Camera-LiDAR-Based 3D Object Detection Methods

Subjects: Computer Science, Artificial Intelligence

Contributor: Pascal Housam Salmane , Josué Manuel Rivera Velázquez , Louahdi Khoudour , Nguyen Anh Minh Mai , Pierre Duthon , Alain Crouzil , Guillaume Saint Pierre , Sergio A. Velastin

Three-dimensional (3D) object detection is a topic that has gained interest within the scientific community dedicated to vehicle automation. Based on LiDAR and stereo cameras, and considering only deep learning-based approaches, 3D object detection methods are classified according to the type of input data: camera-based, LiDAR-based, and fusion-based 3D object detection.

autonomous vehicle 3D object detection LiDAR fusion stereo camera

1. Introduction

Object detection is one of the main components of computer vision aimed at detecting and classifying objects in digital images. Although there is great interest in the subject of 2D object detection, the scope of detection tools has increased with the introduction of 3D object detection, which has become an extremely popular topic, especially for autonomous driving. In this case, 3D object detection is more relevant than 2D object detection since it provides more spatial information: location, direction, and size.

For each object of interest in an image, a 3D object detector produces a 3D bounding box with its corresponding class label. A 3D bounding box can be encoded as a set of seven parameters [1]: (x,y,z,h,w,l,θ), including the coordinates of the object center (x,y,z), the size of the object (height, width, and length), and its heading angle (θ). At the hardware level, the technology involved in the object detection process mainly includes the use of mono and stereo cameras, with visible light or infrared cameras, RADAR (radio detection and ranging), and LiDAR (light detection and ranging), and gated cameras. In fact, the current top-performing methods in 3D object detection are based on the use of LiDAR (**Figure 1**) [2][3].

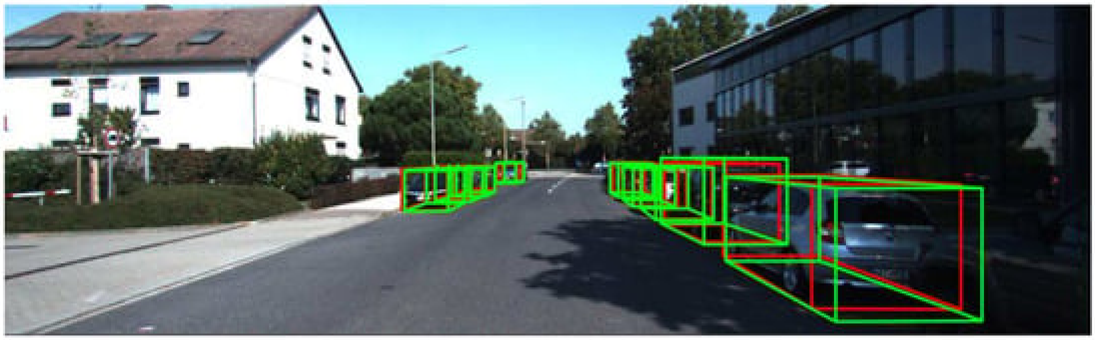

Figure 1. Example of a road scene where detection results using LiDAR technology is used. Detected objects are surrounded by bounding boxes. The green boxes represent detection while the red ones represent ground truth.

However, highly accurate LiDAR sensors are extremely costly (the price of a 64-beam model is around USD 75,000 [4]), which incurs a hefty premium for autonomous driving hardware. Alternatively, systems based only on camera sensors have also received much attention because of their low costs and wide range of use. For example, in [5], the authors claim that instead of using expensive LiDAR sensors for accurate depth information, the alternative is to use pseudo-LiDAR, which has been introduced as a promising alternative at a much lower cost based solely on stereo images. The paper presents the advances to the pseudo-LiDAR framework through improvements in stereo depth estimation. Similarly, in [6], instead of using a LiDAR sensor, the authors provide a simple and effective one-stage stereo-based 3D detection pipeline that jointly estimates depth and detects 3D objects in an end-to-end learning manner. These authors claim that this method outperforms previous stereo-based 3D detectors and even achieves comparable performance to a few LiDAR-based methods on the KITTI 3D object detection leaderboard. Another example is presented in [7]. To tackle the problem of high variance in depth estimation accuracy with a video sensor, the authors propose CG-Stereo, a confidence-guided stereo 3D object detection pipeline that uses separate decoders for foreground and background pixels during depth estimation, and leverages the confidence estimation from the depth estimation network as a soft attention mechanism in the 3D object detector. The authors say that their approach outperforms all state-of-the-art stereo-based 3D detectors on the KITTI benchmark.

Another interesting solution presented in the literature is the combination of LiDAR and a stereo camera. These methods exploit the fact that LiDAR will complete the vision and information provided by the camera, by adding notions of size and distance to the different objects that make up the environment. For example, the proposed method in [8] takes advantage of the fact that it is possible to reconstruct a 3D environment using images from stereo cameras, making it possible to extract a depth map from stereo camera information and enrich it with the data provided by the LiDAR sensor (height, width, length, and heading angle).

2. Camera-LiDAR-Based 3D Object Detection Methods

2.1. Camera-Based Methods

Some of the first algorithms were keypoint/shape-based methods. In 2017, Chabot et al. [9] presented Deep MANTA, one of the first works on camera-based 3D object detection. This architecture recognizes 2D keypoints and uses 2D-to-3D matching. This algorithm has two steps: (1) a RCNN architecture to detect and refine 2D bounding boxes and keypoints, and (2) a predicted template similarity to pick the best matching 3D model inside a 3D dataset. However, the main disadvantage of this method is the excessive time required for 2D/3D matching.

On the other hand, the pseudo-point cloud-based methods arose from the idea of simulating LiDAR from a stereo camera. These methods typically convert data from 2D to 3D by using extrinsic calibration information between the camera and LiDAR. For example, in Xu and Chen [10], a depth map from RGB images was predicted and then concatenated as RGB-D (where D is the datum provided by the depth map) to form a tensor of six channels (RGB image, Z-depth, height, distance) used to regress the 3D bounding boxes. For example, the Pseudo-LiDAR method presented by Wang et al. [8] was inspired by that article, based on the idea of data representation transformation to estimate the depth map from an RGB image using a depth estimation neural network. Next, the predicted depth map is converted into a pseudo-3D point cloud by projecting all pixels with depth information into LiDAR coordinates. Then, the pseudo-point cloud was ready to be used as input in any LiDAR-based object detection method. The ability to reconstruct a 3D point cloud from less expensive monocular or stereo cameras is a valuable feature of this approach. Based on Pseudo-LiDAR [8], the same authors proposed Pseudo-LiDAR++ [5] through a stereo depth estimation neural network (SDN). They proposed a depth cost volume to directly predict the depth map instead of predicting the disparity map, as proposed by Chang and Chen [11]. This boosts the predicted depth map accuracy. In addition, to improve the accuracy of the predicted depth map, they proposed a depth correction phase using a simulated four-beam LiDAR to regularise the predicted depth map.

2.2. LiDAR-Based Methods

Due to the irregular, unstructured, and unordered nature of point clouds [12], they are often handled in one of three ways: projecting point clouds to generate a regular pseudo-image, sub-sampling point cloud cells called voxels, or encoding raw point clouds with a sequence of multi-layer perceptron (MLP), as proposed in [13]. For 3D object detection, LiDAR-based methods are usually classified into four categories: view-based, voxel-based, point-based, and hybrid point-voxel-based detection.

View-based detection methods, such as the one presented by Beltran et al. [14], project a point cloud onto a 2D image space to obtain a regular structure as an initial stage. Generally, a CNN (convolutional neural network) is then used to take advantage of this information [15][16]. The most common types of projection are bird’s eye view (BEV), front view (FV) [17], range view (RV) [18], and spherical view (SV) [19].

The voxel-based method maps a point cloud into 3D grids (voxels) as an initial stage. In 2017, Engelcke et al. [20] presented their LiDAR-based detector Vote3Deep. In this method, LiDAR point clouds are discretized into a sparse 3D grid. Then, items are detected using a sliding-window search with a fixed-size window with *N* different orientations. In each window, a CNN performs binary classification.

Point-based methods usually deal with the raw point cloud directly instead of converting the point cloud to a regular structure. Qi et al. [13] introduced PointNet, a pioneering study on deep learning-based architectures for processing point cloud raw data for classification and semantic segmentation. They argue that as the point cloud is unordered, the architecture should be permutation-invariant for all points.

2.3. Fusion-Based Methods

Image-based object detection is an advanced research area. In addition, cameras are cheap and provide a lot of texture information about objects based on color and edges. However, images lack depth information, which is extremely important for 3D tasks. Even with a stereo camera, the resulting depth map lacks accuracy. On the other hand, although LiDAR does not give texture information, LiDAR-based methods show a very high performance compared to camera-based methods. However, there are still limitations (such as obscure information) regarding object categories. For example, in some cases, it is difficult to distinguish whether it is a car or a bush based on point cloud data alone, while this can be handled more easily by looking at the image data. This is why methods based on data fusion have been developed exploiting the advantages of both sensors. In the literature, there are three main fusion methods: early fusion, where the raw data are fused at the data level or feature level to form a tensor data of numerous channels; late fusion, where the fusion takes place at the decision level; and deep fusion, where fusion is carefully constructed to combine the advantages of both early and late fusion systems.

For example, Qi et al. [1] presented the Frustum-PointNet architecture, which is composed of three phases: 3D frustum proposal, 3D instance segmentation, and 3D bounding box estimation. The first phase of this procedure is to produce 2D region proposals. By extruding the matching 2D region proposal under a 3D projection, a 3D frustum proposal is generated. The instance segmentation stage feeds the frustum proposal point cloud to the PointNet segmentation network [13], which classifies each point and determines if it is linked with the discovered item. In the last stage, all positively classified points are loaded into a new PointNet that estimates 3D bounding box parameters.

Chen et al. [17] introduced MV3D, where the LiDAR point cloud is projected onto both a 2D top view and a 2D front view, from which feature maps are extracted using two separate CNN. The LiDAR top-view feature map is passed to an RPN (Region Proposal Network) to output proposal 3D bounding boxes. Each 3D proposal is projected onto the feature maps of all three views and a fixed-size feature vector is extracted for each view using pooling. The three feature vectors are then fused in a region-based fusion network, which finally outputs class scores and regresses 3D bounding box residuals.

A similar approach, also utilizing the PointNet architecture, was independently presented in Xu et al. [21]. Just as in Frustum-PointNet, a 2D object detector is used to extract 2D region proposals (ResNet), which are extruded to the corresponding frustum point cloud. Each frustum is fed to a PointNet, extracting both point-wise feature vectors and a global LiDAR feature vector. Each 2D image region is also fed to a CNN that extracts an image feature vector. For each point in the frustum, its point-wise feature vector is concatenated with both the global LiDAR feature vector and the image feature vector. This concatenated vector is finally fed to a shared MLP, outputting 8 × 3 values for each point. The output corresponds to predicted $(x,y,z$

$)$ offsets relative to the point for each of the eight 3D bounding box corners. The points in the frustum are thus used as dense spatial anchors. The MLP also outputs a confidence score for each point, and in inference, the bounding box corresponding to the highest-scoring point is chosen as the final prediction.

Ku et al. [22] introduced another fusion architecture named AVOD. Here, the LiDAR point cloud is projected onto a 2D top-view, from which a feature map is extracted by a CNN. A second CNN is used to extract a feature map also from the input image. The two feature maps are shared by two subnetworks: an RPN and a second-stage detection network. The reported 3D detection performance is a slight improvement compared to [17]; it is comparable to that of [23] for cars but somewhat lower for pedestrians and cyclists. The authors also found that using both image and LiDAR features in the RPN, as compared to only using LiDAR features, has virtually no effect on the performance of cars but a significant positive effect for pedestrians and cyclists.

Similar to some of the methods mentioned above, the SLS–Fusion method presented in Mai et al. [24] resulted from this idea. Roughly, the SLS–Fusion method estimates the depth maps from a stereo camera and the projected LiDAR depth maps. However, as Zhu et al. [25] point out, this produces a mismatch between the resolution of point clouds and RGB images. Specifically, taking the sparse points as the multi-modal data aggregation locations causes severe information loss for high-resolution images, which in turn undermines the effectiveness of multi-sensor fusion.

More research is needed on the limits of a 3D object detection model composed of a LiDAR and a stereo camera. Knowing the role of each sensor will allow for the optimization and configuration of the proposed methods.

References

1. Qi, C.R.; Liu, W.; Wu, C.; Su, H.; Guibas, L.J. Frustum PointNets for 3D Object Detection from RGB-D Data. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–23 June 2018; pp. 918–927.
2. Shi, S.; Guo, C.; Jiang, L.; Wang, Z.; Shi, J.; Wang, X.; Li, H. PV-RCNN: Point-Voxel Feature Set Abstraction for 3D Object Detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 13–19 June 2020; pp. 10526–10535.
3. He, C.; Zeng, H.; Huang, J.; Hua, X.S.; Zhang, L. Structure Aware Single-Stage 3D Object Detection From Point Cloud. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 13–19 June 2020; pp. 11870–11879.
4. Velodyne’s HDL-64E Lidar Sensor Looks Back on a Legendary Career. Available online: https://velodynelidar.com/blog/hdl-64e-lidar-sensor-retires/ (accessed on 20 February 2022).
5. You, Y.; Wang, Y.; Chao, W.L.; Garg, D.; Pleiss, G.; Hariharan, B.; Campbell, M.; Weinberger, K.Q. Pseudo-LiDAR++: Accurate Depth for 3D Object Detection in Autonomous Driving. In Proceedings of the International Conference on Learning Representations (ICLR), Virtual Conference, 26 April–1 May 2020.
6. Chen, Y.; Liu, S.; Shen, X.; Jia, J. DSGN: Deep Stereo Geometry Network for 3D Object Detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 13–19 June 2020; pp. 12533–12542.
7. Li, C.; Ku, J.; Waslander, S.L. Confidence Guided Stereo 3D Object Detection with Split Depth Estimation. In Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and

Systems (IROS), Las Vegas, NV, USA, 24 October 2020–24 January 2021; pp. 5776–5783.

8. Wang, Y.; Chao, W.L.; Garg, D.; Hariharan, B.; Campbell, M.; Weinberger, K.Q. Pseudo-LiDAR From Visual Depth Estimation: Bridging the Gap in 3D Object Detection for Autonomous Driving. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019; pp. 8437–8445.
9. Chabot, F.; Chaouch, M.; Rabarisoa, J.; Teuliere, C.; Chateau, T. Deep MANTA: A Coarse-to-Fine Many-Task Network for Joint 2D and 3D Vehicle Analysis from Monocular Image. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 1827–1836.
10. Xu, B.; Chen, Z. Multi-level Fusion Based 3D Object Detection from Monocular Images. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–23 June 2018; pp. 2345–2353.
11. Chang, J.R.; Chen, Y.S. Pyramid Stereo Matching Network. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–23 June 2018; pp. 5410–5418.
12. Bello, S.A.; Yu, S.; Wang, C.; Adam, J.M.; Li, J. Review: Deep Learning on 3D Point Clouds. Remote Sens. 2020, 12, 1729.
13. Qi, C.R.; Su, H.; Mo, K.; Guibas, L.J. PointNet: Deep Learning on Point Sets for 3D Classification and Segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 77–85.
14. Beltran, J.; Guindel, C.; Moreno, F.M.; Cruzado, D.; Garcia, F.; De La Escalera, A. BirdNet: A 3D Object Detection Framework from LiDAR Information. In Proceedings of the IEEE International Conference Intelligent Transportation Systems (ITSC), Maui, HI, USA, 4–7 November 2018; pp. 3517–3523.
15. Liu, T.; Yang, B.; Liu, H.; Ju, J.; Tang, J.; Subramanian, S.; Zhang, Z. GMDL: Toward precise head pose estimation via Gaussian mixed distribution learning for students' attention understanding. Infrared Phys. Technol. 2022, 122, 104099.
16. Liu, T.; Wang, J.; Yang, B.; Wang, X. NGDNet: Nonuniform Gaussian-label distribution learning for infrared head pose estimation and on-task behavior understanding in the classroom. Neurocomputing 2021, 436, 210–220.
17. Chen, X.; Ma, H.; Wan, J.; Li, B.; Xia, T. Multi-view 3D Object Detection Network for Autonomous Driving. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 6526–6534.
18. Meyer, G.P.; Laddha, A.; Kee, E.; Vallespi-Gonzalez, C.; Wellington, C.K. LaserNet: An Efficient Probabilistic 3D Object Detector for Autonomous Driving. In Proceedings of the IEEE/CVF

Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019; pp. 12669–12678.

19. Gigli, L.; Kiran, B.R.; Paul, T.; Serna, A.; Vemuri, N.; Marcotegui, B.; Velasco-Forero, S. Road segmentation on low resolution LiDAR point clouds for autonomous vehicles. arXiv 2020, arXiv:2005.13102.
20. Engelcke, M.; Rao, D.; Wang, D.Z.; Tong, C.H.; Posner, I. Vote3Deep: Fast object detection in 3D point clouds using efficient convolutional neural networks. In Proceedings of the IEEE International Conference on Robotics and Automation (ICRA), Singapore, 29 May–3 June 2017; pp. 1355–1361.
21. Xu, D.; Anguelov, D.; Jain, A. PointFusion: Deep Sensor Fusion for 3D Bounding Box Estimation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–23 June 2018; pp. 244–253.
22. Ku, J.; Mozifian, M.; Lee, J.; Harakeh, A.; Waslander, S.L. Joint 3D Proposal Generation and Object Detection from View Aggregation. In Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Madrid, Spain, 1–5 October 2018; pp. 1–8.
23. Zhou, Y.; Tuzel, O. VoxelNet: End-to-End Learning for Point Cloud Based 3D Object Detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–23 June 2018; pp. 4490–4499.
24. Mai, N.A.M.; Duthon, P.; Khoudour, L.; Crouzil, A.; Velastin, S.A. Sparse LiDAR and Stereo Fusion (SLS-Fusion) for Depth Estimation and 3D Object Detection. In Proceedings of the the International Conference of Pattern Recognition Systems (ICPRS), Curico, Chile, 17–19 March 2021; Volume 2021, pp. 150–156.
25. Zhu, H.; Deng, J.; Zhang, Y.; Ji, J.; Mao, Q.; Li, H.; Zhang, Y. VPFNet: Improving 3D Object Detection with Virtual Point based LiDAR and Stereo Data Fusion. IEEE Trans. Multimedia 2022, 1–14.