3D Object Detection and Tracking Methods using Deep Learning for Computer Vision Applications

Shreyas E
Electronics and Telecommunication Engineering, RV College of Engineering®,
Bengaluru, India

Manav Hiren Sheth
Electronics and Telecommunication Engineering, RV College of Engineering®,
Bengaluru, India

Mohana
Electronics and Telecommunication Engineering, RV College of Engineering®,
Bengaluru, India

2021 International Conference on Recent Trends on Electronics, Information, Communication & Technology (RTEICT) | 978-1-6654-3559-8/21/$31.00 ©2021 IEEE | DOI: 10.1109/RTEICT52294.2021.9573964

***Abstract*- 3D multi-object detection and tracking is an essential constituent for many applications in today's world. Object detection is a technology related to computer vision and image processing that allows us to detect instances of certain classes. There are numerous applications like robotics, autonomous driving and augmented reality. A bounding box often defines the region of interest and then is used to classify into respective categories. Due to identical appearance and shape of various objects and the interference of lighting and shielding, object detection has always been a challenging problem in computer vision. Conventional 2D object detection yields four degrees of freedom axis-aligned bounding boxes with centre (x, y) and 2D size (w, h), the 3D bounding boxes generally have 6 Degrees of freedom: 3D physical size (w, h, l), 3D centre location (x, y, z). 2D object detection and tracking methods do not provide depth information to perform essential tasks in various computer vision applications. One among them is the Autonomous driving. 3D object detection includes depth information that provides more information on the structure of detected object. More information is required to make decisions accurately in different fields where 3D object detection and tracking can be applied. In this paper various 3D object detection and tracking methods are elaborated for various computer vision applications, this includes various fields such as robotics, driving, space field and also in the military.**

Keywords— 3D Object Detection, 3D Object Tracking, Computer Vision, Video Surveillance, Edge Computing, Edge Devices, Graph Neural Network (GNN).

I. Introduction

Presently there are various 3D object detection and tracking methods. These methods use various sensor and data capturing technologies and use them in combination to achieve more accurate results and to improve performance metrics. They can also be classified based on the data points they use. There are several standard datasets developed, some of these datasets are developed for general purpose application, however there are datasets that have been developed for task specific usage. Irrespective of the frameworks and the datasets used, the 3D evaluation metrics remain the same for all the frameworks. Evaluating them on a general benchmark provides us an opportunity to compare these frameworks. There are several popular 3D evaluation metrics developed and there is a need for new evaluation metrics too. Some among them are being developed for evaluating frameworks on their specific use cases. However there are several challenges that the present frameworks face that need to be addressed. Novel technologies are solving some of the present issues. Some of these issues include the need for resources to perform computationally intensive task, occlusions and different weather conditions.

II. Literatures Survey

E. Arnold [1] analyzes different sensor technologies comparing their advantages and disadvantages. They give an overview of different types of 3D object detection and tracking methods particularly those using sensors and dataset for autonomous driving application. X. Weng [2] explains the need for different evaluation techniques for 3D object detection and why old metrics do not give the best indication for 3D applications. They introduce metrics sMOTA, AMOTA and AMOTP which are derived from the old metrics and also give a performance evaluation of the new metrics on KITTI dataset. C. Badue [3] presented a research survey on self-driving cars. It mainly focused on present and available research that has been implemented in the real world. A. Gupta [4] discuss comprehensive methods that address image perception issues and assess recently developed frameworks and tests conducted on self-driving cars using deep learning perspectives [27] [28] [29]. They discuss different techniques in deep learning that can tackle issues in AV's and evaluate these techniques on recent implementations. P. Li [5] have categorized the issue of 3D object detection into two problems. One of them being visual correlation. The presented model is centered on Faster R-CNN for visual correlation, which estimates precise 3D object measurements along with their local orientation. The other problem being geometric correlation, they have introduced a novel geometric constraint for geometrics correlation. This method uses monocular images and outperforms other 3D detection methods. It doesn't use any extra labels. E. A. Tasdelen [6] proposed multiple 3D LiDAR sensors can be used to detect and track objects with high accuracy. They apply and test multiple LIDAR fusion strategies on real vehicles and has also given a comparison of the strategies. R. Khamsehashari [7] presented an approach for robust detection of small and occluded objects. It is based on an early fusion approach and it performs without any need for additional network parameters. The sensory system data includes LiDAR and camera data. A. V. Kozov [8] presented the issues of high performance obstacle detection. It clearly describes problems, different methods of approach and the sensory systems involved. Author analyses the different detection techniques, neural networks and other parameters. M. Kusenbach [9] presented a method for effectively using LiDAR data. It creates a 2D structure of object based on rotation of LiDAR sensor and the way it gathers data. This approach allows quick accessing of neighboring information on spatial and temporal level. Scott McCrae [10] discusses how latency issues and data-hungry nature of temporally-aware feed-forward networks can be solved using recurrent networks. The analysis of this method

978-1-6654-3559-8/21/$31.00 ©2021 IEEE

presents that there is an increased detection accuracy of pedestrians. A. Agafonov [11] analyses how effective 3D objects detection and localization method and also concludes that training models on synthetic data will reduce the accuracy of detecting objects on real data. K. Zhao [12] presented a novel detection method using LiDAR data for occluded vehicles. It has contributions from the following techniques: fusion of different data representation to strongly detect the features of occluded vehicles. It has a slimmer 3DOV network that handles the speed-accuracy tradeoff. Kuk Cho [13] propose a novel detection and tracking mechanism that can successfully track detected objects and prevent wrong detection and tracking using concepts of object classes and relative positioning. Yilin Wang [14] propose a fusion based framework that combines both RGB and point cloud data respectively to detect multiple classes. The region of interest (ROI) is localized by using RGB image using the existing 2D detection models, which is then processed for a pixel mapping strategy in the point cloud. After which the initial 2D bounding box is lifted to 3D space. Z. Yang [15] recognizes that many existing methods focus on using single frames instead of completely using information from multiple frames to perform 3D detection. Author also introduces 3D-MAN: a 3D multi-frame attention network that aggregates features from multiple perspectives to achieve state-of-the-art performance. Zili Liu [16] propose sparse point which is a 3D object detection method that uses sparse methods. In this method, most likely possible positions of 3D objects to be detected are encoded using sparse points, which further encodes shared semantic features of all the objects. S. Luo [17] proposes a two-step method for feature alignment. First step is to enable receptive field of the feature map to focus on predefined anchors to the shape alignment. The features of 2D/3D centers are aligned using center alignment. For the depth prediction of objects, global information and capture long-range relationships are required, and are often difficult to capture. Zhichao Li [18] presented a second stage detector which improves current 3D detector, LiDAR R-CNN. It takes a point based approach to full-fill the real-time and high precision requirements rather than a conventional voxel-based approach. Jongyoun Noh [19] introduce a novel 3D detection method that combines merits of voxel and point based features. Point based features represent the structures more accurately however extracting these features are computationally expensive. On the other hand, Voxel-based features are efficient to extract, however they fail to conserve fine-grained structures. Jin Fang [20] presents Map-fusion, which helps to integrate the present map information into 3D detectors. It is a simple and effective framework. The design includes a module for HD map feature extraction and fusion called the FeatureAgg and other module that forms head for the detection backbone, which is known as MapSeg module. They achieve better points for mean Average Precision (mAP) by the fusion of map information with 3d object detection. G. Zhai [21] presented a LiDAR based 3D MOT framework called FlowMOT. This framework enhances robustness of the motion prediction by integrating point-wise data with the traditional matching problem. Edge-based 3D tracking framework, has many advantages as compared to other tracking methods [26] [27].

III. Popular dataset used for 3D object detection and Tracking

One challenge is the availability of large datasets. There has been significant progress in the creation and maintenance of datasets. Some of the popular datasets available for 3D object detection are -

CamVid - It stands for Cambridge-driving labeled video database. The frames were manually annotated with 32 different classes commonly found on the road.

Objectron - It is a collection of 3D object detection models specially trained for mobile devices. This dataset has been wholly annotated with 3D bounding boxes.

KITTI - Dataset contains various vision tasks built using an autonomous driving platform. Which includes stereo, optical flow, and visual odometry benchmark. And it consists of monocular images and bounding boxes.

H3D (Honda Research Institute 3D) - It is a large-scale dataset. Its features are - Full 360-degree LiDAR dataset, 160 traffic scenes, 1071302 3D bounding box labels, eight common classes of traffic participants which are benchmarked on SOTA algorithms for 3D object detection and tracking algorithms.

nuScenes - Dataset used for Autonomous Vehicles. This dataset has been collected in Boston and Singapore and has 3D bounding boxes for over 1000 videos.

Argoverse - It is a tracking benchmark with over 30,000 scenarios collected in Pittsburgh and Miami. The sequences are split into training, validation, and test sets, which have 205,942, 39,472, and 78,143 sequences. These splits have no geographical overlap.

Lyft L5 - A dataset used for motion prediction. Huge video dataset collected by 20 AVs in California over four months. It captures perception output of self-driving system with precise positions and motions of nearby objects.

Object Net 3D - It is created by integrating existing 2D image repositories and aligning these to 3D shapes by researchers from Stanford University. Images were taken from ImageNet and Google Images with 3D annotations.

*A*3D* - Dataset is used for autonomous driving, contains LiDAR point cloud frames and RGB images. These images are captured at diverse weather conditions.

A2D2 - This dataset contains frames with semantic segmentation of image and point cloud labels and also have annotations for 3D bounding boxes. There is also unlabeled sensor data for sequences with several loops.

IV. 3D Object detection Models and Frameworks

RGB Images - The first method includes 3D object detection using RGB images. RGB images provide sufficient semantic information and hence can be used in 3D object detection. Methods like 3D- GCK employ this method to use available semantic information to generate depth information using neural networks. Hence by estimating depth information, 2D bounding boxes are converted into 3D bounding boxes. Another method of 3D detection using images includes the use of an RGB-D sensor. This sensor first converts input RGB images into grayscale images. Further, it segments the background and foreground hence providing specific information. After segmenting, the noise is removed. Further

classification models are applied for object detection and classification.

Point Cloud 3D object detection – There are two types of methods that use point cloud data. The first method uses all 3D point cloud data. This method has the most negligible information loss; however, it is computationally costly.

The other method includes using a reduced number of 3D point cloud data and making it computationally feasible. The only disadvantage of this method is that it processes less information than other methods. BirdNet+ is a LiDar based 3D object detection method. The input is a 2D bird's eye view with channels from LiDAR point clouds. This method relies on two-stage detectors for 3D detection. Deep Point Cloud Mapping Network (DPC-MN) employs an unsupervised deep learning model to map the 2D view from a 3D point cloud.

Factor graph based 3D Multi Object Tracking in Point Clouds. This approach utilizes 3D point clouds and RGB images. This method of 3D object detection does not rely on explicit data association methods that have been used. Instead, a novel approach has been applied. Robust optimization of the backend leads to solving the assignment problem implicitly and jointly. Hence this approach assigns all the detections simultaneously. The challenging task of assigning the detection in the present frame and prediction of objects in the previous frame has been simplified. Hence this algorithm achieves better results than many state of the art algorithms. The optimization of this assignment has been incorporated in a factor graph framework.

LiDAR - In this method, RGB images and LiDAR cloud points are used for object detection. This method uses 2 dimensional object detectors to reduce the region of interest and reduce the search space and further extract other distinct information from point clouds. Frustum Point Nets is one such framework that uses this method. Another framework that uses a similar method is MV3D.

V. Graph neural network (GNN) in 3D object detection and tracking.

This method can learn more discriminative object features. Learning more discriminative features of different objects reduces confusion during data association. In this new approach, object features are obtained from 2D and 3D space and then the features of the object are updated through the interaction of graph neural networks. This method produces stable trajectories overtime [22].

VI. Edge based 3D object detection and tracking.

Edge based tracking methods offer many advantages as compared to other tracking methods. The methods are mainly categorized into two, methods with and those without straight edges. Methods that follow tracking without explicit edges depend more on detecting strong gradients in each frame. However, methods that follow tracking with explicit edges detect more features.

VII. Quantum convolutional neural network (QCNN) based 3D object detection and tracking

Quantum Convolutional Neural Network offers a powerful computational model. This has applications in many fields including machine learning, deep learning. Many algorithms are being developed for evaluating and training deep convolutional neural networks. This will speed up the evaluation and training over classical CNNs for both passes i.e. the forward and backward pass. QCNN is particularly interesting and promising for deep neural networks and could open up new possibilities in the computer vision field [23]. Hence this would definitely help to use computationally exhaustive methods in the field of 3D object detection and tracking.

VIII. 3D Evaluation Metrics

TABLE. I. POPULAR 3D EVALUATION METRICS FOR COMPUTER VISION

Evaluation metric	Abbreviation
sAMOTA	Scaled Multiple Object Tracking Accuracy
AMOTA	Average Multiple Object Tracking Accuracy
AMOTP	Average Multiple Tracking Precision
IDS	Number of identity switches
FRAG	Number of trajectory fragmentation
FPS	Frames per second

Table 1 shows the popular 3D evaluation metrics used in computer vision applications

IX. Applications of 3D object detection and Tracking

Autonomous driving - 3D object detection and tracking are crucial tasks for AV's. The most important processes during autonomous driving include prediction, planning, and motion control. This requires an accurate understanding of the 3D space around the vehicle.

Augmented reality - is an interactive experience of a real-world environment where the objects that reside in the real world are enhanced by computer-generated perceptual information.

Vision guided robotic systems - The ability of a robot to be a completely autonomous body depends on its ability to perceive its environment and interpret it. The 3D object detection for a vision guided robot requires object point cloud data, class labels, and 3D bounding boxes.

X. Key players and Future technology

Space and Aerospace industry - Object detection in space can work as a GPS on earth. It can also prevent collisions with space obstacles such as asteroids, comets, and meteors.
Security related companies - Security companies invest in technological aids that help them provide better security, reducing human effort and intervention. The availability of suitable hardware materials and technology has been a significant contribution to this shift.

Social Media giants like Google and Facebook are developing 3D detection techniques to provide a user-friendly interactive experience to enter a more immersive and interactive environment.

Automobile Companies - Automobile companies are in a race to produce more efficient self-driving cars. engaging this technology for their cars will definitely help them in a huge way.

Medical Imaging - 3D object detection can be used to semantically interpret the interior of the body for analysis.

This provides an enriched image of the body organs, tissues and other body parts.

XI. Research and Real time implementation challenges

Visual change detection includes automated detection of transitions between frames where the video is first segmented into frames and then processed. There are numerous applications of change detection including behavior analysis, action recognition and video surveillance. The challenges in change detection are the same as in case of object detection and tracking. It also includes background fluctuations, intermittent object motion, shadow, fast/slow object motion, camera motion, heterogeneous object shapes and real-time processing of these illumination variations.

The computer vision application has a varying degree of performance in different light conditions. Hence low light imaging cameras can be used along with other sensory instruments that help retrieve accurate information. However, the calibration of information retrieved through sensors is a challenge.

During Night conditions, this challenge is countered by using night vision cameras. However, using night vision cameras does not give a sense of color. Hence thermal images are used along with night cameras to get information about color perception.

The computing speed is one of the areas to be worked upon. Almost all applications include time as a critical component, and hence there is a need for addressing the computing power issues. In addition, new methods that would help compute large amounts of information with a limited resource are being researched.

3D Object detection range is among the most discussed topics. Various models are produced from a combination of the present models to improve the accuracy of 3D object detection and tracking. The models have faced the following challenges such as fluctuations in background, light variation, change in weather, camera motion and processing real-time data.

Conclusions

With the advancement of new technologies, there are many new methods and opportunities for development and these technologies also solve present issues in the field of computer vision. This paper reviews various 3D object detection and tracking methods. Different sensor and data capturing technologies are elaborated. Standard datasets with their typical use cases are also discussed. Along with this popular 3D evaluation metrics are mentioned. Frameworks used for 3D object detection are explored. Present applications with the future prospects and the key players for this developments are discussed in detail.

References

[1] E. Arnold, et all., "A Survey on 3D Object Detection Methods for Autonomous Driving Applications," *IEEE Transactions on Intelligent Transportation Systems,* 2019.

[2] X. Weng, et all., "3D Multi-Object Tracking: A Baseline and New Evaluation Metrics," IEEE/RSJ International Conference on Intelligent Robots and Systems, 2020.

[3] C. Badue et al., "Self-Driving Cars: A Survey," arXiv [cs.RO], 2019.

[4] A. Gupta et all.,"Deep learning for object detection and scene perception in self-driving cars: Survey, challenges, and open issues," 2021.

[5] P. Li et all.,"Monocular 3D object detection using dual quadric for autonomous driving," 2021.

[6] E. A. Tasdelen et all.,"Comparison and application of multiple 3D LIDAR fusion methods for object detection and tracking,"*International Conference on Robotics and Automation Engineering (ICRAE),* 2020.

[7] R. Khamsehashari et all.,"Improving Deep Multi-modal 3D Object Detection for Autonomous Driving," *International Conference on Automation, Robotics and Applications (ICARA),*2021.

[8] A. V. Kozov et all., "Structural Obstacle Recognition Method and Its Application in Elevated Terrain Objects Search,"*International Russian Automation Conference (RusAutoCon),* 2018.

[9] M. Kusenbach et all., "Enhanced Temporal Data Organization for LiDAR Data in Autonomous Driving Environments," *IEEE Intelligent Transportation Systems Conference (ITSC), Auckland, New Zealand*, 2019.

[10] S. McCrae et all., "3D Object Detection For Autonomous Driving Using Temporal Lidar Data," *IEEE International Conference on Image Processing (ICIP),* 2020.

[11] A. Agafonov et all., "3D Objects Detection in an Autonomous Car Driving Problem," *International Conference on Information Technology and Nanotechnology (ITNT),* Samara, Russia, 2020.

[12] K. Zhao et all., "3D Detection for Occluded Vehicles From Point Clouds,"*IEEE Intelligent Transportation Systems Magazine.*

[13] Kuk Cho et all., "Real-time 3D multiple occluded object detection and tracking," *IEEE ISR 2013*, Seoul, Korea (South), 2013.

[14] Y. Wang et all.,"An Overview Of 3D Object Detection" *arXiv:2010.15614v1 [cs.CV],* 2020

[15] Z.Yang et all.,"3D-MAN: 3D Multi-frame Attention Network for Object Detection "*arXiv:2103.16054v1 [cs.CV],*2021.

[16] Z. Liu et all.,"SparsePoint: Fully End-to-End Sparse 3D Object Detector" *arXiv:2103.10042v1 [cs.CV],* 2021.

[17] Shujie Luo et all.,"M3DSSD: Monocular 3D Single Stage Object Detector"*arXiv:2103.13164v1 [cs.CV],* 2021.

[18] Z. Li et all.,"LiDAR R-CNN:An Efficient and Universal 3D Object Detector"*arXiv:2103.15297v1,* 2021.

[19] J Noh et all.,"HVPR: Hybrid Voxel-Point Representation for Single-stage 3D Object Detection "*arXiv:2104.00902v1,* 2021.

[20] J.Fang et all.,"MapFusion: A General Framework for 3D Object Detection with HDMaps" *arXiv:2103.05929v1,* 2021.

[21] Guangyao Zhai et all.,"FlowMOT: 3D Multi-Object Tracking by Scene Flow Association"*arXiv:2012.07541v3,* 2021.

[22] Pradhyumna P et al.,"Graph Neural Network (GNN) in Image and Video Understanding Using Deep Learning for Computer Vision Applications," *2nd International Conference on Electronics and Sustainable Communication Systems(ICESC)*,2021.

[23] Varadi Rajesh et al.,"Quantum Convolutional Neural Networks (QCNN) Using Deep Learning for Computer Vision Applications," *6th International Conference on Recent Trends in Electronics, Information & Communication Technology (RTEICT),* 2021.

[24] Rohith M et al.,"Comparative Analysis of Edge Computing and Edge Devices: Key Technology in IoT and Computer Vision Applications," *6th International Conference on Recent Trends in Electronics, Information & Communication Technology (RTEICT),* 2021.

[25] Biswas A et al.," Survey on Edge Computing–Key Technology in Retail Industry" *Lecture Notes on Data Engineering and Communications Technologies,* 2021, vol 58. Springer, Singapore.

[26] Mohana et al.,"Performance Evaluation of Background Modeling Methods for Object Detection and Tracking," Fourth *International Conference on Inventive Systems and Control (ICISC)*, 2020.

[27] N. Jain et al.,"Performance Analysis of Object Detection and Tracking Algorithms for Traffic Surveillance Applications using Neural Networks," *International conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC)*, 2019.

[28] C. Kumar B et al.,"YOLOv3 and YOLOv4: Multiple Object Detection for Surveillance Applications," *International Conference on Smart Systems and Inventive Technology (ICSSIT),* 2020.

C. Kumar B et al.," Performance Analysis of Object Detection Algorithm for Intelligent Traffic Surveillance System," *International Conference on Inventive Research in Computing Applications (ICIRCA)*, 2020.