A Survey on Key-point Based Monocular 3D Object Detection for Autonomous Driving

Tao Peng, ByeongWoo Kim
Department of Electrical, Electronic and Computer Engineering, University of Ulsan
e-mail: pengtao1015@naver.com

자율주행을 위한 키포인트 기반 단안 3D 물체 감지 연구

팽도, 김병우
울산대학교 전기전자컴퓨터공학과

Abstract

Monocular 3D object detection, one of the most fundamental and difficult challenges in autonomous driving, has gotten a lot of attention from industry and academia in recent years. Key-point based monocular 3D object detection has made significant progress as a result of the development of monocular 3D object detection. We compiled and organized the most recent research in this field, summarizing and analyzing each component of the most widely used pipelines for key-point based monocular 3D object detection. Furthermore, we propose a classification method that divides monocular 3D object detection into four categories so that monocular 3D object detection algorithms can be systematically sorted and fairly compared. We also discuss current challenges in key-point methods and future directions for key-point based monocular 3D object detection research.

Keypoints: Monocular camera; 3D Object Detection; Key-point.

1. Introduction

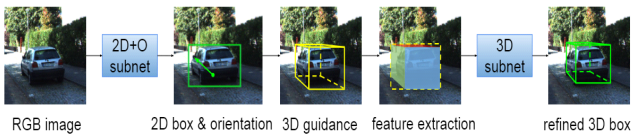
3D detection is a critical task in the field of autonomous driving. Many tasks in this field require the perception and expression of a good 3D space around the unmanned vehicle, such as task decision-making, path planning, and motion control. In recent years, algorithms based on 3D Lidar have greatly improved the accuracy of 3D object detection; however, 3D Lidar has several disadvantages: expensive, easily affected by environmental conditions such as weather, and lidar data is sparse over long distances. 3D detection based on RGB camera can improve the robustness of the system, especially when other more expensive modules fail. Therefore, how to achieve reliable/accurate 3D object detection based on monocular/stereo camera is particularly important.

In comparison to the stereo object detection algorithm, which requires a lot of calculations and is difficult to register, 3D detection using a monocular camera is gradually becoming the main research focus.

2. Monocular 3D Object Detection

According to the method, monocular 3D object detection is roughly divided into four categories:

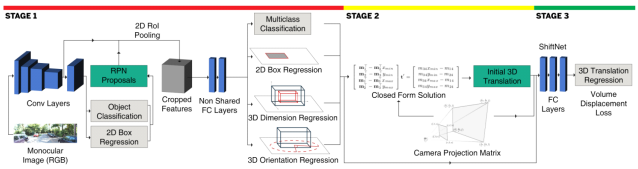
Based on existing 2D detection algorithms, an additional regression branch is added to the ROI of the target to predict 3D parameters. [1] first uses CNN to predict a reliable 2D bouding box and its orientation, then based on the predicted 2D information, use the relevant guidance information to get a 3D bounding box, called 3D guidance, and then refines the guidance to generate the final 3D bounding box. This methods[2,3] is difficult to achieve good results due to the large search space.



[Fig. 1] Overview of [1] proposed 3D object detection paradigm.

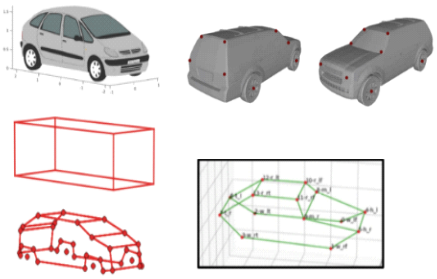
Use 2D bounding box and 3D object properties as supervision. Projection geometry is introduced as a constraint, and a deep regression network is used to predict 3D information, the 2D bounding box can only provide 4 constraints, necessitating extremely accurate bounding box prediction. [4] uses Faster R-CNN[5] to extract features to estimate the dimension of the 2D

bounding box and orientation, and then use the ordinary least-squares method to solve the 2D to 3D inverse geometric projection problem in the camera projection matrix to generate the position of the object. This method[6,7] uses two-stage detection to generate accurate 2D bounding boxes, is difficult to guarantee real-time performance. as show in figure 2.



[Fig. 2] Overview of [4] proposed model.

Select the 3D bounding box of the optimal proposed computational target in a multi-stage fusion module using 3D CAD models, 3D semantic segmentation, object contours, etc. as show in figure 3. But training these networks also requires additional annotations. Such methods[8,9,10] infer the vehicles' full shape from key-points, can improve the mAP of occluders and truncations, and use the CAD models to represent conventionally shaped vehicles. However, in order to train the network, needs to label additional relevant data, and even needs to provide the depth map to enhance the detection ability.



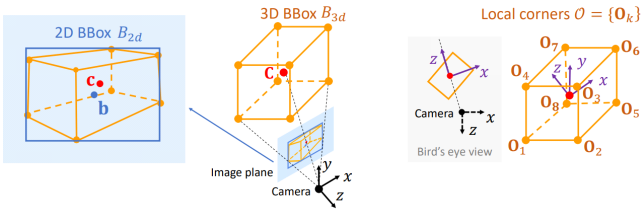
[Fig. 3] Different 3D CAD models, 3D semantic segmentation and object contours. From [9]

Use the key-point detection technology to extract and combine key-points for objects, and then use the camera's internal and external parameters to constrain the image's perspective. This methods of 3D object detection as a key-point detection task does not require predicting the 3D bounding box using existing 2D detectors or other data generators, but instead creates a network to predict the 8 corners and 1 center point of the 3D bounding box while minimizing the reprojection error to generate the best result.

3. Key-point based Monocular 3D Object Detection

One limitation of incorporating geometric information into deep learning methods is the vertices of the 3D box may be related to any edge of the 2D bounding box, but the 4 edges of the 2D box provide only 4 constraints for restoring the 3D box. Because of the over-reliance on the 2D box, even minor errors in the 2D box have a significant impact on the 3D box prediction. For occluded and truncated objects, the key-point method performs well.

MonoGRNet[10] proposed by Qin et al. can regression the projected 3D center point, the instance depth and the approximate 3D position, emphasizing the difference between the 2D bounding box and the 3D bounding box center projected to the 2D image, and the projected 3D center point can be regarded as an additional key-point, as show in figure 4.

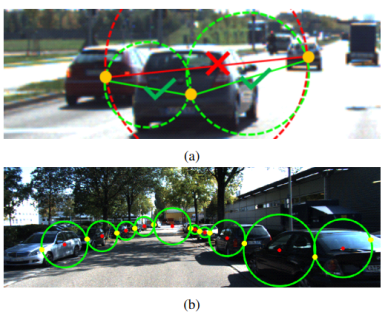


(a) 2D Projection (b) Camera coordinate frame (c) Local coordinate frame

[Fig. 4] Notation for 3D bounding box localization. From [10]

GPP(Ground Plane Polling)[11] Proposed by Akshay et al. Generate standard 2D KP points through 3D bounding box annotation, and the algorithm predicts some attribute values outside the demand to generate 3D bounding box, the purpose is to calculate the closest fit through all the generated predicted values. As a result, the algorithm is more robust in calculating the outer edge, similar to the RANSAC[12] algorithm.

MonoPair[13] Chen et al. proposed it. Get a lot of inspiration from CenterNet, and improve the final detection result through the spatial relationship between paired vehicles. Compared with CenterNet directly predicts a 3D bounding box, and also predicts virtual pairwise matching Constraint points between vehicles. Matching key-points are defined as: the respective center point pairs between the two nearest objects.



(a)

(b)

[Fig. 5] Pair matching strategy for training and inference. From [13]

SMOKE[14] Proposed by Liu et al. used the method of direct regression the 3D bounding box. The center point of the 3D cube generated by projection is used as the encoding method for the 3D bounding box, with other parameters(dimension, depth, orientation) as a supplementary parameters.

FCOS3D[15] proposed by Wang et al. Based on FCOS[16] anchor-free 2D detection method to improve 3D detection, the backbone is ResNet101[17] with DCN, and is equipped with FPN architecture to detect objects of different scales. Same with SMOKE, in order to predict the 3D bounding box, the following variables are regressed: the offset of the projected point of the 3D center point on the 2D plane, the depth, three-dimensional size, heading angle, and the direction discrimination.

Cai et al. proposed Decoupled-3D[18]. It was first proposed that the depth estimation of 3D vertices can be separated from 2D projection estimation. To generate the projected cube vertices, a scheme similar to RTM3D[19] is used, and the virtual edge height is used as a strong prior to getting the depth estimate, allowing the 3D detection frame to be generated.

4. Conclusions

This paper gives an overview of recent advances in monocular 3D object detection using key points for autonomous driving. First, we show how to categorize existing monocular 3D object detection methods, then analyzed and compared the key-point based methods, and discussed each method for key-point 3D detection, such as virtual key-point, key-point dropout module, relationship key-point, etc. Analyzing and comparing the key-point based monocular 3D object detection under different methods, provides a research basis for our subsequent research on key-point based monocular 3D object detection.

Acknowlegements

This research was supported by the Ministry of Trade, Industry & Energy(MOTIE), Korea Institute for Advancement of Technology(KIAT) through the OEM demonstration cluster construction project to support autonomous vehicle parts suppliers.[Project Number: P0014572]

References

[1] Li, Buyu, Ouyang, Wanli, Sheng, Lu, Zeng, Xingyu, Wang, Xiaogang. (2019). GS3D: An Efficient 3D Object Detection Framework for Autonomous Driving. 1019-1028. 10.1109/CVPR.2019.00111.

[2] Shi, Xuepeng, Ye, Qi, Chen, Xiaozhi, Chen, Chuangrong, Chen, Zhixiang, Kim, Tae-Kyun. (2021). Geometry-based Distance Decomposition for Monocular 3D Object Detection. 15152-15161. 10.1109/ICCV48922.2021.01489.

[3] Liu, Lijie, Lu, Jiwen, Xu, Chunjing, Tian, Qi, Zhou, Jie. (2019). Deep Fitting Degree Scoring Network for Monocular 3D Object Detection. 1057-1066. 10.1109/CVPR.2019.00115.

[4] Naiden, Andretti, Paunescu, Vlad, Kim, Gyeongmo, Jeon, ByeongMoon, Leordeanu, Marius. (2019). Shift R-CNN: Deep Monocular 3D Object Detection With Closed-Form Geometric Constraints. 61-65. 10.1109/ICIP.2019.8803397.

[5] Ren, Shaoqing, He, Kaiming, Girshick, Ross, Sun, Jian. (2015). Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. IEEE Transactions on Pattern Analysis and Machine Intelligence. 39. 10.1109/TPAMI.2016.2577031.

[6] Lu, Yan, Ma, Xinzhu, Yang, Lei, Zhang, Tianzhu, Liu, Yating, Chu, Qi, Yan, Junjie, Ouyang, Wanli. (2021). Geometry Uncertainty Projection Network for Monocular 3D Object Detection. 3091-3101. 10.1109/ICCV48922.2021.00310.

[7] Choi, Hee, Kang, Hyoa, Hyun, Yoonsuk. (2019). Multi-View Reprojection Architecture for Orientation Estimation. 2357-2366. 10.1109/ICCVW.2019.00289.

[8] Kundu, Abhijit, Li, Yin, Rehg, James. (2018). 3D-RCNN: Instance-Level 3D Object Reconstruction via Render-and-Compare. 3559-3568. 10.1109/CVPR.2018.00375.

[9] Barabanau, Ivan, Artemov, Alexey, Burnaev, Evgeny, Murashkin, Vyacheslav. (2020). Monocular 3D Object Detection via Geometric Reasoning on Keypoints. 652-659. 10.5220/0009102506520659.

[10] Qin, Zengyi, Wang, Jinglu, Lu, Yan. (2019). MonoGRNet: A Geometric Reasoning Network for Monocular 3D Object Localization. Proceedings of the AAAI Conference on Artificial Intelligence. 33. 8851-8858. 10.1609/aaai.v33i01.33018851.

[11] Rangesh, Akshay, Trivedi, Mohan. (2020). Ground Plane Polling for 6DoF Pose Estimation of Objects on the Road. IEEE Transactions on Intelligent Vehicles. PP. 1-1. 10.1109/TIV.2020.2966074.

[12] Fischler, M.A. and Bolles, R.C. Random Sample Consensus: A Paradigm for Model Fitting with Applications to Image Analysis and Automated Cartography. Communications of the ACM , 24(6): 381‐395, 1981.

[13] Chen, Yongjian, Tai, Lei, Sun, Kai, Li, Mingyang. (2020). MonoPair: Monocular 3D Object Detection Using Pairwise Spatial Relationships. 12090-12099. 10.1109/CVPR42600.2020.01211.

[14] Liu, Zechen, Wu, Zizhang, Tóth, Roland. (2020). SMOKE: Single-Stage Monocular 3D Object Detection via Keypoint

Estimation. 4289-4298. 10.1109/CVPRW50498.2020.00506.

[15] Wang, Tai, Zhu, Xinge, Pang, Jiangmiao, Lin, Dahua. (2021). FCOS3D: Fully Convolutional One-Stage Monocular 3D Object Detection. 913-922. 10.1109/ICCVW54120.2021.00107.

[16] Tian, Zhi, Shen, Chunhua, Chen, Hao, Tong, He. (2019). FCOS: Fully Convolutional One-Stage Object Detection. 9626-9635. 10.1109/ICCV.2019.00972.

[17] He, Kaiming, Zhang, Xiangyu, Ren, Shaoqing, Sun, Jian. (2016). Deep Residual Learning for Image Recognition. 770-778. 10.1109/CVPR.2016.90.

[18] Cai, Yingjie & Li, Buyu & Jiao, Zeyu & Li, Hongsheng & Zeng, Xingyu & Wang, Xiaogang. (2020). Monocular 3D Object Detection with Decoupled Structured Polygon Estimation and Height-Guided Depth Estimation. Proceedings of the AAAI Conference on Artificial Intelligence. 34. 10478-10485. 10.1609/aaai.v34i07.6618.

[19] Li, Peixuan, Zhao, Huaici, Liu, Pengfei, Cao, Feidao. (2020). RTM3D: Real-Time Monocular 3D Detection from Object Keypoints for Autonomous Driving. 10.1007/978-3-030-58580-8_38.