Literature Review for Object Detection in Autonomous Driving

AIZIERJIANG AIERSILAN
MC251013
FST, UM

ABSTRACT

As 3D object detection serves as the core basis of perception stack in autonomous driving especially for the sake of path planning, motion prediction, and collision avoidance etc., there has been lots of technics regarding 3D object detection. This paper summarizes and reviews the state-of-the-art methods and technics by making a simple literature review of more than 10 papers concerning object detection published in recent two years to provide an overall view of the latest trend of object detection in autonomous driving.

INDEX TERMS

Object detection, Autonomous driving, dataset

In autonomous driving, challenging tasks always occur in some corner cases, making the corner cases one of the main research areas to be studied with. To provide better results in safety-critical systems, researchers have come up with various approaches. Contemporary deep-learning object detection methods for autonomous driving usually presume fixed categories of common traffic participants, such as pedestrians and cars. Most existing detectors are unable to detect uncommon objects and corner cases, which may lead to severe accidents in some situations, making the timeline for the real-world application of reliable autonomous driving uncertain. Most of the previous studies on object detection for autonomous driving had been conducted under clear and low-noise conditions and few of them considered difficult situations, simply in terms of weather or illumination differentiation. However, the actual autonomous driving environment has more obstacles than weather and illumination. For example, noises from the camera, motion blur based on driving state, object characteristics in the acquired images and so on. These obstacles are mainly put from corner cases that happens rarely but dangerously. In a paper titled *Anomaly Detection in Autonomous Driving: A Survey*[1], researchers surveyed the area of anomaly detection in autonomous driving and compared multiple modern technics concerning 3D object detection. Besides, the pros and cons of anomaly detection on camera data, LiDAR data, radar data, abstract object data as well as multimodal data are listed through the comparison with lots of state-of-the-art research achievements in autonomous driving to find some solutions for optimizing the object detection especially in corner cases. As is known that the perception of autonomous vehicles performs well under closed-set conditions, they still struggle to handle the unexpected. The survey

[1] Bogdoll, Daniel, Maximilian Nitsche, and J. Marius Zöllner. "Anomaly Detection in Autonomous Driving: A Survey." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022.

provides an extensive overview of anomaly detection techniques based on camera, LiDAR, radar, multimodal and abstract object level data. They provide a systematization including detection approach, corner case level, ability for an online application, and further attributes.

Going deep into one of the specific technics regarding the object detection in some corner cases, the research team of a paper titled *CODA: A Real-World Road Corner Case Dataset for Object Detection in Autonomous Driving*[2] contributes one of the reasons that impedes the development of truly reliably self-driving systems to the lack of public datasets for evaluating the performance of object detectors on corner cases, hence proposing a challenging dataset named CODA that exposes this critical problem of vision-based detectors. The open-source dataset named CODA consists of 1500 carefully selected real-world driving scenes, each containing four object-level corner cases, spanning more than 30 object categories. According to the paper, the construction of CODA is carried out in two main stages. The first stage is an automatic generation of proposals that identifies potential corner cases from initial data, followed by the second stage, a manual selection and labeling process that eliminates the false positives of the proposals, and then classifies the remaining true positives while adjusting their bounding boxes to be more precise.

Though the dataset like CODA which is based on deep learning are robust for detecting objects in some corner cases, the deep learning-based vision systems that utilizes those datasets are vulnerable to perturbation, which contains noise. Thus, robust object detection under harsh autonomous-driving environments is a more difficult than the generic situation. As is pointed in a paper titled *Robust object detection under harsh autonomous-driving environments*[3], not only the accuracy, but also the speed of the non-maximum suppression-based detector can be degraded under harsh environments. Therefore, object detection is handled under a harsh situation with adversarial mechanisms such as adversarial training and adversarial defense. Adversarial defense modules are designed to improve robustness in feature extraction level and define perturbations under a harsh environment for training object detectors to improve the robustness of the model's decision boundary. The researchers of the paper titled *Robust object detection under harsh autonomous-driving environments* proposed adversarial defense and training mechanisms which can improve the object detector in both accuracy and speed. They mentioned that the proposed method shows a 43.7% mean average precision for the COCO2015 dataset in generic object detection and 39.0% mean average precision for the BDD100K dataset in a driving environment. Furthermore, it is said to be able to achieve a real-time capability of 23 frames per second.

Since most of the algorithms heavily rely on data annotation, which is a laborious job, the researchers of the paper titled *A semi-supervised 3D object detection method for*

---

[2] Li, Kaican, et al. "CODA: A Real-World Road Corner Case Dataset for Object Detection in Autonomous Driving." arXiv preprint arXiv:2203.07724 (2022).

[3] Kim, Youngjun, Hyekyoung Hwang, and Jitae Shin. "Robust object detection under harsh autonomous-driving environments." IET Image Processing 16.4 (2022): 958-971.

autonomous driving[4] proposed a semi-supervised 3D object detection method. To reduce the workload of 3D annotations, the proposed method adopted the teacher-student framework to generate pseudo-labels from unlabeled training data, and use a label filtering method to improve the pseudo label quality and it is also validated on the KITTI dataset. For moderate Car detection task, the method is said to be able to achieve 76.28 mAP using half labels compared with 77.34 mAP of the PointPillars using all labels.

However, supervised object detection models based on deep learning technologies cannot perform well in domain shift scenarios where annotated data for training is always insufficient. To this end, domain adaptation technologies for knowledge transfer have emerged to handle the domain shift problems. In a paper titled *Cross-Domain Object Detection for Autonomous Driving: A Stepwise Domain Adaptative YOLO Approach*[5], a stepwise domain adaptive YOLO[6] (S-DAYOLO) framework is developed which constructs an auxiliary domain to bridge the domain gap and uses a new domain adaptive YOLO(DAYOLO) in cross-domain object detection tasks. Different from the previous solutions, the auxiliary domain is composed of original source images and synthetic images that are translated from source images to the similar ones in the target domain. DAYOLO based on YOLOv5s is designed with a category-consistent regularization module and adaptation modules for image-level and instance-level features to generate domain invariant representations. The proposed method is trained and evaluated by using five public driving datasets including Cityscapes, Foggy Cityscapes, BDD100K, KITTI, and KAIST and the experiment results demonstrated that object detection performance is significantly improved when using our proposed method in various domain shift scenarios for autonomous driving applications.

YOLO architecture can also be adopted into low-latency multispectral pedestrian detection and its latest version --- YOLOv4 --- is thoroughly investigated by the researchers of the paper titled *Adopting the YOLOv4 architecture for low-latency multispectral pedestrian detection in autonomous driving*[7]. It is demonstrated that this detector can be adapted to multispectral pedestrian detection. It can achieve accuracy on par with the state-of-the-art while being highly computationally efficient, thereby supporting low-latency decision making.

When it comes to multimodal 3D object detection, the researchers of the paper titled *Spatio-contextual deep network-based multimodal pedestrian detection for autonomous driving*[8] propose an end-to-end multimodal fusion model for

---

[4] Zhang, Jiacheng, Huafeng Liu, and Jianfeng Lu. "A semi-supervised 3D object detection method for autonomous driving." Displays 71 (2022): 102117.

[5] Li, Guofa, et al. "Cross-Domain Object Detection for Autonomous Driving: A Stepwise Domain Adaptative YOLO Approach." IEEE Transactions on Intelligent Vehicles (2022).

[6] A real-time neural network detector architecture: You Only Look Once.

[7] Roszyk, Kamil, Michał R. Nowicki, and Piotr Skrzypczyński. "Adopting the YOLOv4 architecture for low-latency multispectral pedestrian detection in autonomous driving." Sensors 22.3 (2022): 1082.

[8] Dasgupta, Kinjal, et al. "Spatio-contextual deep network-based multimodal pedestrian detection for autonomous

pedestrian detection using RGB and thermal images. The model's novel spatio-contextual deep network architecture is capable of exploiting the multimodal input efficiently and it consists of two distinct deformable ResNeXt-50 encoders for feature extraction from the two modalities. Fusion of these two encoded features takes place inside a multimodal feature embedding module consisting of several groups of a pair of Graph Attention Network and a feature fusion unit. The output of the last feature fusion unit of multimodal feature embedding module is subsequently passed to two CRFs for their spatial refinement. They achieved further enhancement of the features by applying channel-wise attention and extraction of contextual information with the help of four RNNs traversing in four different directions. These feature maps are used by a single-stage decoder to generate the bounding box of each pedestrian and the score map. The proposed framework has been tested on three publicly available multimodal pedestrian detection benchmark datasets, namely KAIST, CVC-14, and UTokyo, and the results on each of them improved the respective state-of-the-art performance. The main contributions of this paper are summarized as follows:

(1) Design of a novel multimodal feature embedding module using graph attention network and feature fusion unit to address the modality imbalance problem.
(2) Design of a spatio-contextual feature aggregation module to improve fusion using CRF-based refinement, channel-wise attention, and 4Dir-IRNN components.
(3) Implementation of an end-to-end trainable network achieving state-of-the-art results on three public datasets, namely KAIST, CVC-14, and UTokyo.
(4) Extensive experimentation of different feature encoders, network components, various augmentation strategies, curriculum learning, and tuning of hyper-parameters.

As a detection model only using the sensing data of a single sensor cannot obtain accurate detection results in a complex environment, existing efforts are divided into the following three subdivisions:

(1) Image based, which is relatively inaccurate but several orders of magnitude cheaper, and more interpretable under the guidance of domain expertise and knowledge priors.
(2) Point cloud based, which has a relatively higher accuracy and lower latency but more prohibitive deployment cost compared with its image based counterparts.
(3) Multimodal fusion based, which currently lags behind its point cloud based counterparts but importantly provides a redundancy to fall back onto in case of a malfunction or outage.

In an investigation report called *Investigating the Impact of Multi-LiDAR Placement on Object Detection for Autonomous Driving*[9], while most of the existing works focus on developing new deep learning algorithms or model

driving." IEEE Transactions on Intelligent Transportation Systems (2022).

[9] Hu, Hanjiang, et al. "Investigating the Impact of Multi-LiDAR Placement on Object Detection for Autonomous Driving." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022.

architectures, the report focused on the problem from the physical design perspective, i.e., how different placements of multiple Li-DARs influence the learning-based perception. An easy-to-compute information-theoretic surrogate metric to quantitatively and quickly evaluate LiDAR placement for 3D detection of different types of objects is introduced and a new data collection, detection model training and evaluation framework in the realistic CARLA simulator to evaluate disparate multi-LiDAR configurations are presented.

The researchers of the paper titled *Multi-scale multi-modal fusion for object detection in autonomous driving based on selective kernel*[10] propose a multi-scale selective kernel fusion method and demonstrate its practical utility by using LiDAR-camera fusion in object detection network. Specifically, a multi-scale feature fusion module that uses multi-scale convolution to separate the feature expression of multi-modal information and calculates the weight of each modal feature channel is proposed. They used the idea of multi-scale convolution and selection kernel to complete multi-modal fusion in object detection, which is conducive to solving the problem that the image and point cloud fusion are difficult to match due to the difference in data structure, and the complementarity of multi-modal information has been fully utilized. Based on abundant experiments, they pointed that the proposed method introduces a new optimization idea for multi-modal fusion in the field of autonomous driving object detection, and the fusion detection efficiency is at over 12 fps on a single GPU. Besides the LiDAR-camera fusion in object detection network, millimeter wave radar and vision fusion has also become a mainstream solution for accurate obstacle detection. In another paper titled *MmWave Radar and Vision Fusion for Object Detection in Autonomous Driving: A Review*[11] presents a detailed survey on millimeter wave radar and vision fusion-based obstacle detection methods. In the paper, the process of millimeter wave radar and vision fusion is divided into three parts: sensor deployment, sensor calibration, and sensor fusion, which are reviewed comprehensively in the paper. Specifically, they classified the fusion methods into data level, decision level, and feature level fusion methods.

Compared with LiDAR system and millimeter wave radar, monocular cameras are cheap, stable, and flexible, favored by mass-produced cars. However, monocular 3D object detection is a natural ill-posed problem for the lack of depth information, making it difficult to estimate an accurate and stable state of the 3D target. A typical solution is to smooth the previous and current state through 2D multiple object trackers. In a paper titled *Time3D: End-to-End Joint Monocular 3D Object Detection and Tracking for Autonomous Driving*[12], jointly training 3D detection and 3D tracking from only monocular videos in an

[10] Gao, Xin, Guoying Zhang, and Yijin Xiong. "Multi-scale multi-modal fusion for object detection in autonomous driving based on selective kernel." Measurement 194 (2022): 111001.

[11] Wei, Zhiqing, et al. "MmWave Radar and Vision Fusion for Object Detection in Autonomous Driving: A Review." Sensors 22.7 (2022): 2542.

[12] Li, Peixuan, and Jieyu Jin. "Time3D: End-to-End Joint Monocular 3D Object Detection and Tracking for Autonomous Driving." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022.

end-to-end manner is proposed. In this method, the key component is a novel spatial-temporal information flow module that aggregates geometric and appearance features to predict robust similarity scores across all objects in current and past frames. Specifically, the proposed method aims to leverage the attention mechanism of the transformer, in which self-attention aggregates the spatial information in a specific frame, and cross-attention exploits relation and affinities of all objects in the temporal domain of sequence frames. The affinities are then supervised to estimate the trajectory and guide the flow of information between corresponding 3D objects. Besides, the stated temporal-consistency loss that explicitly involves 3D target motion modeling into the learning, making the 3D trajectory smooth in the world coordinate system.

Another end-to-end method which is similar to the one stated above in general framework but totally different in detail is proposed in another paper titled *End-to-end deep learning-based autonomous driving control for high-speed environment* [13]. Inspired by the UNet [14] architecture of semantic image segmentation, this paper presents a lightweight UNet using depth-wise separable convolutions for end-to-end learning of lane detection and path prediction in autonomous driving.

Despite existing efforts, 3D object detection for autonomous driving is still regarded to be staying in its infancy currently. Starting from grouping literature based on network architecture derives from 2D object detection, the paper called *3D Object Detection for Autonomous Driving* [15] introduces background associated with foundations, sensors, datasets, and performance metrics, reviews 3D object detection methods with their corresponding pros and cons in the context of autonomous driving and makes comprehensive comparisons of the state-of-the-arts. The researchers attributed current challenges to the visual appearance recovery in the absence of depth information from images, representation learning from partially occluded unstructured point clouds, and semantic alignments over heterogeneous features from cross modalities. Benefiting from the rapid development of deep learning technologies, image-based 3D detection has achieved remarkable progress. Particularly, more than 200 works have studied this problem from 2015 to 2021, encompassing a broad spectrum of theories, algorithms, and applications. The survey titled *3D object detection for autonomous driving* [16] not only filled the gap of collecting and organizing the knowledge of image-based 3D detection but also provided the first comprehensive survey of this novel and continuously growing research field, summarizing the most commonly used pipelines for image-based 3D detection and deeply analyzing each of

[13] Kim, Cheol-jin, et al. "End-to-end deep learning-based autonomous driving control for high-speed environment." The Journal of Supercomputing 78.2 (2022): 1961-1982.

[14] UNet is an encoder–decoder convolutional neural network (CNN) for semantic segmentation.

[15] Shi, Yuguang, et al. "Stereo CenterNet-based 3D object detection for autonomous driving." Neurocomputing 471 (2022): 219-229.

[16] Qian, Rui, Xin Lai, and Xirong Li. "3D object detection for autonomous driving: a survey." Pattern Recognition (2022): 108796.

their components. The main contributions of the survey can be summarized as follows:

(1) The survey provided a comprehensive review and an insightful analysis on the key aspects of the problem, including datasets, evaluation metrics, detection pipelines, and technical details.
(2) The survey also proposed two novel taxonomies of the state-of-the-art methods, with the purpose of helping the readers to easily acquire knowledge on this new and growing research field.
(3) The survey summarized the main issues and future challenges in image-based 3D detection, outlining some potential research directions for future work.

The proposed two novel taxonomies to group the existing image-based 3D detectors are:

(1) The methods based on 2D features. These methods first estimate the 2D locations (and other items such as orientation, depth, etc.) of the objects in the image plane from the 2D features, and then lift the 2D detections into the 3D space. Based on this, these methods can also be called "result lifting-based methods". Besides, because these methods generally share the similar architecture with the 2D detection models, they can be further classified by the common taxonomy used in 2D detection.
(2) The methods based on 3D features. These methods predict the objects based on the 3D features and thus can directly localize the objects in the 3D space. Furthermore, according to how to get the 3D features, these methods can be further grouped into "feature lifting-based methods" and "data lifting-based methods". As the names suggest, the former get the 3D features by lifting the 2D features, while the latter directly extract the 3D features from the 3D data transferred from the 2D images.

Apart from the survey mentioned above, a 3D object detection method --- Stereo CenterNet --- is proposed in another paper titled *Stereo CenterNet-based 3D object detection for autonomous driving*[17]. This method uses geometric information in stereo imagery and predicts the four semantic key points of the 3D bounding box of the object in space and utilizes 2D left and right boxes, orientation, and key points to restore the bounding box of the object in the 3D space. To further optimize the position of the 3D bounding box, an improved photometric alignment module is integrated.

Another method for 3D object detection is a voxel-based method, which is proposed in a paper titled *GVnet: Gaussian model with voxel-based 3D detection network for autonomous driving*[18]. It proposed a two-stage Voxel-based 3D Object detector which named GVnet. In the first stage, Gaussian-Voxel Feature Encoding is used for the raw point cloud, then voxelization

[17] Shi, Yuguang, et al. "Stereo CenterNet-based 3D object detection for autonomous driving." Neurocomputing 471 (2022): 219-229.

[18] Qin, Peilin, Chuanwei Zhang, and Meng Dang. "GVnet: Gaussian model with voxel-based 3D detection network for autonomous driving." Neural Computing and Applications 34.9 (2022): 6637-6645.

is carried out, next, 3D CNN is used to generate high-quality feature maps, the feature map is passed as input to the RPN network to generate a series of 3D proposals. In the second stage, voxel-ROI pooling was used to improve the RPN performance. The corresponding receptive field is obtained through the mapping relationship between raw point and feature map in voxel. Then, the receptive field is regulated by sampling any point of the Gaussian model corresponding to the raw point. This makes the features corresponding to the proposal stronger, and improves the effect of classification and regression tasks. Compared with the traditional voxel feature encoder methods which cannot adjust the quality of detection, it is an improvement to the existing voxel feature encoder as it calculates the corresponding Gaussian distribution of the original point cloud data, and then sampling any number of points by controlling the confidence value to improve the performance of voxel encoder and further improve the quality of the feature map output by the 3D CNN. In addition, a voxel ROI pooling method is also proposed in the paper. In ROI Pooling, the receptive field in the original space and the corresponding raw point are obtained through the mapping relationship between feature and ROI, then change the raw point to adjust the receptive field to improve the performance of classification and regression. The core part of the method is to perform Gaussian clustering on the raw data to obtain a series of Gaussian models that the class obeys, and then voxelize the overall data. When sampling each non-empty voxel, use the Gaussian model corresponding to the voxel to perform this operation. Their experimental results on the KITTI, nuScenes and Waymo dataset has shown that the performance of GVnet under most of the evaluation indexes was better than the current detection methods, at the cost of only a small amount of inference time.

To sum up, there are lots of approaches to optimize the 3D object detection in autonomous driving while using different methods. Based on the approaches reviewed in this paper, general solutions of solving and optimizing the problems in object detection in autonomous driving can be utilized to ensure more safe traffic environment. But since there are lots of challenges along with the different methods stated above, finding a novel approach for 3D object detection to eliminate the problem caused by object detection system in autonomous vehicles still has a long way to go.

REFERENCE

[1] Bogdoll, Daniel, Maximilian Nitsche, and J. Marius Zöllner. "Anomaly Detection in Autonomous Driving: A Survey." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022.

[2] Li, Kaican, et al. "CODA: A Real-World Road Corner Case Dataset for Object Detection in Autonomous Driving." arXiv preprint arXiv:2203.07724 (2022).

[3] Kim, Youngjun, Hyekyoung Hwang, and Jitae Shin. "Robust object detection under harsh autonomous-driving environments." IET Image Processing 16.4 (2022): 958-971.

[4] Zhang, Jiacheng, Huafeng Liu, and Jianfeng Lu. "A semi-supervised 3D object detection method for

autonomous driving." Displays 71 (2022): 102117.

[5] Li, Guofa, et al. "Cross-Domain Object Detection for Autonomous Driving: A Stepwise Domain Adaptative YOLO Approach." IEEE Transactions on Intelligent Vehicles (2022).

[6] Roszyk, Kamil, Michał R. Nowicki, and Piotr Skrzypczyński. "Adopting the YOLOv4 architecture for low-latency multispectral pedestrian detection in autonomous driving." Sensors 22.3 (2022): 1082.

[7] Dasgupta, Kinjal, et al. "Spatio-contextual deep network-based multimodal pedestrian detection for autonomous driving." IEEE Transactions on Intelligent Transportation Systems (2022).

[8] Hu, Hanjiang, et al. "Investigating the Impact of Multi-LiDAR Placement on Object Detection for Autonomous Driving." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022.

[9] Gao, Xin, Guoying Zhang, and Yijin Xiong. "Multi-scale multi-modal fusion for object detection in autonomous driving based on selective kernel." Measurement 194 (2022): 111001.

[10] Wei, Zhiqing, et al. "MmWave Radar and Vision Fusion for Object Detection in Autonomous Driving: A Review." Sensors 22.7 (2022): 2542.

[11] Li, Peixuan, and Jieyu Jin. "Time3D: End-to-End Joint Monocular 3D Object Detection and Tracking for Autonomous Driving." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022.

[12] Kim, Cheol-jin, et al. "End-to-end deep learning-based autonomous driving control for high-speed environment." The Journal of Supercomputing 78.2 (2022): 1961-1982.

[13] Shi, Yuguang, et al. "Stereo CenterNet-based 3D object detection for autonomous driving." Neurocomputing 471 (2022): 219-229.

[14] Qian, Rui, Xin Lai, and Xirong Li. "3D object detection for autonomous driving: a survey." Pattern Recognition (2022): 108796.

[15] Shi, Yuguang, et al. "Stereo CenterNet-based 3D object detection for autonomous driving." Neurocomputing 471 (2022): 219-229.

[16] Qin, Peilin, Chuanwei Zhang, and Meng Dang. "GVnet: Gaussian model with voxel-based 3D detection network for autonomous driving." Neural Computing and Applications 34.9 (2022): 6637-6645.