Deep Multi-modal Object Detection for Autonomous Driving

Amal Ennajar, Nadia Khouja, Rémi Boutteau, Fethi Tlili

Deep Multi-modal Object Detection for Autonomous Driving

Amal Ennajar
Grescom Laboratory
Sup'Com
Tunis, Tunisia
amal.ennajar@supcom.tn

Nadia Khouja
Grescom Laboratory
Sup'Com
Tunis, Tunisia
nadia.khouja@gmail.com

Rémi Boutteau
Normandie Univ, UNIROUEN,
UNILEHAVRE, INSA Rouen, LITIS
76000 Rouen, France
remi.boutteau@univ-rouen.fr

Fethi Tlili
Grescom Laboratory
Sup'Com
Tunis, Tunisia
fethi.tlili@supcom.tn

***Abstract*—Robust perception in autonomous vehicles is a huge challenge that is the main tool for detecting and tracking the different kinds of objects around the vehicle. The aim is to reach the capability of the human level, which is frequently realized by taking utility of several sensing modalities. This lead to make the sensor combination a main part of the recognition system. In this paper, we present methods that have been proposed in the literature for the different deep multi-modal perception techniques. We focus on works dealing with the combination of radar information with other sensors. The radar data are in fact very important mainly when weather conditions and precipitation affect the quality of the data. In this case, it is crucial to have at least some sensors that are immunized against different weather conditions and radar is one of them.**

Keywords — multi-modality, object detection, deep learning, autonomous driving, simulators, datasets, sensors fusion.

I. Introduction

Self driving vehicle also known as an autonomous vehicle that is capable of sensing its environment and can detect obstacles and stop when necessary with the help of sophisticated technology. These vehicles are ordinarily equipped with distinctive sorts of sensors to require advantage of their complementary characteristics. Using numerous sensor modalities increments robustness and precision, The system can perceive, predict, decide and execute the necessary actions required for the autonomous vehicle to navigate in the real world without having any collisions but each of these sensor-families has its corresponding strengths and weaknesses, that's why it requires an intelligent fusion and combination of their data as shown in Fig. 1.

In this paper, we first provide background information on uni-modal recognition sensors and modern deep learning methods for object detection. In section III, we summarize the deep fusion methods for sensors (LIDAR, CAMERA and RADAR) and we discuss the problem that only few researchers have focused on combining radar information with other sensors, although it's highly important. In section IV, we represent the main multimodal datasets with RADAR data, and explain the reason behind using them. In the conclusion, we summarize and emphasize the importance of the sensors above, and mention the importance of focusing more on developing them through more research.

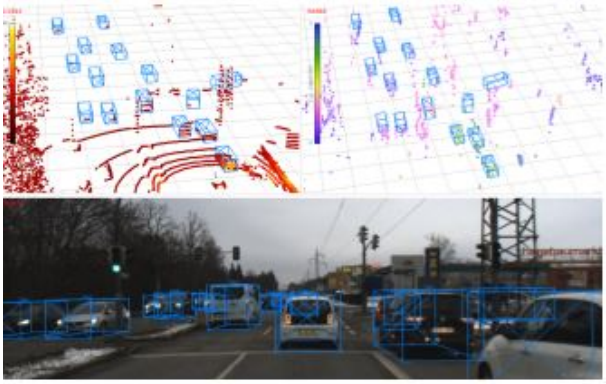

Fig. 1. Multi-sensor 3D detection to ground truth objects [1].

II. Uni-modal Recognition

A. Deep Learning for Images

Camera is the most popular used sensor due to its simplicity and least expensive cost. It is the most popular sensor in autonomous driving vehicles that comes with the highest resolution. Camera gives images with 2D information that can be used in object classification, and in some other tasks like road line tracking.

Traditionally, we are all aware of the importance of camera data, however, it became better exploited thanks to deep learning. It was not properly used mainly for two reasons. The first one is that the models themselves did not exist. The second one is that reaching an efficient computation of image data was not possible. This is no longer true, because now cameras are considered to be an extremely powerful sensor that can be used for object detection with the emergence of the deep Learning algorithms.

In fact, the usage of Convolutional Neural Networks (CNN) [2] has been highly effective within the field of object detection. The CNN is a detection algorithm that aims to require an image and identify accurately where the most objects are found through a selection frame. Detection could be a strategy that looks for classification and finding regions/areas of an image or

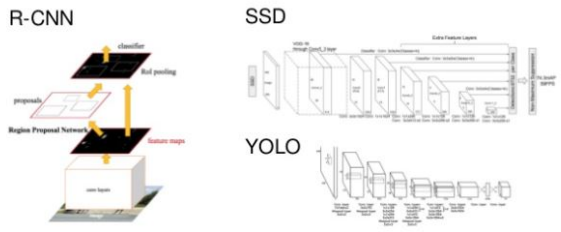

Fig. 2. Architectures of deep learning networks for camera (R-CNN, YOLO, SSD).

a video stream. The detection strategy can mostly be separated into two categories.

In the two-step method, Region-CNN (R-CNN) [3] works in three basic steps. In a first step, it analyzes the input image to extract 2000 Region proposals and, after that, employs a glutton algorithm to recursively combine similar regions into bigger regions. The next step is to run a (CNN) on each of these regions to extract features. Finally, it uses a Support Vector Machine (SVM) to classify the region. Fast R-CNN [4] employs an external proposition generator and exempts from excess feature extractions by using the global features extracted from the complete image to classify each proposition within the second stage. Faster RCNN [5] unifies the proposition generation and classification by introducing the RPN (Region Proposal Network), which employs the global features extracted from the picture to produce object proposals. The SSD [6] (Single Shot MultiBox Finder) performs as one-step method is another object detection algorithm that uses a Region Proposal Network (RPN) and multi-scale representation strategies which employ a default set of anchor boxes with diverse perspective proportions to more precisely locate the object.

Unlike SSD, YOLO (”You Only Look Once”) is a very different object detection algorithm which sees the entire image during the training and test stages. In YOLO [7], a single convolutional network predicts the bounding boxes and the class probabilities for those boxes. YOLO trains on all pictures and specifically optimizes detection performances. Fig.2 presents an overview of the overall the architecture.

B. Deep Learning for Lidar

Lidar or laser detection and ranging is a remote measurement technique based on the analysis of the properties of a beam of light sent back to its transmitter. Over the last ten years this technique has become the most popular sensor for self-driving. It is an unparalleled sensor, in the sense that it combines really accurate, dense depth, and it is an active sensor that has additional benefits for safety and accuracy that can be obtained in object modeling (at least at a geometric level). For better understanding, we see the need to list Many algorithms for Vehicle 3D Detection using Lidar in the next levels.

VoxelNet [8] is a generic 3D detection network that unifies feature extraction and bounding box prediction

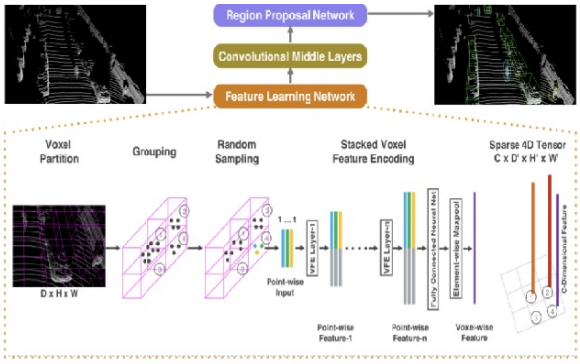

Fig. 3. VoxelNet architecture to generates the 3D detection [8].

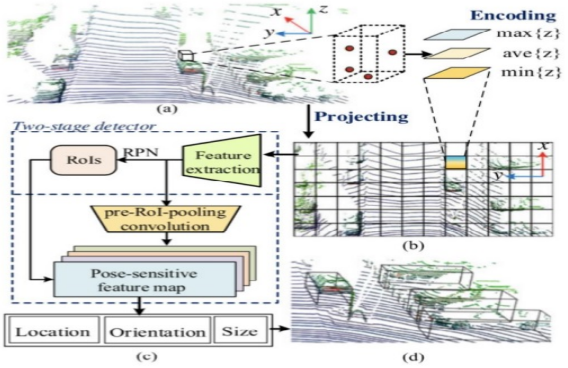

Fig. 4. RT3D architecture is a method using 3D point clouds from a LiDAR to detect vehicles with their oriented 3D bounding boxes. [9].

into a single stage, end-to-end trainable deep network. Specifically, VoxelNet separates a point cloud into similarly dispersed 3D voxels and changes a gathering of points inside each voxel into a unified feature representation through the voxel feature encoding (VFE) layer, as shown in Fig.3. In this way, the point cloud is encoded as a graphic volumetric representation, which is at that point associated to a RPN to produce detections. The Real-time-3-dimensional (RT3D) vehicle detection method extracts features from only a Bird’s Eye View (BEV) representation of the LiDAR data [9]. It then utilizes a CNN-based two-stage detector to generate region proposals with a Region Proposal Network (RPN), and uses pre-RoI-pooling convolutions on pose-sensitive feature maps to classify the region as shown in Fig.4.

PointNet [10] is a deep network architecture that directly operates on the raw point clouds obtained from a LiDAR providing a simple, competent and effective approach for a number of 3D recognition tasks like object classification, part segmentation and semantic segmentation (see Fig. 5). It takes n points as input and applies input and feature transformations. The next step is to aggregates point features using max pooling and finally it provides classification scores for k classes.

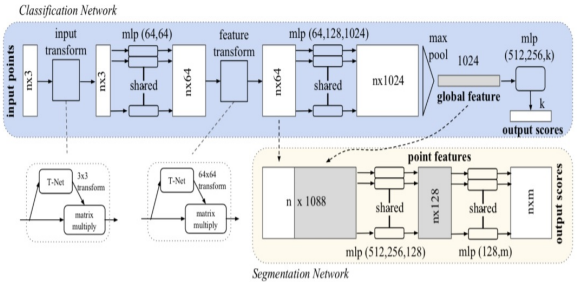

Fig. 5. The PointNet classification network [10].

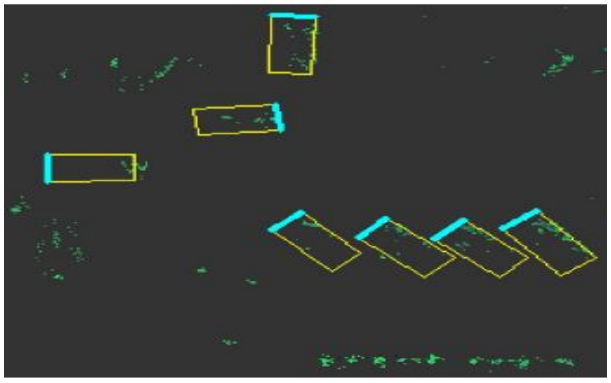

Fig. 6. An Example of bird's eye view detection on radar data [11].

C. Deep Learning for RADAR

Radar or Radio Detection and Ranging, which suggests identifying objects using radio. This system uses electromagnetic waves to detect the presence and determine the position and velocity of objects. For the most part, the working frequency of the vehicle radar framework is 24GHz or 77GHz. Compared with 24GHz, 77GHz appears higher exactness in distance and speed measurements. Moreover, 77GHz features a smaller antenna, and it has less impedance than 24GHz. For 24GHz radars, the most extreme range is 70 meters, while the range increments to 200 meters for 77GHz radars. Compared with a camera, radar is less influenced by the climate and lighting environment, making it very valuable in many applications. In addition, the amount of data is humbler than the camera.

In [11], authors use a deep learning-based method to generate 3D object detection with radar only. They use a public radar dataset "ASTYX HIRES [12]" to train their model. Due to a shortage of radar labeled information, they suggest a novel method by taking advantage of the abundant LiDAR information by transforming it into radar-like point cloud information and then use aggressive radar augmentation strategies (rotation, flip, perturbation, global noise, ...). The result is illustrated in Fig.6.

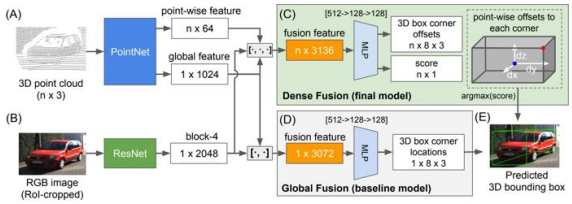

Fig. 7. PointFusion: Deep Sensor Fusion for 3D Bounding Box Estimation [13].

III. MULTI-MODAL RECOGNITION

A. Introduction

Even the best perception algorithms are limited by the quality of their sensor data and to simplify the task the use of multiple sensor modalities can improve the autonomous perception and moreover present new challenges for recognition systems. The fusion of sensors is one of these challenges, which has motivated many researches on object detection (3D or 2D) in recent years. In this section, we summarize deep, multi-modal perception technologies.

B. Data Fusion for LiDAR and camera sensors

Cameras and LiDARs have complementary characteristics that make camera-LiDAR combination models more viable and Well-known compared to other sensor combination setups (radar-camera, LiDAR-radar, etc.,). To be more specific, vision-based recognition frameworks accomplish palatable performance at low-cost, regularly beating human experts. Nevertheless, a mono-camera discernment framework cannot give a solid 3D geometry, which is required for self-driving. On the other hand, stereo cameras can give 3D geometry, but do so at a high computational cost and fail in high-occlusion and texture-less situations. Most later sensor combination strategies focus on harnessing LiDAR and camera for 3D object detection. PointFusion [13] is a generic 3D object detection method that exploits both image and 3D point cloud information. It processes the image and LiDAR information using a CNN and a PointNet architectures and then generates 3D bounding boxes using the extracted features. The result is illustrated in Fig. 7.

In [14], the authors fuse a Bird's Eye View LiDAR point cloud and a front view camera image for object detection in deep Convolutional Neural Networks (CNN). The method creates a layer called sparse non-homogeneous pooling layer to transform features between the bird's eye view and the front view. However, the sparse point cloud is used to construct the mapping between the two views and the pooling layer allows efficient fusion of the multi-view features at any stage of the network as shown in Fig. 8. This method is designed and tested on the KITTI bird's eye view object detection dataset, which produces 3D bounding boxes from the bird's eye view map.

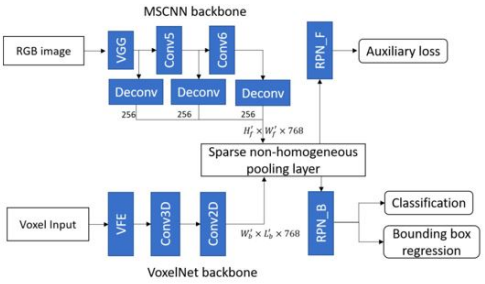

Fig. 8. The fusion-based one-stage object detection network proposed by [14].

C. Fusion (Radar and camera)-(Radar and LiDAR)

LiDARs give precise 3D estimations at near range, but the coming about point cloud gets to be scanty at long extend, decreasing the system capacity to precisely identify far off objects. Cameras offer wealthy appearance characteristics, but are not a great source of data for depth estimation. These extra highlights have made LiDAR-camera sensor combination a theme of investigation in recent years. This combination has been demonstrated to attain high accuracy in 3D object detection for numerous applications, counting autonomous driving, but it has its impediments. Both cameras and LiDARs are touchy to unfavorable climate conditions (eg snow, fog, rain), which can radically decrease their perception range and detection capabilities. Moreover, LiDARs and cameras are not able to identify the speed of objects without utilizing temporal data. Estimating the speed of objects may be a necessity to maintain a security distance to avoid collisions in numerous scenarios, and depending on temporal information may not be a doable arrangement in time. For a long time, radars have been utilized in vehicles for ADAS (Advanced Driver Assistance System) applications to avoid collision and control velocity. Compared to LiDARs and cameras, radars are exceptionally strong to unfavorable climate conditions and are able to distinguish objects at exceptionally long extend (up to 200 meters for car radars). Radars utilize the Doppler effect to precisely gauge the speeds of all identified objects, without requiring any temporal data. In addition, compared to LiDARs, radar point clouds require less handling time and recently they can be utilized as object detection results. These highlights and their lower cost compared to LiDARs have made radars a prevalent sensor in autonomous driving applications. Few research has focused on combining radar information with other sensors.

Radar is very complementary to cameras in unfavorable climate conditions. The sensor quality of the camera is constrained in severe weather condition e.g. if water beads adhere to the camera lens and block the view, or if expanded sensor noise in meagerly lit regions and at night. The CRF-Net (Camera Radar Fusion net) [15] is an architecture to fuse radar detection with images. The image information is increased with the radar data and utilized in a CNN to extract 2D object detection.

RadarNet [16] is a method to exploit Radar in combination with LiDAR for the detection of 3D objects. It employs an early fusion technique to memorize the joint representations of the two sensors and a late fusion to refine object velocities. In [17], a method is introduced to fuse radar detections with images and utilize them to boost the object accuracy. The real-time performances are very important for autonomous driving vehicles. Thus, in [18], the authors use a real-time region proposal network (RRPN) for object detection. First, they use radar detections to propose ROIs, and then, they project them onto the image plane to obtain faster and more accurate detections.

IV. DATASETS AND SIMULATORS

A. Introduction

Various algorithms of deep multi-modal object detection are based on supervised learning. For that reason, multi-modal datasets with labeled ground-truth are required to train the deep learning neural networks. Labeling the data is a fastidious process, that is why it is preferable to use an already existing dataset or simulator. In the following, we introduce the simulators and datasets containing RADAR data.

B. Simulators

To generate the data needed for learning, or to test the algorithms on new scenarios, there are now a few open-source vehicle simulators available.

CARLA (Intel): an open-source simulator for autonomous driving research. CARLA is a platform for the evaluation of autonomous urban driving systems. In this platform agents are tasked to safety drive from a set of starting points to target destinations, following route instructions and respecting traffic rules. CARLA lets users configure a sensor suite choosing from a range of sensors and offers the access to several maps [19].

LGSVL (LG): The LGSVL simulator is a test system that encourages testing and advancement of self-driving program frameworks. It empowers designers to simulate billions of miles and self-assertive edge case scenarios to speed up algorithm advancement and framework integration [20].

DeepDrive (Voyage): Voyage Deepdrive is a simulator that permits anybody with a PC to thrust the state-of-the-art in self-driving. [21]

AirSim (Microsoft): AirSim is a simulator for drones, cars and more, built on Unreal Engine. It is open-source, cross stage, and underpins software-in-the-loop reenactment with popular flight controllers such as PX4 ArduPilot and hardware-in-loop with PX4 for physically and visually reasonable recreations. It is created as an Unreal plugin that can essentially be dropped into any Unreal environment. [22]

C. Datasets

Most multi-modal recognition strategies are based on supervised learning. Subsequently, multi-modal datasets with labeled ground-truth are required for preparing such deep neural networks. While there are many datasets for autonomous vehicles containing LiDAR and camera data, there are relatively few datasets containing radar data as well.

nuScenes dataset [23]: NuScenes is a large-scale open dataset for autonomous vehicles. This dataset empowers researchers to ponder urban driving circumstances utilizing the complete sensor suite of a real self-driving car: 6 cameras, 5 radars, 1 lidar, and an RTK GPS. NuScenes comprises more than 1000 scenes, with annotated 3D bounding boxes. Moreover, this dataset contains annotations for 23 different classes.

Astyx HiRes2019 [12]: The Astyx Dataset HiRes2019 is a radar dataset for deep learning-based 3D object detection. The motivation behind this dataset is to provide high-resolution radar information to the scientific community, encouraging and stimulating research on using radar sensor information.

Oxford RobotCar [24]: The Oxford RobotCar Dataset contains over 100 repetitions of a reliable course through Oxford, UK, captured over a period of over a year. The dataset captures numerous diverse combinations of climate, traffic and pedestrians, together with longer term changes such as development and roadworks.

V. CONCLUSION

In this paper, we have presented the different sensors used in autonomous vehicles and their various algorithms for deep uni-modal object detection. Furthermore, we have shown that these sensors are complementary and that it is necessary to fuse their information to obtain more accurate and robust detections. We have presented the different multi-modal fusion methods and showed their interest.

We have pointed out the lack of work on multi-modal fusion involving radar data, even though it is of great interest since it is less sensitive to environmental conditions and provides additional information such as speed. We have presented the simulators and datasets that would allow work in this direction, i.e. work on multi-modal fusion with radar data.

We have also shown that many algorithms perform a fusion on the output of the detections in the different modalities, but that there is little work on the early-fusion of these modalities (at the input of the network). As a future work, we intend to propose a radar-camera-lidar early-fusion algorithm for 3D object detection to obtain better results.

REFERENCES

[1] M. Meyer and G. Kuschk, "Deep learning based 3d object detection for automotive radar and camera," in *2019 16th European Radar Conference (EuRAD)*, pp. 133–136, IEEE, 2019.

[2] R. Yamashita, M. Nishio, R. K. G. Do, and K. Togashi, "Convolutional neural networks: an overview and application in radiology," *Insights into imaging*, vol. 9, no. 4, pp. 611–629, 2018.

[3] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 580–587, 2014.

[4] R. Girshick, "Fast r-cnn," in *Proceedings of the IEEE international conference on computer vision*, pp. 1440–1448, 2015.

[5] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," in *Advances in neural information processing systems*, pp. 91–99, 2015.

[6] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, "Ssd: Single shot multibox detector," in *European conference on computer vision*, pp. 21–37, Springer, 2016.

[7] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 779–788, 2016.

[8] Y. Zhou and O. Tuzel, "Voxelnet: End-to-end learning for point cloud based 3d object detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4490–4499, 2018.

[9] Y. Zeng, Y. Hu, S. Liu, J. Ye, Y. Han, X. Li, and N. Sun, "Rt3d: Real-time 3-d vehicle detection in lidar point cloud for autonomous driving," *IEEE Robotics and Automation Letters*, vol. 3, no. 4, pp. 3434–3440, 2018.

[10] C. R. Qi, H. Su, K. Mo, and L. J. Guibas, "Pointnet: Deep learning on point sets for 3d classification and segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 652–660, 2017.

[11] S. Lee, "Deep learning on radar centric 3d object detection," *arXiv preprint arXiv:2003.00851*, 2020.

[12] *M. Meyer and G. Kuschk, "Automotive radar dataset for deep learningbased 3d object detection," in Proceedings of the 16th European RadarConference, 2019.*

[13] D. Xu, D. Anguelov, and A. Jain, "Pointfusion: Deep sensor fusion for 3d bounding box estimation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 244–253, 2018.

[14] Z. Wang, W. Zhan, and M. Tomizuka, "Fusing bird view lidar point cloud and front view camera image for deep object detection," *arXiv preprint arXiv:1711.06703*, 2017.

[15] F. Nobis, M. Geisslinger, M. Weber, J. Betz, and M. Lienkamp, "A deep learning-based radar and camera sensor fusion architecture for object detection," in *2019 Sensor Data Fusion: Trends, Solutions, Applications (SDF)*, pp. 1–7, IEEE, 2019.

[16] B. Y. et R. Guo et Ming Liang et S. Casas et R. Urtasun, "Radarnet: Exploiting radar for robust perception of dynamic objects," *ArXiv*, vol. abs / 2007.14366, 2020.

[17] S. Chadwick, W. Maddetn, and P. Newman, "Distant vehicle detection using radar and vision," in *2019 International Conference on Robotics and Automation (ICRA)*, pp. 8311–8317, IEEE, 2019.

[18] R. Nabati and H. Qi, "Rrpn: Radar region proposal network for object detection in autonomous vehicles," in *2019 IEEE International Conference on Image Processing (ICIP)*, pp. 3093–3097, IEEE, 2019.

[19] A. Dosovitskiy, G. Ros, F. Codevilla, A. Lopez, and V. Koltun, "CARLA: An open urban driving simulator," in *Proceedings of the 1st Annual Conference on Robot Learning*, pp. 1–16, 2017.

[20] L. E. A. R. Center, *LGSVL Simulator - An Autonomous Vehicle Simulator*, 2020 (accessed November 6, 2020).

[21] Voyage, *Deepdrive from Voyage - Push the state-of-the-art in self-driving*, 2020 (accessed November 7, 2020).

[22] Microsoft, *AirSim - Microsoft Open Source*, 2020 (accessed November 7, 2020).

[23] H. Caesar, V. Bankiti, A. H. Lang, S. Vora, V. E. Liong, Q. Xu, A. Krishnan, Y. Pan, G. Baldan, and O. Beijbom, "nuscenes: A multimodal dataset for autonomous driving," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11618–11628, 2020.

[24] *W. Maddern, G. Pascoe, C. Linegar, and P. Newman, "1 year, 1000km: The Oxford RobotCar dataset," Int. J. Robotics Research, vol. 36,no. 1, pp. 3–15, 2017.*