors and use them for key-point positions.

C. 6D POSE ESTIMATION DIRECTLY FROM RGB-D IMAGES

A scene coordinate regression (SCoRe) forest is used, trained in a specific scene, employs only RGB-D image pixel comparison features and has fast calculation accuracy. The proposed method is an RNSAC-based pose optimization algorithm where SCoRe Forest is evaluated by the RNSAC algorithm and makes accurate posture estimates [222]. Since the additional depth channel of the RGB-D image helps extracts the entire 6D pose (3D rotation and 3D translation) of rigid object instances present in the scene. The core objective of the approach is the intermediate representation of the form of a dense 3D object coordinate labelled and paired with a dense class.

On the other hand, Taylor [245] did not predict 6 DoF directly from an RGB image but instead followed the object's coordinates in that image. Each pixel in this image points to a coordinate of the canonical body in a canonical position called VM (Vitruvian Manifold). The popular RF (Random Forest) [3] classifier is used to vote here, and geometric validity is used.

Brachmann et al., [13] provided an idea that is both an extension and combination of [222] and [245]. This hybrid concept estimates the 6D pose of a specific object from a single RGB-D image. Wang et al., [257] initiated a compelling method in cluttered scenes, which can successfully predict the object's posture. It has mixed colour and depth data from the RGB-D image and then integrates repetitive refinement methods into neural network architectures.

Li et al., [140] applied CNN to process the depth image as an additional image channel. However, the built-in 3D structure in the depth channel was neglected. In contrast, the geometric features of the dense fusion method convert pixels into sectional depths into 3D point clouds by applying built-in cameras. The proposed DPOD (Dense Pose Object Detector) applies PNP and RANSAC to compute an input image and a map of dense multi-class 2D / 3D correspondence between available 3D models [281].

D. INSTANCE-LEVEL 6 DOF POSE ESTIMATION

Collet et al., [37] created 3D object metric models using local descriptors of different images. Each model was optimized to easily fit a sequential training image, resulting in the best possible alignment between the 3D model and the original object. It combines the well-known RANSAC [64] and Mean Shift algorithm [35] to register multiple instances of each object that can successfully guess the 6-DOF pose for any complex and chaotic scene. In addition, it can handle randomly complex non-planning objects, powerful to handle outliers and occlusions, and able to control illumination, scale and rotation change.

The vision-based system, which is actually an extension

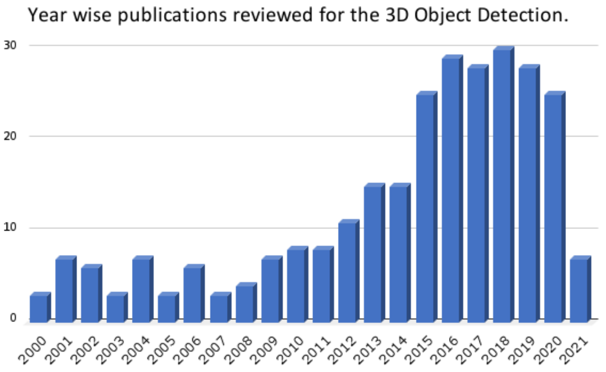

FIGURE 2. Increasing amount of efforts in literature on monocular 3D object detection in recent years.

of Gordon's method [78], enables the accurate localization initialization step called POSESEQ and enables full pose inference in object recognition in a complete cluttered environment. Thanh et al. presented LieNet [50], as a unique template-based pose estimation method that uses the Lie algebraic rotation matrix to estimate the rotation matrix of an object. It estimates the translation vector by predicting the distance of the object from the centre of the camera. This method takes the input of an image and then outputs the object's identification with a 6D pose, including a bounding box, label, and segmentation mask.

Vidal et al., [255] developed a method that followed the basic structure of the point pair feature (PPF) method introduced by Drost [55], which is a combination of two levels, such as global modelling and local matching. The main structure identifies the rotation points, model points and angles of each scene. The expansion of Vidal's work is the concept of the posture of free-form objects, critical work in favour of a highly confused autonomous system. A novel pre-processing step has been added here, transforming the classification into a better efficient feature matching method.

E. CATEGORY-LEVEL 6 DOF POSE ESTIMATION

Sahin et al., [207] covers various challenges for 6D pose estimation such as inconsistency of viewpoint, objects (both texture and texture-less), curbs, cluttered scene and identical objects. Wang et al., [260] has created a method that assumes both 6D poses of hidden object instances without an object CAD model in an RGB-D image. Furthermore, a novel concept called NOCS (Normalised Object Coordinate Space) has been introduced here, representing a partnership principle for all possible instances of an object.

Schuster et al., [215] evaluates dense 3D data located in multiple light situations and applies online graph SLAM to generate a dense 3D composite map and estimates 6D poses. This technique also creates a fancy graph topology for incorporating the results of local reference filters and overall high-bandwidth sensor data into sub-maps.

F. FEATURE MATCHING METHODS

To solve the 6D object pose hypothesis and ensure the best possible accuracy, Krull [125] successfully applied Reinforcement Learning to the pose agent classification for the first time. Each decision here follows the potential distribution of a stochastic policy gradient approach that takes a direct gradient in terms of the expected loss of interest.

G. TEMPLATE-MATCHING TECHNIQUES

Hinterstoisser et al., [90] built a framework called LineMod for automatic detection and tracking of 3D objects based on the latest template-based approach that uses both depth and colour images to capture the object's presence and 3D shape on a set of templates with different aspects of an object. Also, the 3D model can be used for the accurate estimation of the position of the object. Tejani et al., [246] developed a novel patch-based framework where a Latent-Class Hough Forests method for 3D object detection was introduced, and estimations were made in a heavily cluttered and occluded environment. This method absorbs the classification labels during training, and as a by-product, it creates the right image-ground mask.

H. CNN/ DEEP LEARNING - BASED APPROACHES

Krull [124] presented a model for 6D pose estimation, which applied a CNN to map images and revealed that training on a single object was sufficient and that CNN successfully generalized all the different objects and backgrounds of an image. Rangesh et al., [199] applied an exclusive idea for a 3D identification box suitable for the object on the ground to combine 2D visual context, 3D dimension and ground plane. Eppner et al., [59] presented and evaluated the winning system for the 2015 Amazon Picking Challenge, where they created four key aspects of system building: integration, manipulation, manipulation planning, and estimation.

Google has announced the release of MediaPipe, a 3D object detection pipeline that identifies objects in 2D images on everyday objects and estimates their poses and sizes. MediaPipe is a cross-platform structure that builds ML pipelines and creates 3D bounding boxes with augmented reality (AR) [5] and identifies additional information such as camera pose, 3D point cloud, lighting and planar surfaces [84, 267]. Basically, MediaPipe performs object detection, face detection, hand tracking, hair segmentation with ML frameworks called Tensorflow and Tensorflow Lite [1].

A novel model [100] designed to predict the pose and size of an object from a monocular RGB image has applied a multi-task-learning approach named MobileNetv2 [211] and predicts object size. The Gaussian regression task applies a pose estimation algorithm (EPnP) [135] to the final 3D coordinates for the bounding box. A novel model [100] designed to predict the pose and size of an object from a monocular RGB image has applied a multi-task-learning approach named MobileNetv2 [211] and predicts object size. The Gaussian regression task applies a pose estimation al-

TABLE 6. Advantages and disadvantages of some state-of-the-art 6D pose estimation methods.

No.	Algorithms	Advantages	Disadvantages	DataSets
1	LieNet [50]	Quite simple and inexpensive pose refiner	Poor performance for rotational symmetry objects.(e.g., coffee mug)	LINEMOD and Tejani's dataset
2	DenseFusion[257]	Shows robustness when the scene is extremely cluttered.	Unable to grasp a specific type of object.	YCB-Video and LineMOD
3	Brachmann [12]	Swift, scalable, powerful and exceptionally perfect method for generic objects.	Direct using an auto-context random forest hampered test performance.	RGB - D Images [91]
4	POSESEQ [37]	Handles complex non-planar objects with great speed.	Poor latency scale into household environments.	RGB - Images
5	MOPED [161]	Improved scalability and optimized speed with high robustness, accuracy and low latency.	Computationally expensive and large Image show low performance.	4 Real World Benchmark
6	Brachmann [13]	Covers the textured and textureless, rigid and non-rigid objects, symmetrical objects.	The pipeline is linear in the number of objects.	RGB - D Images
7	LINEMOD [90]	Easy to deploy, reliable, and fast.	Poor performances and suffers for false positives data.	LINEMOD [90]
8	SSD-6D [115]	Stable training and robust prediction.	Computationally expensive.	LINEMOD and OCCLUSION
9	BB8 [197]	Perform better on rotational symmetrical and similar objects.	Performs badly for Texture-LESS dataset.	LINEMOD, OCC, T-LESS.
10	Iryna [78]	An efficient, incremental and jitters reduction method.	Unscalable for operation in large environments	Live video sequence
11	Posecnn [275]	ShapeMatch-Loss for symmetric objects pose estimation.	Estimate pose for only color images.	YCB-Video, LINEMOD and OCC
12	PoseAgent [125]	Dramatically reduces computation time and variance.	Estimate pose for only color images.	Krull et al [126]
13	Deep-6DPose [49]	Better trade-off between speed and accuracy.	Not very strong to nearly rotational symmetric objects.	LINEMOD [90]
14	DeepIM [142]	Accurately manage poses and shapes for texture-less and unseen objects.	Estimate pose for color images only.	LINEMOD and OCCLUSION
15	BB8 [197]	Effective method.	Relatively Slow for large image.	LINEMOD and OCCLUSION

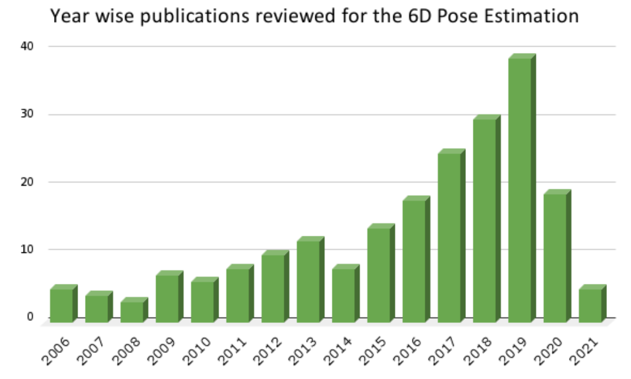

FIGURE 3. Increasing amount of efforts in literature on 6D pose estimation in recent years

gorithm (EPnP) [135] to the final 3D coordinates for the bounding box.

Tremblay et al., [250] introduced the first one-shot deep neural network for robotic manipulation trained only on synthetic data capable of achieving 6-DoF object pose estimates of 3D objects. The system is called DOPE (Deep Object Pose Estimation), which applies the Perspective-N-Point (PNP) algorithm, which combines 3D bounding boxes with 2D images. Li [140] has proposed a pose correction algorithm where the solution is to correct the pose because the object is being observed from the centre line of the camera. It is a multi-philosophy fusion framework with a single philosophical ambiguity and quick guess selection based on a voting scheme.

On the other hand, a model called DeepIM [142] is able to predict a relational pose transformation by applying 3D location and 3D orientation and a repetitive training process. The network FlowNetSimple architecture uses the backbone network project as DeepHMap++, which centres a two-stage pipeline and integrates two learning concepts to estimate 6D poses of invisible objects in challenging scenes [70].

I. TEMPLATE-CLUSTERING APPROACHES

Zhang et al., [287] has proposed City-LineMod (advanced to Cognitive Template-Clustering Line mode) method. The technique applies a 7D (4D geometry + 3D texture) cognitive feature vector to restore the standard 3D spacing points in the patch-linemode clustering method. Moreover, the distance of different 3D spatial points will also be affected by the 4D additional information regarding the direction and width of

TABLE 7. 6D Full Object Pose Estimation Paper Collection

No.	Approach	References
1	Pose from RGB images.	[207], DenseFusion[257], MPOED[38], SSD-6D[115], BB8 [197], [247], [100], [18], [25], 6D-VNet [268], GPP [199], [288], [11], Deep-6DPose [49], RePOSE [105], DPOD [282], E2E6DoF [80]
2	Pose from depth / point cloud method.	[289], [167], PointPoseNet [82], [23], [254], [285].
2	Pose from RGB-D data.	[216], [98], [96], [95], [97], [281], [245], [54], [203], [36]. [210], [179], [230].
3	Instance-Level 6 DoF Pose Estimation	POSESEQ [37], MOPED [161], LieNet [50] , [23], [207], [255], [231] [2], GSNet [114].
4	Category-Level 6 DoF Pose Estimation	[78], [184], C3DPO [180], [207], [212], [260], [215], [148], Pix2Pose [187].
5	Feature matching methods	[156], [38], [256], [13], [125], [165], [239], EfficientPose [17],
7	Template-matching techniques	[90], [116], [204], [246], [188],
8	CNN/ Deep Learning -based approaches.	[124],[275], GPP [199], [26], Deep3DBox [171], [59], [94], [250], [140], [74], DeepIM [142], DeepHMap++ [70], HRNet [29].
9	Template-clustering approaches.	PVNet [190], CT-LineMod[287], HybridPose [233]

the features.

J. HYBRID POSE METHOD

Martinez et al., [161] presented a hybrid GPU / CPU architecture that uses parallelism at all levels named MOPED (Multiple Object Position Estimation and Detection), a bright and measurable perception concept for both object recognition and fracture estimation. Furthermore, a mode based on another object recognition algorithm known as POSESEQ [37] showed a massive increase in scalability and accuracy and optimizes the algorithm's speed. Technically, MOPED has employed a new feature-matching algorithm that optimizes databases to handle complexity and a robust pose merge algorithm capable of efficiently rejecting outsiders with matching K-NN (K-Nearest Neighbour) method where $k > 2$ [244]. The default classification algorithm and SIFTGPU is the MOPED feature extraction algorithm.

VI. FUTURE RESEARCH DIRECTIONS

From the above discussion, it is clear that plenty of work has been done on 3D object detection, which forms a solid foundation for this field. Nevertheless, further research is needed as 6D pose estimation systems have not yet performed adequately. Therefore, this part of the article will give some possible ideas of the future directions for both sectors, which will help understand the status and involvement of 3DOD and 6 DPE.

A. FUTURE RESEARCH DIRECTIONS FOR 3D OBJECT DETECTION

1) Detecting a rigid object:

Several existing work [270], [150], [31], [32] showed the efficacy of deep learning in detecting a rigid object. Even though the classifiers are mainly focused on the "car" category, the concept of these methods can be contextual to all other solid and inflexible types of objects. The accurate detection of the 3D rigid object is a complex job and is very significant in the domain of computer vision. Currently, using some proposed deep learning techniques, we can detect inflexible objects, but still, lots of works need to do to make the process flawless. In developing an autonomous car, accurately identifying rigid objects can be a significant research idea.

2) Handing rotationally symmetric objects:

Identifying half and full symmetry objects like a coffee mug or glass is a confusing and complex matter that classifiers usually fail to give accurate results. Do et al., [50] has suggested an idea to overcome this complicated problem, but not entirely successful. Although the work on symmetric object detection is starting to get deeper, not much work has been done to date. More attention and research needs to be done on this case of symmetrical object identification.

3) Tracking a object from video:

One possible incoming direction is to simultaneously explore its application as part of a system that uses repetitive neural networks to detect and track objects in video [115]. In addition, the intense colour variation between the CAD model and the visual avatar is a significant work. Another potential research aspect in this context is online model learning and relocalization [222]. A hypothesis can be developed to represent both single and multiview with an extended update of a new frame [140]. In addition, to avoid the problem of proper loss, the term-balancing required for upcoming potential research direction [115].

B. FUTURE RESEARCH DIRECTIONS FOR 6D POSE ESTIMATION

1) Improving the VO (visual odometry)

Appropriate VO (Visual Odometry) is mandatory in the context of autonomous driving; Thus, the future design of both automated car and street scene construction needs to be improved [290]. It is not possible to apply driverless cars without the proper implementation of VO.

2) Improving the 6D pose estimate accuracy

One can improve the version of the DeepIM method [142] for autonomous applications to produce accurate 6D pose estimates from high-resolution camera images (colour only) at high frame rates with a large field view. The authors also mentioned using stereotype camera images as input to improve the quality of this method. Another work can be done by combining the advanced two-step method to transform it into a new pose tracking framework where the pose parameters from the previous frame can be reused to replace the pose detection step in DeepHMap [70]. Adding a branch to the back for object segmentation in DeepHMap may provide some additional regularization.

3) Improving the Map or VPS

It is an open challenge to efficiently and consistently merge sub-maps into multi-robot systems to create a long-term mapping system, aiming at improving the algorithm that matches the map [215]. The globalization strategy combines visual positioning services (VPS), street view, and machine learning for more accurate location and adaptation detection. Mutual technology is essential to enhance the correct positioning and orientation of blue dots on digital maps in our cars, smartphones, and up-to-date interactions.

4) Improving the pose of symmetrical objects

Since managing the poses of symmetrical or symmetrical objects is a complex task, relevant classifiers should be improved to accomplish the task efficiently [275]. In order to properly manage symmetrical properties, methods need to learn the symmetry of objects and update their capabilities [26]. One of the notable tasks may be to manage the symmetry property of objects and pose estimation automatically.

5) Improving the function of mobile manipulators

The efficiency of the POSSEQ [37] classifier can be enriched by enabling mobile manipulators to work more perfectly to communicate with the crowd's internal environment. Some hypothetical 6DoF pose [268] reprocessing techniques will be filtered using repetitive closet point-based algorithms or repetitive retrieval networks. Also, classifiers need to successfully model and recognize scenes of different sizes and complexities in large environments [78] (campus, laboratory, shopping centre or a museum).

6) Improving the 3D point cloud networks

A number of 3D point cloud networks can be replaced directly by the PointNet network [82] for potential improvement in accurate 3D object detection and 6DF pose estimation. A computational budget can be created to know the appropriate time for the softer version of the PoseAgent [125] classification. For parallelism, multiple computational cores can be applied by advanced PoseAgent. In addition, training can be provided to replace the processing steps of an existing CNN method and improve the results by observing and predicting updated postures from the given images [124].

7) Improving the Dataset

To deal with the common challenges of objects, such as reflective and texture-less objects, and the adverse conditions, such as occlusion and changing lighting conditions, we can integrate some multi-dimensional object models into the dataset packages. To facilitate the reconstruction of indoor and outdoor dynamic scenes, 4D or 5D models can be added to the dataset, which can play an important role in any visual applications such as navigational systems for moving objects (for example: autonomous car) [280].

8) Removing the visible gap between machine performance and that of human's

In ApolloCar3D, researchers [237] mentioned four visible surfaces and manually defines a correspondence between critical points and surfaces. They suggested that a total of 66 key points were assigned to every single car model (for both SUVs and Sedans). According to [237], since people cannot memorize the semantic meaning of 66 key points correctly, there is a noticeable gap (10 %) in between algorithms/machines with humans. Henceforth, correctly resolving visible gaps between machines and humans can be a future inspiration for research.

9) Explore geometric properties

Estimating the 6DoF pose of an object from a single RGB image is a significant and challenging task, especially under heavy occlusion and for the Texture-Less object. In such a case, the exploring of geometric features needs to be improved to estimate the 6 DOF object more efficiently [80].

VII. CONCLUSION

This review paper studies the state-of-the-art deep learning techniques for 3D object detection and 6D pose estimation. Most current object detection methods identify images with a 2D bounding box technique that can recognize both the position and range of the objects in the image. However, recognizing a vehicle as a 2D BB is not always sufficient for perfect autonomous driving. Therefore, predicting the position of the 3D object from the images is just as important as determining the 2D position of the vehicle. For 3D object detection, current works report sophisticated results using RGB / RGB-D imagery, point cloud, and fusion-based techniques.

Here, with the help of this review, we have addressed the advantages and disadvantages of each of the basic techniques, both 3D object detection and 6D pose estimation techniques. We have also mentioned some traditional theoretical evaluation metrics and summarised the popular Big Image datasets applied by well-known object identification and pose estimation methods. Since the deep learning method of 3D object detection and 6D pose estimation are not as mature as 2D object detection, research is needed for real-time operation. From now on, a significant improvement needs to be made to manage a fast and reliable 3 DOD and 6 DPE system across a broad set of real-time practical applications. Although RGB-D is much simpler than RGB, it faces problems for some depth issues, such as not being able to recognize small objects properly.

Several classifications have been proposed in the 6D Pose estimation functions, such as the point addition method, the template matching method, the Hough forest method, and the deep learning method. However, the effectiveness of the proposed classifiers is still far from the level of actual application, which should be able to successfully predict 6D poses of multi-objects, including severe occurrence and chaos scene situations. Therefore, this article presents an in-depth review of the most significant work to date on in-depth learning-based 3D object detection and 6D pose estimation systems. Until then, we believe that this review article can be cited and used as a sample source of reference and forms an important endorsement to the research community.

• • •

REFERENCES

[1] An end-to-end open source machine learning platform.

[2] Georgios Nikolaos Albanis, Nikolaos Zioulis, Anargyros Chatzitofis, Anastasios Dimou, Dimitrios Zarpalas, and Petros Daras. On end-to-end 6DOF object pose estimation and robustness to object scale. In ML Reproducibility Challenge 2020, 2021.

[3] M. Y. Arafat, S. Hoque, and D. M. Farid. Cluster-based under-sampling with random forest for multi-class imbalanced classification. In 2017 11th International Conference on Software, Knowledge, Information Management and Applications (SKIMA), pages 1–6, 2017.

[4] Md Yasir Arafat, Sabera Hoque, and Dewan Md Farid. An Under-Sampling Method with Support Vectors in Multi-class Imbalanced Data Classification. 13'th International Conference on Software, Knowledge, Information Management and Applications (SKIMA), 2019.

[5] Ronald T. Azuma. A Survey of Augmented Reality, volume 6. Hughes Research Laboratories, 1997.

[6] D.H. Ballard. Generalizing the hough transform to detect arbitrary shapes. Pattern Recognition, 13(2):111 – 122, 1981.

[7] M. R. Banham and A. K. Katsaggelos. Digital image restoration. IEEE Signal Processing Magazine, 14(2):24–41, March 1997.

[8] W. Bao, B. Xu, and Z. Chen. Monofenet: Monocular 3d object detection with feature enhancement networks. IEEE Transactions on Image Processing, 29:2753–2765, 2020.

[9] Barabás, A Todoruţ, N Cordoş, and A Molea. Current challenges in autonomous driving. IOP, (012096):252, 2017.

[10] J. L. Barron, D. J. Fleet, and S. S. Beauchemin. Performance of optical flow techniques. International Journal of Computer Vision, 12(1):43–77, 1994.

[11] Gideon Billings and Matthew Johnson-Roberson. Silhonet: An rgb method for 6d object pose estimation. IEEE ROBOTICS AND AUTOMATION LETTERS, 2018.

[12] E. Brachmann, F. Michel, A. Krull, M. Y. Yang, S. Gumhold, and C. Rother. Uncertainty-driven 6d pose estimation of objects and scenes from a single rgb image. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 3364–3372, 2016.

[13] Eric Brachmann, Alexander Krull, Frank Michel, Stefan Gumhold, Jamie Shotton, and Carsten Rother. Learning 6d object pose estimation using 3d object coordinates. In David Fleet, Tomas Pajdla, Bernt Schiele, and Tinne Tuytelaars, editors, Computer Vision - ECCV 2014, pages 536–551, Cham, 2014. Springer International Publishing.

[14] Garrick Brazil and Xiaoming Liu. M3d-rpn: Monocular 3d region proposal network for object detection, 2019.

[15] Marjolein Bruijning, Marco Visser, Caspar Hallmann, and Eelke Jongejans. trackdem : Automated particle tracking to obtain population counts and size distributions from videos in r. Methods in Ecology and Evolution, 01 2018.

[16] Fan Bu, Trinh Le, Xiaoxiao Du, Ram Vasudevan, and Matthew Johnson-Roberson. Pedestrian planar lidar pose (pplp) network for oriented pedestrian detection based on planar lidar and monocular images. IEEE Robotics and Automation Letters, PP:1–1, 12 2019.

[17] Yannick Bukschat and Marcus Vetter. Efficientpose: An efficient, accurate and scalable end-to-end 6d multi object pose estimation approach. CVPR 2020, 2020.

[18] M. Burenius, J. Sullivan, and S. Carlsson. 3d pictorial structures for multiple view articulated pose estimation. In 2013 IEEE Conference on Computer Vision and Pattern Recognition, pages 3618–3625, 2013.

[19] 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. ArXiv 2019, 2019.

[20] Zhaowei Cai and Nuno Vasconcelos. Cascade r-cnn: Delving into high quality object detection. Computer Vision and Pattern Recognition, 2017.

[21] Berk Calli, Arjun Singh, James Bruce, Aaron Walsman, Kurt Konolige, Siddhartha Srinivasa, Pieter Abbeel, and Aaron Dollar. Yale-cmu-berkeley dataset for robotic manipulation research. The International Journal of Robotics Research, 36:027836491770071, 04 2017.

[22] Berk Calli, Arjun Singh, Aaron Walsman, Siddhartha Srinivasa, Pieter Abbeel, and Aaron Dollar. The ycb object and model set: Towards common benchmarks for manipulation research. In IEEE, editor, Proceedings of IEEE International Conference on Advanced Robotics (ICAR), July 2015.

[23] D. Campbell, L. Petersson, L. Kneip, and H. Li. Globally-optimal inlier set maximisation for camera pose and correspondence estimation. IEEE Transactions on Pattern Analysis and Machine Intelligence, 42(2):328–342, 2020.

[24] Richard J. Campbell and Patrick J. Flynn. A survey of free-form object representation and recognition techniques. Computer Vision and Image Understanding, 81(2):166 – 210, 2001.

[25] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Realtime multi-person 2d pose estimation using part affinity fields, 2016.

[26] Catherine Capellen, Max Schwarz, and Sven Behnke. Convposecnn: Dense convolutional 6d object pose estimation. Computer Vision and Pattern Recognition , CVPR, 2019.

[27] Florian Chabot, Mohamed Chaouch, Jaonary Rabarisoa, Céline Teulière, and Thierry Chateau. Deep manta: A coarse-to-fine many-task network for joint 2d and 3d vehicle analysis from monocular image, 2017.

[28] Angel X. Chang, Thomas Funkhouser, Leonidas Guibas, Pat Hanrahan, Qixing Huang, Zimo Li, Silvio Savarese, Manolis Savva, Shuran Song, Hao Su, Jianxiong Xiao, Li Yi, and Fisher Yu. Shapenet: An information-rich 3d model repository. Computer Vision and Pattern Recognition, 2015.

[29] Bo Chen, Alvaro Parra, Jiewei Cao, Nan Li, and Tat-Jun Chin. End-to-end learnable geometric vision by backpropagating pnp optimization, 2020.

[30] Junying Chen and Tongyao Bai. Saanet: Spatial adaptive alignment network for object detection in automatic driving. Image and Vision Computing, 94:103873, 2020.

[31] X. Chen, K. Kundu, Z. Zhang, H. Ma, S. Fidler, and R. Urtasun. Monocular 3d object detection for autonomous driving. In 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 2147–2156, 2016.

[32] X. Chen, H. Ma, J. Wan, B. Li, and T. Xia. Multi-view 3d object detection network for autonomous driving. In Proc. 2017 IEEE Conf. on Comput. Vision and Pattern Recognition (CVPR), 2017.

[33] Xiaozhi Chen, Kaustav Kundu, Yukun Zhu, Andrew G Berneshawi, Huimin Ma, Sanja Fidler, and Raquel Urtasun. 3d object proposals for accurate object class detection. In C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, editors, Advances in Neural Information Processing Systems 28, pages 424–432. Curran Associates, Inc., 2015.

[34] Ming-Ming Cheng, Yun Liu, Wen-Yan Lin, Ziming Zhang, Paul L. Rosin, and Philip H. S. Torr. Bing: Binarized normed gradients for objectness estimation at 300fps. Computational Visual Media, 5(1):3–20, 2019.

[35] Yizong Cheng. Mean shift, mode seeking, and clustering. IEEE Transactions on Pattern Analysis and Machine Intelligence, 17(8):790–799, Aug 1995.

[36] Chi Li, J. Bohren, E. Carlson, and G. D. Hager. Hierarchical semantic parsing for object pose estimation in densely cluttered scenes. In 2016 IEEE International Conference on Robotics and Automation (ICRA), pages 5068–5075, 2016.

[37] A. Collet, D. Berenson, S. S. Srinivasa, and D. Ferguson. Object recognition and full pose registration from a single image for robotic manipulation. IEEE International Conference on Robotics and Automation, pages 48–55, 2009.

[38] Alvaro Collet, Manuel Martinez, and Siddhartha S Srinivasa. The moped framework: Object recognition and pose estimation for manipulation. The International Journal of Robotics Research, 30(10):1284–1306, 2011.

[39] Adam Conner-Simons and Rachel Gordon. Self-driving cars for country roads, May 2018.

[40] C Cortes and V. Vapnik. Support-vector networks. In Machine Learning, volume 20, pages 273–297. doi:10.1023/A:1022627411411, 1995.

[41] Robert Cupec, Ivan Vidović, Damir Filko, and Petra Đurović. Object recognition based on convex hull alignment. Pattern Recognition, 102:107199, 2020.

[42] Jifeng Dai, Yi Li, Kaiming He, and Jian Sun. R-fcn: Object detection via region-based fully convolutional networks. arXiv, 2016.

[43] Jifeng Dai, Haozhi Qi, Yuwen Xiong, Yi Li, Guodong Zhang, Han Hu, and Yichen Wei. Deformable convolutional networks. Computer Vision and Pattern Recognition, 2017.

[44] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05), volume 1, pages 886–893 vol. 1, 2005.

[45] E. R. Davies. Machine Vision: Theory, Algorithms, Practicalities. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 2004.

[46] Jeffrey Dean, Greg Corrado, Rajat Monga, Kai Chen, Matthieu Devin, Mark Mao, Marc'aurelio Ranzato, Andrew Senior, Paul Tucker, Ke Yang, Quoc V. Le, and Andrew Y. Ng. Large scale distributed deep networks. In F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, editors, Advances in Neural Information Processing Systems 25, pages 1223–1231. Curran Associates, Inc., 2012.

[47] Jia Deng, Wei Dong, Richard Socher, Li jia Li, Kai Li, and Li Fei-fei. Imagenet: A large-scale hierarchical

image database. In In CVPR, 2009.

[48] Wenhao Ding, Shuaijun Li, Guilin Zhang, Xiangyu Lei, and Huihuan Qian. Vehicle pose and shape estimation through multiple monocular vision, 2018.

[49] Thanh-Toan Do, Ming Cai, Trung Pham, and Ian Reid. Deep-6dpose: Recovering 6d object pose from a single rgb image. arXiv, 2018.

[50] Thanh-Toan Do, Trung Pham, Ming Cai, and Ian D. Reid. Lienet: Real-time monocular object instance 6d pose estimation. In BMVC, 2018.

[51] Jadranka Dokic, Beate Müller, and Gereon Meyer. European roadmap smart systems for automated driving. European Technology Platform on Smart Systems Integration (EPoSS), 04 2015.

[52] P. Dollar, C. Wojek, B. Schiele, and P. Perona. Pedestrian detection: An evaluation of the state of the art. IEEE Transactions on Pattern Analysis and Machine Intelligence, 34(4):743–761, 2012.

[53] Piotr Dollár, Ron D. Appel, Serge J. Belongie, and Pietro Perona. Fast feature pyramids for object detection. IEEE Transactions on Pattern Analysis and Machine Intelligence, 36:1532–1545, 2014.

[54] Andreas Doumanoglou, Rigas Kouskouridas, Sotiris Malassiotis, and Tae-Kyun Kim. Recovering 6d object pose and predicting next-best-view in the crowd. CVPR, 2015.

[55] B. Drost, M. Ulrich, N. Navab, and S. Ilic. Model globally, match locally: Efficient and robust 3d object recognition. In 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pages 998–1005, 2010.

[56] Bertram Drost, Markus Ulrich, Paul Bergmann, Philipp Härtinger, and Carsten Steger. Introducing mvtec itodd - a dataset for 3d object recognition in industry. In Proc. IEEE Int. Conf. Comput. Vis. Workshops (ICCV), 10 2017.

[57] Martin Engelcke, Dushyant Rao, Dominic Zeng Wang, Chi Hay Tong, and Ingmar Posner. Vote3deep: Fast object detection in 3d point clouds using efficient convolutional neural networks. CVPR 2016, 2016.

[58] M. Enzweiler and D. M. Gavrila. A multilevel mixture-of-experts framework for pedestrian classification. In IEEE Transactions on Image Processing, volume 20, pages 2967–2979, Oct 2011.

[59] Clemens Eppner, Sebastian Höfer, Rico Jonschkowski, Roberto Martin-Martin, Arne Sieverling, Vincent Wall, and Oliver Brock. Lessons from the amazon picking challenge: Four aspects of building robotic systems. In Twenty-Sixth International Joint Conference on Artificial Intelligence, pages 4831–4835, 08 2017.

[60] M. Everingham, S. M. A. Eslami, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The pascal visual object classes challenge: A retrospective. International Journal of Computer Vision, 111(1):98–136, January 2015.

[61] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes Challenge 2007 (VOC2007) Results. http://www.pascal-network.org/challenges/VOC/voc2007/workshop/index.html.

[62] Jiaojiao Fang, Lingtao Zhou, and Guizhong Liu. 3d bounding box estimation for autonomous vehicles by cascaded geometric constraints and depurated 2d detections using 3d results, 2019.

[63] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part-based models. IEEE Transactions on Pattern Analysis and Machine Intelligence, 32(9):1627–1645, 2010.

[64] Martin A. Fischler and Robert C. Bolles. Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography. Commun. ACM, 24(6):381–395, June 1981.

[65] Robert Fisher. Cvonline: The evolving, distributed, non-proprietary, on-line compendium of computer visio. School of Informatics University of Edinburgh, 2019.

[66] David A. Forsyth and Jean Ponce. Computer Vision: A Modern Approach. Number 792 in ISBN-13: 978-0136085928. Pearson Higher Ed USA, 2011.

[67] Yoav Freund and Robert E Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. Journal of Computer and System Sciences, 55(1):119 – 139, 1997.

[68] Andrea Frome, Daniel Huber, Ravi Kolluri, Thomas Bülow, and Jitendra Malik. Recognizing objects in range data using regional point descriptors. In Tomáš Pajdla and Jiří Matas, editors, Computer Vision - ECCV 2004, pages 224–237, Berlin, Heidelberg, 2004. Springer Berlin Heidelberg.

[69] Huan Fu, Mingming Gong, Chaohui Wang, Kayhan Batmanghelich, and Dacheng Tao. Deep ordinal regression network for monocular depth estimation. CVPR 2018, 2018.

[70] Mingliang Fu and Weijia Zhou. Deephmap++: Combined projection grouping and correspondence learning for full dof pose estimation. Sensors, 19(5):1032, Feb 2019.

[71] Leon A. Gatys, Alexander S. Ecker, and Matthias Bethge. A neural algorithm of artistic style. arXiv:1508.06576, 2015.

[72] Stefan K Gehrig and Fridtjof J Stein. Dead reckoning and cartography using stereo vision for an autonomous car. IEEE or RSJ International Conference on Intelligent Robots and Systems Human and Environment Friendly Robots with High Intelligence and Emotional Quotients (Cat. No.99CH36289), 3:1507–1512, 1999.

[73] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In IEEE Computer Society Conference on Computer Vision and Pattern Recognition,

(CVPR), pages 3354–3361, 05 2012.

[74] Nils Gessert, Matthias Schlüter, and Alexander Schlaefer. A deep learning approach for pose estimation from volumetric oct data. Medical Image Analysis, 46:162–179, May 2018.

[75] R. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. IEEE Conference on Computer Vision and Pattern Recognition, pages 580–587, 2014.

[76] Ross Girshick. Fast r-cnn, 2015.

[77] A. González, D. Vázquez, A. M. López, and J. Amores. On-board object detection: Multicue, multimodal, and multiview random forest of local experts. In IEEE Transactions on Cybernetics, volume 47, pages 3980–3990, 07 2017.

[78] Iryna Gordon and David Lowe. What and where: 3d object recognition with accurate pose. In Heidelberg Springer, Berlin, editor, Ponce J., Hebert M., Schmid C., Zisserman A. (eds) Toward Category-Level Object Recognition, volume 4170, pages 67–82, 01 2006.

[79] Yulan Guo, Hanyun Wang, Qingyong Hu, Hao Liu, Li Liu, and Mohammed Bennamoun. Deep learning for 3d point clouds: A survey. CVPR 2020, 2020.

[80] Anshul Gupta, Joydeep Medhi, Aratrik Chattopadhyay, and Vikram Gupta. End-to-end differentiable 6dof object pose estimation with local and global constraints, 2020.

[81] Fredrik K. Gustafsson, Martin Danelljan, and Thomas B. Schön. Accurate 3d object detection using energy-based models. CVPR 2020, 2020.

[82] Frederik Hagelskjaer and Anders Buch. Pointposenet: Accurate object detection and 6 dof pose estimation in point clouds. Computer Vision and Pattern Recognition (CVPR), 12 2019.

[83] Andrew J. Hawkins. Mit built a self-driving car that can navigate unmapped country roads, May 2015.

[84] Michael Hays and Tyler Mullen. Mediapipe on the web. Blog, January 2020.

[85] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In Facebook AI Research (FAIR). Computer Vision and Pattern Recognition (CVPR), 2017.

[86] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. CVPR 2016, 2015.

[87] Qingdong He, Zhengning Wang, Hao Zeng, Yi Zeng, Shuaicheng Liu, and Bing Zeng. Svga-net: Sparse voxel-graph attention network for 3d object detection from point clouds. CVPR 2020, 2020.

[88] Zaixing He, Wuxi Feng, Xinyue Zhao, and Yongfeng Lv. 6d pose estimation of objects: Recent technologies and challenges. Applied Sciences, 11(1), 2021.

[89] G. Hetzel, B. Leibe, P. Levi, and B. Schiele. 3d object recognition from range images using local feature histograms. In Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2001, volume 2, pages II–II, 2001.

[90] Stefan Hinterstoisser, Vincent Lepetit, Slobodan Ilic, Stefan Holzer, Gary Bradski, Kurt Konolige, and Nassir Navab. Model based training, detection and pose estimation of texture-less 3d objects in heavily cluttered scenes. In Kyoung Mu Lee, Yasuyuki Matsushita, James M. Rehg, and Zhanyi Hu, editors, Computer Vision – ACCV 2012, pages 548–562, Berlin, Heidelberg, 2013. Springer Berlin Heidelberg.

[91] Stefan Hinterstoisser, Vincent Lepetit, Naresh Rajkumar, and Kurt Konolige. Going further with point pair features. Lecture Notes in Computer Science, pages 834–848, 2016.

[92] G. Hinton, L. Deng, D. Yu, G. E. Dahl, A. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. N. Sainath, and B. Kingsbury. Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. IEEE Signal Processing Magazine, 29(6):82–97, 2012.

[93] G. E. Hinton and R. R. Salakhutdinov. Reducing the dimensionality of data with neural networks. Science, 313(5786):504–507, 2006.

[94] D. Hoang, A. J. Lilienthal, and T. Stoyanov. Panoptic 3d mapping and object pose estimation using adaptively weighted semantic information. IEEE Robotics and Automation Letters, 5(2):1962–1969, 2020.

[95] Tomas Hodan, Pavel Haluza, Stepan Obdrzalek, Jiri Matas, Manolis Lourakis, and Xenophon Zabulis. T-less: An rgb-d dataset for 6d pose estimation of texture-less objects. WACV, 2017.

[96] Tomás Hodan, Juan E. Sala Matas, and Stepán Obdrzálek. On evaluation of 6d object pose estimation. In ECCV Workshops, 2016.

[97] Tomas Hodan, Frank Michel, Eric Brachmann, Wadim Kehl, Anders Glent Buch, Dirk Kraft, Bertram Drost, Joel Vidal, Stephan Ihrke, Xenophon Zabulis, Caner Sahin, Fabian Manhardt, Federico Tombari, Tae-Kyun Kim, Jiri Matas, and Carsten Rother. Bop: Benchmark for 6d object pose estimation, 2018.

[98] T. Hodaň, X. Zabulis, M. Lourakis, Š. Obdržálek, and J. Matas. Detection and fine 3d pose estimation of texture-less objects in rgb-d images. In 2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pages 4421–4428, 2015.

[99] Sabera Hoque, Dewan Md Farid, and Md Yasir Arafat. Advanced Data Balancing Method with SVM Decision Boundary and Bagging. 6th IEEE CSDE, CQUniversity Australia, Melbourne, Australia, 2019.

[100] Tingbo Hou, Adel Ahmadyan, Liangkai Zhang, Jianing Wei, and Matthias Grundmann. Mobilepose: Realtime pose estimation for unseen objects with weak shape supervision. Computer Vision and Pattern Recognition(CVPR), 2020.

[101] Hou-Ning Hu, Qi-Zhi Cai, Dequan Wang, Ji Lin, Min Sun, Philipp Krähenbühl, Trevor Darrell, and Fisher Yu. Joint monocular 3d vehicle detection and tracking, 2018.

[102] Jonathan Huang, Vivek Rathod, Chen Sun, Menglong Zhu, Anoop Korattikara, Alireza Fathi, Ian Fischer, Zbigniew Wojna, Yang Song, Sergio Guadarrama, and Kevin Murphy. Speed/accuracy trade-offs for modern convolutional object detectors. CVPR 2017, 2016.

[103] Veli ilçi and Charles Toth. High definition 3d map creation using gnss/imu/lidar sensor integration to support autonomous vehicle navigation. Sensors, 20:899, 02 2020.

[104] Anastasia Ioannidou, Elisavet Chatzilari, Spiros Nikolopoulos, and Ioannis Kompatsiaris. Deep learning advances in computer vision with 3d data: A survey. ACM Computing Surveys, 50, 06 2017.

[105] Shun Iwase, Xingyu Liu, Rawal Khirodkar, Rio Yokota, and Kris M. Kitani. Repose: Real-time iterative rendering and refinement for 6d object pose estimation, 2021.

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

[107] Bernd Jähne and Horst Haußecker. Computer vision and applications a guide for students and practitioners. ACADEMIC PRESS, 2000.

[108] Joel Janai, Fatma Güney, Aseem Behl, and Andreas Geiger. Computer vision for autonomous vehicles: Problems, datasets and state of the art, 2017.

[109] Yangqing Jia, Evan Shelhamer, Jeff Donahue, Sergey Karayev, Jonathan Long, Ross Girshick, Sergio Guadarrama, and Trevor Darrell. Caffe: Convolutional architecture for fast feature embedding, 2014.

[110] Thorsten Joachims, Thomas Finley, and Chun-Nam John Yu. Cutting-plane training of structural svms. Mach. Learn., 77(1):27–59, October 2009.

[111] A. E. Johnson and M. Hebert. Using spin images for efficient object recognition in cluttered 3d scenes. IEEE Transactions on Pattern Analysis and Machine Intelligence, 21(5):433–449, 1999.

[112] Eskil Jörgensen, Christopher Zach, and Fredrik Kahl. Monocular 3d object detection and box fitting trained end-to-end using intersection-over-union loss, 2019.

[113] Hanwen Kang and Chao Chen. Fruit detection, segmentation and 3d visualisation of environments in apple orchards. Computers and Electronics in Agriculture, 171:105302, Apr 2020.

[114] Lei Ke, Shichao Li, Yanan Sun, Yu-Wing Tai, and Chi-Keung Tang. Gsnet: Joint vehicle pose and shape reconstruction with geometrical and scene-aware supervision. CVPR 2020, 2020.

[115] Wadim Kehl, Fabian Manhardt, Federico Tombari, Slobodan Ilic, and Nassir Navab. Ssd-6d: Making rgb-based 3d detection and 6d pose estimation great again, 2017.

[116] Wadim Kehl, Federico Tombari, Nassir Navab, Slobodan Ilic, and Vincent Lepetit. Hashmod: A hashing method for scalable 3d object detection. In British Machine Vision Conference, 09 2015.

[117] Y. Kim and D. Kum. Deep learning based vehicle position and orientation estimation via inverse perspective mapping image. In 2019 IEEE Intelligent Vehicles Symposium (IV), pages 317–323, 2019.

[118] Roberta L. Klatzky. Allocentric and egocentric spatial representations: Definitions, distinctions, and interconnections. In Lecture Notes in Computer Science, volume 1404. Springer, 1998.

[119] Reinhard Klette. Concise computer vision, an introduction into theory and algorithms. Springer, 2014.

[120] H Kobatake and Y Yoshinaga. Detection of spicules on mammogram based on skeleton analysis. IEEE Trans Med Imaging, 15(3):235–245, 1996.

[121] L. Koval, J. Vaňuš, and P. Bilík. Distance measuring by ultrasonic sensor. IFAC-PapersOnLine, 49(25):153 – 158, 2016. 14th IFAC Conference on Programmable Devices and Embedded Systems PDES 2016.

[122] Alex Krizhevsky, Ilya Sutskever, and Geoffrey Hinton. Imagenet classification with deep convolutional neural networks. Neural Information Processing Systems, 25, 01 2012.

[123] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. Imagenet classification with deep convolutional neural networks. Commun. ACM, 60(6):84–90, May 2017.

[124] Alexander Krull, Eric Brachmann, Frank Michel, Michael Ying Yang, Stefan Gumhold, and Carsten Rother. Learning analysis-by-synthesis for 6d pose estimation in rgb-d images, 2015.

[125] Alexander Krull, Eric Brachmann, Sebastian Nowozin, Frank Michel, Jamie Shotton, and Carsten Rother. Poseagent: Budget-constrained 6d object pose estimation via reinforcement learning. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 12 2016.

[126] Alexander Krull, Frank Michel, Eric Brachmann, Stefan Gumhold, Stephan Ihrke, and Carsten Rother. 6-dof model based tracking via object coordinate regression. In Daniel Cremers, Ian Reid, Hideo Saito, and Ming-Hsuan Yang, editors, Computer Vision – ACCV 2014, pages 384–399, Cham, 2015. Springer International Publishing.

[127] A. Kundu, Y. Li, and J. M. Rehg. 3d-rcnn: Instance-level 3d object reconstruction via render-and-compare. In 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 3559–3568, 2018.

[128] K. Lai, L. Bo, X. Ren, and D. Fox. A large-scale hierarchical multi-view rgb-d object dataset. In 2011 IEEE International Conference on Robotics and Automation, pages 1817–1824, 2011.

[129] Y. Lamdan and H. J. Wolfson. Geometric hashing: A general and efficient model-based recognition scheme. In [1988 Proceedings] Second International Conference on Computer Vision, pages 238–249, 1988.

[130] Todd Lassa. The beginning of the end of driving, November 2012.

[131] Fahad Lateef and Yassine Ruichek. Survey on semantic segmentation using deep learning techniques. Neurocomputing, 338:321 – 348, 2019.

[132] Hei Law and Jia Deng. Cornernet: Detecting objects as paired keypoints, 2018.

[133] Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. D. Jackel. Backpropagation applied to handwritten zip code recognition. Neural Comput., 1(4):541–551, December 1989.

[134] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. Nature, 521(7553):436–444, 2015.

[135] Vincent Lepetit, Francesc Moreno-Noguer, and Pascal Fua. Epnp: An accurate o(n) solution to the pnp problem. International Journal of Computer Vision, 81, 02 2009.

[136] Vincent Lepetit, Luca Vacchetti, Daniel Thalmann, and Pascal Fua. Fully automated and stable registration for augmented reality applications. In Proceedings of the 2nd IEEE and ACM International Symposium on Mixed and Augmented Reality, pages 93– 102, 11 2003.

[137] Bo Li. 3d fully convolutional network for vehicle detection in point cloud. In International Conference on Intelligent Robots and Systems (IROS), pages 1513–1518, 2017.

[138] Bo Li, Tianlei Zhang, and Tian Xia. Vehicle detection from 3d lidar using fully convolutional network, 2016.

[139] Buyu Li, Wanli Ouyang, Lu Sheng, Xingyu Zeng, and Xiaogang Wang. Gs3d: An efficient 3d object detection framework for autonomous driving, 2019.

[140] Chi Li, Jin Bai, and Gregory D. Hager. A unified framework for multi-view multi-class object pose estimation, 2018.

[141] Peixuan Li, Huaici Zhao, Pengfei Liu, and Feidao Cao. Rtm3d: Real-time monocular 3d detection from object keypoints for autonomous driving, 2020.

[142] Yi Li, Gu Wang, Xiangyang Ji, Yu Xiang, and Dieter Fox. Deepim: Deep iterative matching for 6d pose estimation. International Journal of Computer Vision, 128(3):657–678, Nov 2019.

[143] Ming Liang, Bin Yang, Yun Chen, Rui Hu, and Raquel Urtasun. Multi-task multi-sensor fusion for 3d object detection. CVPR 2019, 2020.

[144] Rainer Lienhart and Jochen Maydt. An extended set of haar-like features for rapid object detection. In Proceedings / ICIP ... International Conference on Image Processing, volume 1, pages I–900, 02 2002.

[145] J. J. Lim, H. Pirsiavash, and A. Torralba. Parsing ikea objects: Fine pose estimation. In 2013 IEEE International Conference on Computer Vision, pages 2992–2999, 2013.

[146] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection, 2017.

[147] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft coco: Common objects in context. In David Fleet, Tomas Pajdla, Bernt Schiele, and Tinne Tuytelaars, editors, Computer Vision – ECCV 2014, pages 740–755, Cham, 2014. Springer International Publishing.

[148] Fuchang Liu, Pengfei Fang, Zhengwei Yao, Ran Fan, Zhigeng Pan, Weiguo Sheng, and Huansong Yang. Recovering 6d object pose from rgb indoor image based on two-stage detection network with multi-task loss. Neurocomputing, 337:15 – 23, 2019.

[149] Lijie Liu, Jiwen Lu, Chunjing Xu, Qi Tian, and Jie Zhou. Deep fitting degree scoring network for monocular 3d object detection. CVPR 2019, 2019.

[150] Patrick Langechuan Liu. Monocular 3d object detection in autonomous driving — a review. Blog, 2019.

[151] Patrick Langechuan Liu. Orientation estimation in monocular 3d object detection. blog, October 2019.

[152] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C. Berg. Ssd: Single shot multibox detector. Lecture Notes in Computer Science, pages 21–37, 2016.

[153] Yu-Chih Liu, Kai-Ying Lin, and Yong-Sheng Chen. Bird's-eye view vision system for vehicle surrounding monitoring. In Gerald Sommer and Reinhard Klette, editors, Robot Vision, pages 207–218, Berlin, Heidelberg, 2008. Springer Berlin Heidelberg.

[154] Z. Liu, L. Wang, G. Hua, Q. Zhang, Z. Niu, Y. Wu, and N. Zheng. Joint video object discovery and segmentation by coupled dynamic markov networks. IEEE Transactions on Image Processing, 27(12):5840–5853, 2018.

[155] D. G. Lowe. Object recognition from local scale-invariant features. In Proceedings of the Seventh IEEE International Conference on Computer Vision, volume 2, pages 1150–1157 vol.2, Sep. 1999.

[156] David G. Lowe. Distinctive image features from scale-invariant keypoints. International Journal of Computer Vision, 60(2):91–110, 2004.

[157] Qianhui Luo, Huifang Ma, Yue Wang, Li Tang, and Rong Xiong. 3d-ssd: Learning hierarchical features from rgb-d images for amodal 3d object detection, 2017.

[158] Xinzhu Ma, Zhihui Wang, Haojie Li, Pengbo Zhang, Xin Fan, and Wanli Ouyang. Accurate monocular object detection via color-embedded 3d reconstruction for autonomous driving, 2019.

[159] G. Mamic and M. Bennamoun. Representation and recog- nition of 3d free-form objects. In Digital Signal Processing, volume 12, pages 47–76, 2002.

[160] Fabian Manhardt, Wadim Kehl, and Adrien Gaidon. Roi-10d: Monocular lifting of 2d detection to 6d pose and metric shape, 2018.

[161] Martinez Manuel, Collet Alvaro, and Srinivasa Siddhartha. Moped: A scalable and low latency object recognition and pose estimation system. IEEE International Conference on Robotics and Automation, ICRA 2010, pages 2043–2049, 05 2010.

[162] E. Marchand, H. Uchiyama, and F. Spindler. Pose estimation for augmented reality: A hands-on survey. IEEE Transactions on Visualization and Computer Graphics, 22(12):2633–2651, 2016.

[163] K. Matzen and N. Snavely. Nyc3dcars: A dataset of 3d vehicles in geographic context. In 2013 IEEE International Conference on Computer Vision, pages 761–768, 2013.

[164] Jamshed Memon, Maira Sami, and Rizwan Ahmed Khan. Handwritten optical character recognition (ocr): A comprehensive systematic literature review (slr), 2020.

[165] Frank Michel, Alexander Kirillov, Eric Brachmann, Alexander Krull, Stefan Gumhold, Bogdan Savchynskyy, and Carsten Rother. Global hypothesis generation for 6d object pose estimation, 2016.

[166] Shervin Minaee, Yuri Boykov, Fatih Porikli, Antonio Plaza, Nasser Kehtarnavaz, and Demetri Terzopoulos. Image segmentation using deep learning: A survey. CVPR, 2020.

[167] Chaitanya Mitash, Abdeslam Boularias, and Kostas Bekris. Improving 6d pose estimation of objects in clutter via physics-aware monte carlo tree search. In 2018 IEEE International Conference on Robotics and Automation (ICRA), pages 1–8, 05 2018.

[168] F. Mokhtarian, N. Khalili, and P. Yuen. Multi-scale free-form 3d object recognition using 3d models. Image and Vision Computing, 19(5):271 – 281, 2001.

[169] Tim Morris. Computer vision and image processing. Red Globe Press, 2003.

[170] Tim Morris. Enlarge Computer Vision and Image Processing. Number 320 in ISBN 978-0-333-99451-1. Red Globe Press, 2004.

[171] Arsalan Mousavian, Dragomir Anguelov, John Flynn, and Jana Kosecka. 3d bounding box estimation using deep learning and geometry, 2016.

[172] A. Mukhtar, L. Xia, and T. B. Tang. Vehicle detection techniques for collision avoidance systems: A review. IEEE Transactions on Intelligent Transportation Systems, 16(5):2318–2338, 2015.

[173] Raul Mur-Artal and Juan D. Tardos. Orb-slam2: An open-source slam system for monocular, stereo, and rgb-d cameras. IEEE Transactions on Robotics, 33(5):1255–1262, Oct 2017.

[174] Andretti Naiden, Vlad Paunescu, Gyeongmo Kim, ByeongMoon Jeon, and Marius Leordeanu. Shift r-cnn: Deep monocular 3d object detection with closed-form geometric constraints, 2019.

[175] Vinod Nair and Geoffrey E. Hinton. Rectified linear units improve restricted boltzmann machines. In Proceedings of the 27th International Conference on International Conference on Machine Learning, ICML'10, pages 807–814, Madison, WI, USA, 2010. Omnipress.

[176] Muzammal Naseer, Salman Khan, and Fatih Porikli. Indoor scene understanding in 2.5/3d for autonomous agents: A survey. IEEE Access, 7:1859–1887, 2019.

[177] Chris Nicholson. A beginner's guide to important topics in ai, machine learning, and deep learning, 2019.

[178] Hyeonwoo Noh, Seunghoon Hong, and Bohyung Han. Learning deconvolution network for semantic segmentation, 2015.

[179] D. Nospes, K. Safronov, S. Gillet, K. Brillowski, and U. E. Zimmermann. Recognition and 6d pose estimation of large-scale objects using 3d semi-global descriptors. In 2019 16th International Conference on Machine Vision Applications (MVA), pages 1–6, 2019.

[180] David Novotny, Nikhila Ravi, Benjamin Graham, Natalia Neverova, and Andrea Vedaldi. C3dpo: Canonical 3d pose networks for non-rigid structure from motion. IEEE/CVF International Conference on Computer Vision 2019 (ICCV), 2019.

[181] oceanservice.noaa.gov. What is lidar. National Oceanic and Atmospheric Administration, February 2021.

[182] University of Tasmania. International theses.

[183] University of Tasmania. Open access repository.

[184] M. Ozuysal, V. Lepetit, and P. Fua. Pose estimation for category specific multiview object localization. In 2009 IEEE Conference on Computer Vision and Pattern Recognition, pages 778–785, 2009.

[185] Jiangmiao Pang, Kai Chen, Jianping Shi, Huajun Feng, Wanli Ouyang, and Dahua Lin. Libra r-cnn: Towards balanced learning for object detection. Computer Vision and Pattern Recognition, 2019.

[186] Chavdar Papazov and Darius Burschka. An efficient ransac for 3d object recognition in noisy and occluded scenes. In ACCV 2010 - 10th Asian Conference on Computer Vision, pages 135–148, 01 2010.

[187] Kiru Park, Timothy Patten, and Markus Vincze. Pix2pose: Pixel-wise coordinate regression of objects for 6d pose estimation. 2019 IEEE/CVF International Conference on Computer Vision (ICCV), Oct 2019.

[188] N. Payet and S. Todorovic. From contours to 3d object detection and pose estimation. In 13th International Conference on Computer Vision (ICCV), pages 983–990, 2011.

[189] Scott Pendleton, Hans Andersen, Xinxin Du, Xiaotong Shen, Malika Meghjani, You Eng, Daniela Rus, and Marcelo Jr. Perception, planning, control, and coordination for autonomous vehicles. Machines, 5:6, 02 2017.

[190] Sida Peng, Yuan Liu, Qixing Huang, Hujun Bao, and

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2021.3114399, IEEE Access

Xiaowei Zhou. Pvnet: Pixel-wise voting network for 6dof pose estimation, 2018.

[191] Kelsey Piper. It's 2020. where are our self-driving cars?, 2020.

[192] David A. Forsyth; Jean Ponce. Computer Vision, A Modern Approach. ISBN 978-0-13, 2003.

[193] Victor A. Prisacariu and Ian D. Reid. Pwp3d: Real-time segmentation and tracking of 3d objects. International Journal of Computer Vision, 98(3):335–354, 2012.

[194] Charles R. Qi, Wei Liu, Chenxia Wu, Hao Su, and Leonidas J. Guibas. Frustum pointnets for 3d object detection from rgb-d data, 2017.

[195] Jiaming Qian, Shijie Feng, Tianyang Tao, Yan Hu, Yixuan Li, Qian Chen, and Chao Zuo. Deep-learning-enabled geometric constraints and phase unwrapping for single-shot absolute 3d shape measurement. APL Photonics, 5(4):046105, Apr 2020.

[196] Zengyi Qin, Jinglu Wang, and Yan Lu. Monogrnet: A geometric reasoning network for monocular 3d object localization, 2018.

[197] Mahdi Rad and Vincent Lepetit. Bb8: A scalable, accurate, robust to partial occlusion method for predicting the 3d poses of challenging objects without using depth. ICCV 2017, 2017.

[198] M. M. Rahman, Y. Tan, J. Xue, and K. Lu. Recent advances in 3d object detection in the era of deep neural networks: A survey. IEEE Transactions on Image Processing, 29:2947–2962, 2020.

[199] Akshay Rangesh and Mohan M. Trivedi. Ground plane polling for 6dof pose estimation of objects on the road, 2018.

[200] Esteban Real, Jonathon Shlens, Stefano Mazzocchi, Xin Pan, and Vincent Vanhoucke. Youtube-boundingboxes: A large high-precision human-annotated data set for object detection in video, 2017.

[201] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection, 2015.

[202] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In Advances in neural information processing systems, 2015.

[203] Colin Rennie, Rahul Shome, Kostas E. Bekris, and Alberto F. De Souza. A dataset for improved rgbd-based object detection and pose estimation for warehouse pick-and-place, 2015.

[204] R. Rios-Cabrera and T. Tuytelaars. Discriminatively trained templates for 3d object detection: A real time scalable approach. In 2013 IEEE International Conference on Computer Vision, pages 2048–2055, 2013.

[205] J. Rouillard. Contextual qr codes. In 2008 The Third International Multi-Conference on Computing in the Global Information Technology (iccgi 2008), pages 50–55, 2008.

[206] D. E. Rumelhart, G. E. Hinton, and R. J. Williams. Learning Internal Representations by Error Propagation, pages 318–362. MIT Press, Cambridge, MA, USA, 1986.

[207] Caner Sahin, Guillermo Garcia-Hernando, Juil Sock, and Tae-Kyun Kim. Instance- and category-level 6d object pose estimation, 2019.

[208] Caner Sahin, Guillermo Garcia-Hernando, Juil Sock, and Tae-Kyun Kim. A review on object pose recovery: from 3d bounding box detectors to full 6d pose estimators, 2020.

[209] Caner Sahin and Tae-Kyun Kim. Recovering 6d object pose: A review and multi-modal analysis. In Laura Leal-Taixé and Stefan Roth, editors, Computer Vision – ECCV 2018 Workshops, pages 15–31, Cham, 2019. Springer International Publishing.

[210] H. Sahloul, S. Shirafuji, and J. Ota. 3d affine: An embedding of local image features for viewpoint invariance using rgb-d sensor data. In sensors, volume 19 of https://doi.org/10.3390/s19020291, 2019.

[211] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks, 2018.

[212] S. Savarese and Li Fei-Fei. 3d generic object categorization, localization and pose estimation. In 2007 IEEE 11th International Conference on Computer Vision, pages 1–8, 2007.

[213] Cordelia Schmid and Roger Mohr. Local grayvalue invariants for image retrieval. IEEE Trans. Pattern Anal. Mach. Intell., 19(5):530–535, May 1997.

[214] Henry Schneiderman and Takeo Kanade. A statistical method for 3d object detection applied to faces andcars. In Computer Vision and Pattern Recognition (CVPR), volume 1, pages 746–751 vol.1, 02 2000.

[215] Martin J. Schuster, Korbinian Schmid, Christoph Brand, and Michael Beetz. Distributed stereo vision-based 6d localization and mapping for multi-robot teams. J. Field Robotics, 36:305–332, 2019.

[216] Max Schwarz, Hannes Schulz, and Sven Behnke. Rgb-d object recognition and pose estimation based on pre-trained convolutional neural network features. In IEEE International Conference on Robotics and Automation ((ICRA)), volume 2015, 06 2015.

[217] Pierre Sermanet, David Eigen, Xiang Zhang, Michael Mathieu, Rob Fergus, and Yann LeCun. Overfeat: Integrated recognition, localization and detection using convolutional networks, 2013.

[218] Linda G. Shapiro and George C. Stockman. Computer Vision. Prentice Hall, ISBN 978-0-13-030796-5., 2001.

[219] Linda G. Shapiro and George C. Stockman. Computer Vision. Number ISBN-13: 978-0130307965. Pearson, 2001.

[220] 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 Computer Vision and Pattern Recognition, 2019.

[221] Shaoshuai Shi, Li Jiang, Jiajun Deng, Zhe Wang, Chaoxu Guo, Jianping Shi, Xiaogang Wang, and Hongsheng Li. Pv-rcnn++: Point-voxel feature set abstraction with local vector representation for 3d object detection, 2021.

[222] J Shotton, B Glocker, C. Zach, S. Izadi, A. Criminisi, and A Fitzgibbon. Scene coordinate regression forests for camera relocalization in rgb-d images. In CVPR, 2013.

[223] Abhinav Shrivastava, Rahul Sukthankar, Jitendra Malik, and Abhinav Gupta. Beyond skip connections: Top-down modulation for object detection. CVPR, 2016.

[224] Jennifer Shuttleworth. Automated driving – levels of driving automation are defined in new sae international standard j3016" (pdf). SAE International, J3016, page 2, 2018.

[225] Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. Indoor segmentation and support inference from rgbd images. In Andrew Fitzgibbon, Svetlana Lazebnik, Pietro Perona, Yoichi Sato, and Cordelia Schmid, editors, Computer Vision – ECCV 2012, pages 746–760, Berlin, Heidelberg, 2012. Springer Berlin Heidelberg.

[226] Andrea Simonelli, Samuel Rota Rota Bulò, Lorenzo Porzi, Manuel López-Antequera, and Peter Kontschieder. Disentangling monocular 3d object detection, 2019.

[227] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In ICLR 2015, 2015.

[228] S. Sivaraman and M. M. Trivedi. Looking at vehicles on the road: A survey of vision-based vehicle detection, tracking, and behavior analysis. IEEE Transactions on Intelligent Transportation Systems, 14(4):1773–1795, 2013.

[229] A. W. M. Smeulders, M. Worring, S. Santini, A. Gupta, and R. Jain. Content-based image retrieval at the end of the early years. IEEE Transactions on Pattern Analysis and Machine Intelligence, 22(12):1349–1380, 2000.

[230] J. Sock, S. H. Kasaei, L. S. Lopes, and T. Kim. Multi-view 6d object pose estimation and camera motion planning using rgbd images. In 2017 IEEE International Conference on Computer Vision Workshops (ICCVW), pages 2228–2235, 2017.

[231] Juil Sock, Pedro Castro, Anil Armagan, Guillermo Garcia-Hernando, and Tae-Kyun Kim. Tackling two challenges of 6d object pose estimation: Lack of real annotated rgb images and scalability to number of objects, 03 2020.

[232] Amir Soltani, Huang Haibin, Jiajun Wu, Tejas Kulkarni, and Joshua Tenenbaum. Synthesizing 3d shapes via modeling multi-view depth maps and silhouettes with deep generative networks. In Synthesizing 3D Shapes via Modeling Multi-View Depth Maps and Silhouettes with Deep Generative Networks, 07 2017.

[233] Chen Song, Jiaru Song, and Qixing Huang. Hybridpose: 6d object pose estimation under hybrid representations. CVPR 2020, 2020.

[234] Shuran Song and Jianxiong Xiao. Sliding shapes for 3d object detection in depth images. In David Fleet, Tomas Pajdla, Bernt Schiele, and Tinne Tuytelaars, editors, Computer Vision – ECCV 2014, pages 634–651, Cham, 2014. Springer International Publishing.

[235] Shuran Song and Jianxiong Xiao. Deep sliding shapes for amodal 3d object detection in rgb-d images. CVPR, 2016.

[236] Xibin Song, Peng Wang, Dingfu Zhou, Rui Zhu, Chenye Guan, Yuchao Dai, Hao Su, Hongdong Li, and Ruigang Yang. Apollocar3d: A large 3d car instance understanding benchmark for autonomous driving, 2018.

[237] Xibin Song, Peng Wang, Dingfu Zhou, Rui Zhu, Chenye Guan, Yuchao Dai, Hao Su, Hongdong Li, and Ruigang Yang. Apollocar3d: A large 3d car instance understanding benchmark for autonomous driving. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 5452–5462, 2019.

[238] Milan Sonka, Vaclav Hlavac, and Roger Boyle. Image Processing, Analysis, and Machine Vision. Number 3216-7. Springer, Boston, MA, second edition, 1993.

[239] G. Spampinato, J. Lidholm, C. Ahlberg, F. Ekstrand, M. Ekström, and L. Asplund. An embedded stereo vision module for 6d pose estimation and mapping. In 2011 IEEE/RSJ International Conference on Intelligent Robots and Systems, pages 1626–1631, 2011.

[240] Martin Sundermeyer, Zoltan-Csaba Marton, Maximilian Durner, Manuel Brucker, and Rudolph Triebel. Implicit 3d orientation learning for 6d object detection from rgb images. CVPR 2019, 2019.

[241] Christian Szegedy, Sergey Ioffe, Vincent Vanhoucke, and Alex Alemi. Inception-v4, inception-resnet and the impact of residual connections on learning. CVPR, 2016.

[242] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions, 2014.

[243] Christian Szegedy, Scott Reed, Dumitru Erhan, Dragomir Anguelov, and Sergey Ioffe. Scalable, high-quality object detection. arXiv preprint arXiv:1412.1441, 2014, 2014.

[244] Araz Taeihagh and Hazel Si Min Lim. Governing autonomous vehicles: emerging responses for safety, liability, privacy, cybersecurity, and industry risks. Transport Reviews, 39(1):103–128, 2019.

[245] J. Taylor, J. Shotton, T. Sharp, and A Fitzgibbon. The

vitruvian manifold: Inferring dense correspondences for one-shot human pose estimation. In CVPR, 2012.

[246] Alykhan Tejani, Danhang Tang, Rigas Kouskouridas, and Tae-Kyun Kim. Latent-class hough forests for 3d object detection and pose estimation. In David Fleet, Tomas Pajdla, Bernt Schiele, and Tinne Tuytelaars, editors, Computer Vision – ECCV 2014, pages 462–477, Cham, 2014. Springer International Publishing.

[247] Bugra Tekin, Sudipta N. Sinha, and Pascal Fua. Real-time seamless single shot 6d object pose prediction. CVPR, 2017.

[248] D. Tomè, F. Monti, L. Baroffio, L. Bondi, M. Tagliasacchi, and S. Tubaro. Deep convolutional neural networks for pedestrian detection. Signal Processing: Image Communication, 47:482–489, Sep 2016.

[249] Jonathan Tremblay, Thang To, and Stan Birchfield. Falling things: A synthetic dataset for 3d object detection and pose estimation. CVPR, 04 2018.

[250] Jonathan Tremblay, Thang To, Balakumar Sundaralingam, Yu Xiang, Dieter Fox, and Stan Birchfield. Deep object pose estimation for semantic robotic grasping of household objects, 2018.

[251] Ioannis Tsochantaridis, Thomas Hofmann, Thorsten Joachims, and Yasemin Altun. Support vector machine learning for interdependent and structured output spaces. In Proceedings of the Twenty-First International Conference on Machine Learning, ICML '04, page 104, New York, NY, USA, 2004. Association for Computing Machinery.

[252] J. R. R. Uijlings, K. E. A. van de Sande, T. Gevers, and A. W. M. Smeulders. Selective search for object recognition. International Journal of Computer Vision, 104(2):154–171, 2013.

[253] Greg Van Houdt, Carlos Mosquera, and Gonzalo Nápoles. A review on the long short-term memory model. Artificial Intelligence Review, 53(8):5929–5955, 2020.

[254] Joel Vidal, Chyi-Yeu Lin, Xavier Llado, and Robert Martí. A method for 6d pose estimation of free-form rigid objects using point pair features on range data. Sensors, 18:2678, 08 2018.

[255] Joel Vidal, Chyi-Yeu Lin, and Robert Marti. 6d pose estimation using an improved method based on point pair features. In 2018 4th International Conference on Control, Automation and Robotics (ICCAR), pages 405–409, 04 2018.

[256] D. Wagner, G. Reitmayr, A. Mulloni, T. Drummond, and D. Schmalstieg. Pose tracking from natural features on mobile phones. In 2008 7th IEEE/ACM International Symposium on Mixed and Augmented Reality, pages 125–134, 2008.

[257] Chen Wang, Danfei Xu, Yuke Zhu, Roberto Martín-Martín, Cewu Lu, Li Fei-Fei, and Silvio Savarese. Densefusion: 6d object pose estimation by iterative dense fusion, 2019.

[258] Chien-Yao Wang, I-Hau Yeh, and Hong-Yuan Mark Liao. You only learn one representation: Unified network for multiple tasks, 2021.

[259] Dominic Zeng Wang and Ingmar Posner. Voting for voting in online point cloud object detection. In Robotics: Science and Systems, 2015.

[260] He Wang, Srinath Sridhar, Jingwei Huang, Julien Valentin, Shuran Song, and Leonidas J. Guibas. Normalized object coordinate space for category-level 6d object pose and size estimation, 2019.

[261] Tai Wang, Xinge Zhu, Jiangmiao Pang, and Dahua Lin. Fcos3d: Fully convolutional one-stage monocular 3d object detection. CVPR 2020, 2021.

[262] Xinlong Wang, Wei Yin, Tao Kong, Yuning Jiang, Lei Li, and Chunhua Shen. Task-aware monocular depth estimation for 3d object detection, 2019.

[263] 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, 2018.

[264] Xinshuo Weng and Kris Kitani. Monocular 3d object detection with pseudo-lidar point cloud, 2019.

[265] HUGH DURRANT WHYTE and TIM BAILEY. Simultaneous localization and mapping. IEEE Robotics and Automation Magazine, 2(13):99–110, 2006.

[266] KYLE WIGGERS. Facebook highlights ai that converts 2d objects into 3d shapes. Online Blog, October 2019.

[267] Kyle Wiggers. Google brings cross-platform ai pipeline framework, January 2020.

[268] Di Wu, Zhaoyong Zhuang, Canqun Xiang, Wenbin Zou, and Xia Li. 6d-vnet: End-to-end 6-dof vehicle pose estimation from monocular rgb images. In The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, June 2019.

[269] Zuxuan Wu, Xi Wang, Yu-Gang Jiang, Hao Ye, and Xiangyang Xue. Modeling spatial-temporal clues in a hybrid deep learning framework for video classification, 2015.

[270] Y. Xiang, Wongun Choi, Y. Lin, and S. Savarese. Data-driven 3d voxel patterns for object category recognition. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 1903–1911, 2015.

[271] Yu Xiang, Wongun Choi, Yuanqing Lin, and Silvio Savarese. Subcategory-aware convolutional neural networks for object proposals and detection, 2016.

[272] Yu Xiang, Wonhui Kim, Wei Chen, Jingwei Ji, Chris Choy, Hao Su, Roozbeh Mottaghi, Leonidas Guibas, and Silvio Savarese. Objectnet3d: A large scale database for 3d object recognition. In European Conference on Computer Vision, volume 9912, pages 160–176, 10 2016.

[273] Yu Xiang, Roozbeh Mottaghi, and Silvio Savarese. Beyond pascal: A benchmark for 3d object detection in the wild. In IEEE Winter Conference on Applica-

tions of Computer Vision (WACV), pages 75–82, 03 2014.

[274] Yu Xiang and Silvio Savarese. Estimating the aspect layout of object categories. In IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pages 3410–3417, 06 2012.

[275] Yu Xiang, Tanner Schmidt, Venkatraman Narayanan, and Dieter Fox. Posecnn: A convolutional neural network for 6d object pose estimation in cluttered scenes. Computer Vision and Pattern Recognition (CVPR), 2017.

[276] X.Mao, D.Inoue, S.Kato, and M.Kagami. mplitude-modulatedlaser radar for range and speed measurement in car applications. IEEE Trans. Intell. Transp. Syst, 13(1):408–413, 2012.

[277] Bin Xu and Zhenzhong Chen. Multi-level fusion based 3d object detection from monocular images. In CVPR, pages 2345–2353, 06 2018.

[278] Bin Yang, Wenjie Luo, and Raquel Urtasun. Pixor: Real-time 3d object detection from point clouds, 2019.

[279] 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, 2019.

[280] Honglin Yuan, Tim Hoogenkamp, and Remco C. Veltkamp. Robotp: A benchmark dataset for 6d object pose estimation. Sensors, 21(4), 2021.

[281] Sergey Zakharov, Ivan Shugurov, and Slobodan Ilic. Dpod: 6d pose object detector and refiner, 2019.

[282] Sergey Zakharov, Ivan Shugurov, and Slobodan Ilic. Dpod: 6d pose object detector and refiner, 2019.

[283] Matthew D Zeiler and Rob Fergus. Visualizing and understanding convolutional networks, 2013.

[284] Matthew D. Zeiler, Dilip Krishnan, Graham W. Taylor, and Rob Fergus. Deconvolutional networks. Number 2528–2535 in 10.1109/CVPR.2010.5539957. Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 9781424469840 edition, 08 2010.

[285] Haoruo Zhang and Qixin Cao. Fast 6d object pose refinement in depth images. Applied Intelligence, 49(6):2287–2300, 2019.

[286] Shifeng Zhang, Longyin Wen, Xiao Bian, Zhen Lei, and Stan Z. Li. Single-shot refinement neural network for object detection, 2017.

[287] Tielin Zhang, Yang Yang, Yi Zeng, and Yuxuan Zhao. Cognitive template-clustering improved linemod for efficient multi-object pose estimation. Cognitive Computation, 2020.

[288] Xin Zhang, Zhiguo Jiang, and Haopeng Zhang. Real-time 6d pose estimation from a single rgb image. Image and Vision Computing, 89:1 – 11, 2019.

[289] Xin Zhang, Zhiguo Jiang, and Haopeng Zhang. Out-of-region keypoint localization for 6d pose estimation. Image and Vision Computing, 93:103854, 2020.

[290] Y. Zhang, H. Zhang, G. Wang, J. Yang, and J. Hwang. Bundle adjustment for monocular visual odometry based on detections of traffic signs. IEEE Transactions on Vehicular Technology, 69(1):151–162, 2020.

[291] Yanan Zhang, Di Huang, and Yunhong Wang. Pc-rgnn: Point cloud completion and graph neural network for 3d object detection. CVPR 2020, 2020.

[292] Jianfeng Zhao, Bodong Liang, and Qiuxia Chen. The key technology toward the self-driving car. In International Journal of Intelligent Unmanned Systems, 2018.

[293] Zhong-Qiu Zhao, Haiman Bian, Donghui Hu, Wenjuan Cheng, and Hervé Glotin. Pedestrian detection based on fast r-cnn and batch normalization. In Intelligent Computing Theories and Application. ICIC 2017, pages 735–746, 07 2017.

[294] Zhong-Qiu Zhao, Peng Zheng, Shou tao Xu, and Xindong Wu. Object detection with deep learning: A review, 2018.

[295] Zhe Cao, Y. Sheikh, and N. K. Banerjee. Real-time scalable 6dof pose estimation for textureless objects. In 2016 IEEE International Conference on Robotics and Automation (ICRA), pages 2441–2448, 2016.

[296] Wu Zheng, Weiliang Tang, Sijin Chen, Li Jiang, and Chi-Wing Fu. Cia-ssd: Confident iou-aware single-stage object detector from point cloud. CVPR 2020, 2020.

[297] Wu Zheng, Weiliang Tang, Li Jiang, and Chi-Wing Fu. Se-ssd: Self-ensembling single-stage object detector from point cloud. CVPR 2020, 2021.

[298] Zhenheng Yang and R. Nevatia. A multi-scale cascade fully convolutional network face detector. In 2016 23rd International Conference on Pattern Recognition (ICPR), pages 633–638, 2016.

[299] Y. Zhong. Intrinsic shape signatures: A shape descriptor for 3d object recognition. In 2009 IEEE 12th International Conference on Computer Vision Workshops, ICCV Workshops, pages 689–696, 2009.

[300] Dingfu Zhou, Yuchao Dai, and Hongdong li. Ground plane based absolute scale estimation for monocular visual odometry. In IEEE Transactions on Intelligent Transportation Systems, 03 2019.

[301] Dingfu Zhou, Jin Fang, Xibin Song, Liu Liu, Junbo Yin, Yuchao Dai, Hongdong Li, and Ruigang Yang. Joint 3d instance segmentation and object detection for autonomous driving. In 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pages 1836–1846, 2020.

[302] T. Zhou, M. Brown, N. Snavely, and D. G. Lowe. Unsupervised learning of depth and ego-motion from video. In 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 6612–6619, 2017.

[303] Xingyi Zhou, Dequan Wang, and Philipp Krähenbühl. Objects as points, 2019.

[304] Xingyi Zhou, Jiacheng Zhuo, and Philipp Krähenbühl. Bottom-up object detection by grouping extreme and

center points, 2019.

[305] Wentao Zhu, Jun Miao, Jiangbi Hu, and Laiyun Qing. Vehicle detection in driving simulation using extreme learning machine. Neurocomputing, 128:160 – 165, 2014.