Multimodal Deep Learning for Advanced Driving Systems

Nerea Aranjuelo[1], Luis Unzueta[1], Ignacio Arganda-Carreras[2,3,4], and Oihana Otaegui[1]

[1] Vicomtech, San Sebastian, Spain
naranjuelo@vicomtech.org
[2] Basque Country University (UPV/EHU), San Sebastian, Spain
[3] Ikerbasque, Basque Foundation for Science, Bilbao, Spain
[4] Donostia International Physics Center (DIPC), San Sebastian, Spain

Abstract. Multimodal deep learning is about learning features over multiple modalities. Impressive progress has been made in deep learning solutions that rely on a single sensor modality for advanced driving. However, these approaches are limited to cover certain functionalities. The potential of multimodal sensor fusion has been very little exploited, although research vehicles are commonly provided with various sensor types. How to combine their data to achieve a complex scene analysis and improve therefore robustness in driving is still an open question. While different surveys have been done for intelligent vehicles or deep learning, to date no survey on multimodal deep learning for advanced driving exists. This paper attempts to narrow this gap by providing the first review that analyzes existing literature and two indispensable elements: sensors and datasets. We also provide our insights on future challenges and work to be done.

Keywords: Autonomous Driving, ADAS, Deep Learning, Sensor Fusion

1 Introduction

In the last decade self-driving vehicle research has gained large attention. Advances in algorithms, along with improvements in sensor technology, are promoting the race to develop driverless vehicles. Efforts are not only focused on achieving total autonomy. Advanced Driver Assistance Systems (ADAS) have become part of most recent vehicles, such as automatic parking assistance or traffic sign recognition. They promote a human-machine interaction environment that combines complementary strengths of humans and machines in order to achieve higher performance than each one by themselves. Impressive progresses in machine learning have played a key role in the up-rising of both paradigms, most of them thanks to high-level feature extraction based on deep learning.

Deep learning algorithms have quickly become a method of choice for most machine learning problems. The renaissance and evolution of Artificial Neural Networks (ANNs) in the form of Deep Neural Networks (DNNs) has changed the way the researchers work. Moving from hand-crafted features to machine-learned features has marked a milestone in computer vision tasks [1, 2], where state-of-the-art results achieved by traditional algorithms have been rapidly overcome. Deep learning has been widely applied to the field of driving in last years, frequently to process data coming from a single sensor modality, most of times cameras, and in some cases to process data from various modalities, for example LiDAR and camera. Even if some works have been recently developed for the last case, there is still a lot of uncertainty on how to process and combine data from heterogeneous sensors in the best way. This topic is an open question to which this paper aims at bringing some light.

Even though some works have attempted to use multimodal deep learning approaches in driving scenarios [3, 4] and some reviews have studied independently autonomous driving [5] and multimodal deep learning [6], no deep study has ever been done for its application in the automotive field. This topic entails two additional challenges that therefore have not been analyzed either: which are the key sensors to bring intelligence to vehicles and how can researchers obtain large-scale data from them. To the best of our knowledge, this is the first work that analyzes sensor modalities, datasets and current state of the art.

This paper contributes and responds to the following needs:

1. Analysis of most frequent sensors for advanced driving research: characteristics, strengths and limitations
2. Data sources for multisensory approaches: real-world and synthetic data
3. Comparison of multimodal data coverage in public datasets
4. Current state of the art in multimodal deep learning approaches for ADAS and autonomous driving systems
5. Challenges and future work to be done

2 Sensor Modalities for Driving

Intelligent vehicles understand their surrounding based on the information fed to their computer system by processing signals from their external sensors. Various technologies are needed in order to have enough data to detect, predict and react to the surrounding environment factors, such as other road users. The range of possible sensors is wide, as shown in Fig. 1, and each one presents specific strengths and limitations. Nevertheless, the combination of sensor modalities provides loads of both complementary and redundant data, which favors their use in safety-critical environments. The most common sensor types are depicted in Fig. 1 and described in the sequel.

- Light Detection And Ranging (LiDAR): This laser-based system measures distances to surrounding objects sending out high-speed pulses of laser-light and calculating the reflection time of the beams. The collision location between an object and the laser beam is represented as a point in a 3D point cloud. There are two main trends in models: systems which use a rotating laser, or solid-state LiDARs, which have no moving parts.

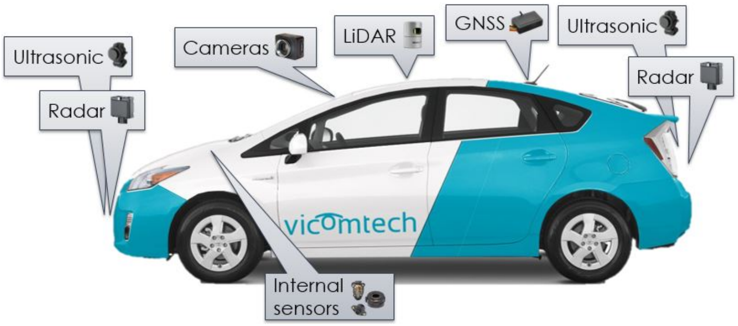

Fig. 1. Frequent sensors self-driving research vehicles are equipped with.

They cover high distances accurately, usually more than 100 meters up to 200 meters range. It is not affected by different lighting conditions, however it is not able to perceive color or textures. It can provide noisy measurements when suspended particles are present in the air, such as rain, fog or snow. LiDARs allow an accurate 3D analysis and are mainly used for mapping, obstacle avoidance, free space detection on road and localization [7, 8].

- Vision Cameras: Multiple cameras are often installed in vehicles to have a detailed sight of the environment, covering the front or back view or even 360° around the vehicle.

They cover a medium distance. Cameras preserve detailed semantic information of the surrounding, making it possible to interpret objects such as traffic signs. They are sensitive to lighting conditions, they do not work well at nighttime or in dazzling sunlight and often acquire poor-quality images under unfavorable weather conditions. Cameras need to be calibrated to address 3D measurements. They are mainly used for object detection, road guidance and park assistance [9, 10]. Internal cameras are also common for driver monitoring.

- Thermal Cameras: Thermal cameras detect the heat from pedestrians, animals and objects, and consequently the differences in temperatures emitted by living and inanimate objects. Although this sensor has been widely applied to different fields for years, its use is not so widespread for self-driving research.

They cover a medium distance. These cameras are advantageous to help edge cases, where some sensors might have difficulties, for example differentiating between images of humans and real ones, as well as poor lighting scenes when vision cameras have problems. They are mainly used for object detection [11].

- Ultrasonic Sensors: Ultrasonic sensors send out sound waves at a high frequency imperceptible by the human ear and measure the time it takes for the signal to return. This way, the distance to an object can be calculated.

They cover short distances. Due to the sensing fundamentals, air temperature, humidity and wind can affect the accuracy of the sensor, as they affect the speed of sound in air. They are employed for short-distance applications at low speeds, such as park assistance or close obstacle and blind spot detection [12].

- Radio Detection And Ranging (Radar): This sensor emits radio waves which are reflected when they hit an obstacle, revealing the distance to the object and how fast it is approaching. Radar can be categorized based on different operating distance ranges, starting from 0.2m to more than 200m, in Short Range Radar (SRR), Medium Range Radar (MRR) and Long Range Radar (LRR).

They are affected much less than other sensors by weather conditions, such as rain, fog, dust and snow. Nonetheless, they can be confused by small very reflective metal objects. They do not provide any information about detected object types. They are usually used for very close obstacle avoidance [13].

- Global Navigation Satellite System (GNSS): It is a global localization system which triangulates multi constellation satellite signals to calculate the 3D position of the receiver (its latitude, longitude and altitude). Currently, the considered GNSS providers are GPS, GLONASS and Galileo.

The absolute position provided by this technology is affected by several error sources, such as lonosphere, multipath effect or urban canyons, and therefore it is not enough to achieve lane level accuracy. The position is usually enhanced by either using Differential GPS or by fusing its information with inertial sensors like Inertial Measurement Units (IMUs) and accelerometers. It is used for localizing the ego vehicle itself and for path planning [14].

In addition to external sensors, internal vehicle parameters also provide a very relevant information source for driving. These signals are available through the vehicle's Controller Area Network (CAN) bus and include parameters such as wheel speed, acceleration, steering and powertrain values. In the last years the vehicle communication with the cloud has been included in the intelligent driving scenario, due to the possibility of sharing real-time map data and anticipating to different situations. Information exchange with other vehicles or infrastructures is also considered in cooperative systems [15], through the use of vehicle-to-vehicle (V2V) and infrastructure-to-vehicle (I2V) communication.

Signal capturing and processing from the aforementioned sensor modalities is a complex task, which requires addressing various aspects such as sensor capturing synchronization. There exist some libraries and tools that help in this procedure, mainly in capturing, recording and managing sensor timestamps or integrating developed approaches. Among the most frequent ones we find RTMaps [16], ADTF [17] and ROS [18]. At the same time, some public datasets provide already captured multisensory data that is ready to be used by researchers who want to start deploying their approaches. This is described in next section.

3 Real and CGI-Generated Datasets and Simulators

Annotated data is indispensable to develop and train deep learning models, but also to generate a quantitative evaluation of them. However, collecting large amounts of annotated data with quality is a tedious and complex work and it is often beyond the reach of researchers. In an effort to alleviate these needs, some large datasets have been made public. Despite this, multimodal approaches present an additional difficulty for data gathering. Many of the open datasets are focused on solving a specific problem and do not include data from all desired sensor modalities. In the following, we will go through relevant datasets dedicated to advanced driving research and the data type they include.

Real-World Data Real-world datasets, although very costly to obtain, are crucial to deploy and test algorithms under real conditions. Table 1 summarizes the most recent and relevant real-world datasets for the driving context, within the sensor modalities they include and most relevant information. None of the datasets contains data from ultrasonic or Radar sensor.

Table 1. Sensor modalities in large-scale datasets

Dataset	Relevant information	Vision Camera	Thermal Camera	LiDAR	GNSS /IMU	Internal params.
KITTI [19]	6 hours of recordings, multi task annotations	✓		✓	✓	
Cityscapes [20]	segmentation benchmark, coarse and accurate labels	✓				
TorontoCity [21]	multitask annotations, various perspectives	✓		✓		
Paris-Lille [22]	point cloud segmentation and classification			✓	✓	
RobotCar [23]	recorded in Oxford through a year	✓		✓	✓	
Comma.ai [24]	11 videos, mostly on highway	✓			✓	✓
BDDV [25]	400 hours of HD video, multitask annotations	✓			✓	✓
Mapillary [26]	segmentation annotations, 66 classes	✓				
KAIST [27]	multispectral, bounding box annotations	✓	✓	✓	✓	

CGI-Generated Data The large costs and difficulties of creating large enough datasets to train deep learning models have led researchers to look for alternatives in the field of Computer-Generated Imagery (CGI).

The creation of realistic virtual worlds facilitates the automatic generation of ground truth annotations. For example, Synthia dataset [28], based on a virtual city, includes automatically extracted pixel-level and instance-level semantic annotations in both videos and independent snapshots. Motivated by the fact that generating a realistic virtual world can be a very arduous task, some works have proposed to take advantage of already developed video games [29].

In self-driving and ADAS, simulation has an additional application. Developed approaches must be evaluated during a vast number of kilometers and varying conditions to ensure and demonstrate they are safe, which could be impractical in its entirety. In addition, dangerous or uncommon driving situations must be evaluated. As a solution, various simulation environments have been released, such as AirSim [30] or Carla [31]. They facilitate the user collecting data from different sensors or integrating developed approaches to be tested.

4 Multimodal Deep Learning Approaches

In this section, we investigate the multisensory deep learning, popularly known as multimodal deep learning, for the two major paradigms for automated driving.

Mediated perception approaches Mediated perception approaches [32] rely on the decomposition of autonomous driving into multiple sub-components that are combined to obtain a comprehensive understanding of the vehicle surroundings. This perception information is often used to feed on-board world models such as Local Dynamic Maps (LDM) [15]. A large variety of tasks is included in which different sensor types are present, combined or alone. In this diversity, vision-based systems are maturing within their limitations of the data type [1, 2], while models that incorporate other data modalities are still emerging with no standardized methodologies to follow.

For road segmentation, [8] trained a model that fuses camera and LiDAR data through a hybrid Conditional Random Field (CRF). In the field of path prediction, [14] proposes a LSTM architecture that estimates the future position of obstacles given a sequence of their past trajectory data obtained from sensors like LiDAR and GPS. In [11], multispectral pedestrian detection is proposed using a CNN that fuses color and thermal images. They also explore results obtained by early, halfway and late fusion of images in the architecture.

Among the different tasks, object detection plays a key role in mediated perception approaches. It has advanced notably in the domain of 2D in the last years, but self-driving vehicles also need 3D information. For this task, advantage of various sensor modalities and their strengths can be taken.

To date, different strategies have been proposed. Some works suggest using already mature 2D detectors to perform 3D object detection from monocular images. For example, Deep3DBox [9] estimates 3D bounding boxes from 2D detections using a DNN and geometric constraints. Deep MANTA [33] does 2D detection, part localization and 3D dimension estimation from monocular images, using also a dataset of 3D models of different types of vehicles. Even if these methods work well compared to other monocular based approaches, they perform poorer than models that use point cloud data to complete the task.

There are other works that rely only on point cloud data to perform detection. 3D-FCN [34] uses 3D convolutions to generate the detections, which require expensive computations. VeloFCN [35] projects LiDAR point clouds to the front view in order to apply a Fully Convolutional Network (FCN) and generate 3D bounding boxes. VoxelNet [7] encodes voxels with point-wise features and includes 3D convolutions in its network architecture.

However, models that fuse both images and point clouds achieve better results. In MV3D [36], Chen et al. propose an object detection model that fuses data from images and LiDAR point clouds, which they project on a bird's eye view and a frontal view. They extend the image based Region Proposal Network (RPN) of Faster R-CNN [37] to 3D, so that 3D proposals are created based on the bird's eye view. The features of these candidates in all views are then combined through a deep fusion to produce the final result. AVOD [38] also feeds a RPN with the extracted features, but in this case not only from bird's eye view, also from image, so that candidates are generated and transfered to a second detection network, which estimates bounding box orientation, refines dimensions and classifies them. Wang et al. [39] as well focus on an image and point cloud based feature fusion before the region proposal stage, through a new layer they called non-homogeneous pooling layer.

Behavior reflex approaches Behavior reflex approaches build a direct mapping from sensory input to a driving reaction, such as turning left or braking. Here belong the so-called end-to-end driving approaches.

In 2016, [40] trained a CNN to map directly images from single frontal-facing camera to steering commands. 72 hours of driving were collected to train the model. Other works have extended the input information so that not only images feed the model. In [41], Xu et al. propose a FCN-LSTM architecture trained with large-scale crowd-sourced data [25]. The input images are processed by a dilated FCN [42] and the segmented images are concatenated with previous sensor information, such as speed and angular velocity, to feed a LSTM and predict driver actions. In the same research direction, Deep Steering [43] proposes an architecture of DNNs that combines spatial and temporal information and uses a convolutional LSTM to predict the steering actions. Extracted feature vectors from consecutive frames are combined with previous steering actions and vehicle status in two different concatenation layers.

Point cloud data is also present in some multimodal approaches. The method proposed in [44] fuses the depth and vision from LiDAR and camera to predict steering commands. Features from RGB images and depth range images are extracted independently through a series of convolutional operations, then combined to be transfered to fully connected layers and finally predict the commands. The network was trained with some corrupted samples from one of the sensors to achieve robustness in case of sensor failure. [4] also considers the handling of partial failure in the sensor-set and proposes a solution by introducing Sensor Dropout, which randomly drops sensor modules in the training stage. They consider a Deep Reinforcement Learning (DRL) setup for their tests.

Chowdhuri et al. [3] trained a CNN to predict steering vehicle control based on the fusion of images and data packets they called data moments. The model takes as input 4 RGB images, left and right images from a stereo camera and a pair of images from current and past timestamps to perceive motion, behavioral information and a collection of speed and steering angle values. All this information is fused after applying the first convolutional layer to the images, so that basic image processing is done without considering the rest of the data.

5 Discussion and Conclusions

We studied most relevant sensors used in advanced driving research. None of them alone is enough to complete a complex perception of the environment, but a right combination of them brings the opportunity to overcome their limitations and benefit from their strengths. In addition, the use of various sensor modalities entails some data redundancy, which is a basic feature until the driving technologies are mature enough. However, the quantity of processed data cannot be unlimited due to real-time, resources and energy restrictions. Which are the sensors that best complement each other and whether any of them is dispensable is an open question that will be answered as the different approaches progress.

Gathering enough data to train models, based on the analyzed sensor modalities, is not only difficult but often impracticable, therefore we went through the different available datasets, focusing on the sensor types they include. Most of current public datasets do not contain all the desired sensor modalities, so simulation was presented as an alternative to collect specific data under different conditions. Released simulation environments are also an interesting tool to test developed approaches safely in various use cases and a vast number of kilometers.

Current literature in multimodal deep learning was presented for the main paradigms in self-driving vehicles, which propose different data fusions that may benefit from being combined. LiDAR and image based fusion approaches stand out, specially in mediated perception approaches, as a way to obtain more accurate results. Nonetheless, approaches still do not share a clear pipeline on how the features of both should be best combined. In behavior reflex approaches, some models benefit from having different sensors to include in their architectures techniques that face specific sensor failures. This is a crucial idea that should be integrated also in mediated perception developments.

We hope that our survey will encourage researchers to continue working on multimodal deep learning, as it has the potential to be a key methodology for self-driving vehicles and ADAS but still entails a lot of research to be done. We strongly believe that developing methods that benefit from the combination of sensor modalities and provide an optimal and efficient way to integrate their information is a highly relevant open research topic.

Acknowledgments

This work has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement n 6880099, project Cloud-LSVA).

References

1. Guo, Y., Liu, Y., Georgiou, T., Lew, M.S.: A review of semantic segmentation using deep neural networks. IJMIR (2017) 1–7

2. Han, J., Zhang, D., Cheng, G., Liu, N., Xu, D.: Advanced Deep-Learning Techniques for Salient and Category-Specific Object Detection: A Survey. IEEE Signal Processing Magazine **35**(1) (2018) 84–100
3. Chowdhuri, S., Pankaj, T., Zipser, K.: Multi-Modal Multi-Task Deep Learning for Autonomous Driving. CoRR **abs/1709.05581** (2017)
4. Liu, G.H., Siravuru, A., Prabhakar, S., Veloso, M., Kantor, G.: Learning End-to-end Multimodal Sensor Policies for Autonomous Navigation. arXiv preprint arXiv:1705.10422 (2017)
5. Janai, J., Güney, F., Behl, A., Geiger, A.: Computer vision for autonomous vehicles: Problems, datasets and state-of-the-art. arXiv preprint arXiv:1704.05519 (2017)
6. Ngiam, J., Khosla, A., Kim, M., Nam, J., Lee, H., Ng, A.Y.: Multimodal deep learning. In: Proceedings of the 28th ICML-11. (2011) 689–696
7. Zhou, Y., Tuzel, O.: VoxelNet: End-to-End Learning for Point Cloud Based 3D Object Detection. arXiv preprint arXiv:1711.06396 (2017)
8. Xiao, L., Wang, R., Dai, B., Fang, Y., Liu, D., Wu, T.: Hybrid conditional random field based camera-LiDAR fusion for road detection. Information Sciences (2017)
9. Mousavian, A., Anguelov, D., Flynn, J., Košecká, J.: 3d bounding box estimation using deep learning and geometry. In: Computer Vision and Pattern Recognition (CVPR), 2017 IEEE Conference on, IEEE (2017) 5632–5640
10. Oliveira, G.L., Burgard, W., Brox, T.: Efficient deep models for monocular road segmentation. In: Intelligent Robots and Systems (IROS), 2016 IEEE/RSJ International Conference on, IEEE (2016) 4885–4891
11. Liu, J., Zhang, S., Wang, S., Metaxas, D.N.: Multispectral deep neural networks for pedestrian detection. arXiv preprint arXiv:1611.02644 (2016)
12. Carullo, A., Parvis, M.: An ultrasonic sensor for distance measurement in automotive applications. IEEE Sensors journal **1**(2) (2001) 143–147
13. Lombacher, J., Hahn, M., Dickmann, J., Wöhler, C.: Potential of radar for static object classification using deep learning methods. In: Microwaves for Intelligent Mobility (ICMIM), IEEE MTT-S International Conference on, IEEE (2016) 1–4
14. Virdi, J.: Using Deep Learning to Predict Obstacle Trajectories for Collision Avoidance in Autonomous Vehicles. PhD thesis, UC San Diego (2017)
15. Shimada, H., Yamaguchi, A., Takada, H., Sato, K.: Implementation and evaluation of local dynamic map in safety driving systems. JTTs **5**(02) (2015) 102
16. Intempora: RTMaps. https://intempora.com/products/rtmaps.html (Accessed on 03/18/2018).
17. Elektrobit: EB Assist ADTF. https://www.elektrobit.com/products/eb-assist/adtf/ (Accessed on 03/18/2018).
18. ROS. http://www.ros.org/ (Accessed on 03/18/2018).
19. Geiger, A., Lenz, P., Urtasun, R.: Are we ready for Autonomous Driving? The KITTI Vision Benchmark Suite. In: Conference on CVPR. (2012)
20. Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., Franke, U., Roth, S., Schiele, B.: The cityscapes dataset for semantic urban scene understanding. In: Proc. of the IEEE conference on CVPR. (2016) 3213–3223
21. Wang, S., Bai, M., Mattyus, G., Chu, H., Luo, W., Yang, B., Liang, J., Cheverie, J., Fidler, S., Urtasun, R.: Torontocity: Seeing the world with a million eyes. arXiv preprint arXiv:1612.00423 (2016)
22. Roynard, X., Deschaud, J.E., Goulette, F.: Paris-Lille-3D: a large and high-quality ground truth urban point cloud dataset for automatic segmentation and classification. arXiv preprint arXiv:1712.00032 (2017)
23. Maddern, W., Pascoe, G., Linegar, C., Newman, P.: 1 year, 1000 km: The Oxford RobotCar dataset. International Journal of Robotics Research **36**(1) (2017) 3–15

24. Santana, E., Hotz, G.: Learning a driving simulator. arXiv preprint arXiv:1608.01230 (2016)
25. Xu, H., Gao, Y., Yu, F., Darrell, T.: End-to-end learning of driving models from large-scale video datasets. arXiv preprint (2017)
26. Neuhold, G., Ollmann, T., Bulò, S.R., Kontschieder, P.: The mapillary vistas dataset for semantic understanding of street scenes. In: Proc. ICCV. (2017) 22–29
27. Choi, Y., Kim, N., Hwang, S., Park, K., Yoon, J.S., An, K., Kweon, I.S.: KAIST Multi-Spectral Day/Night Data Set for Autonomous and Assisted Driving. IEEE Transactions on Intelligent Transportation Systems (2018)
28. Ros, G., Sellart, L., Materzynska, J., Vazquez, D., Lopez, A.: The SYNTHIA Dataset: A Large Collection of Synthetic Images for Semantic Segmentation of Urban Scenes. (2016)
29. Richter, S.R., Vineet, V., Roth, S., Koltun, V.: Playing for data: Ground truth from computer games. In: European Conference on CV, Springer (2016) 102–118
30. Shah, S., Dey, D., Lovett, C., Kapoor, A.: AirSim: High-Fidelity Visual and Physical Simulation for Autonomous Vehicles. In: Field and Service Robotics. (2017)
31. Dosovitskiy, A., Ros, G., Codevilla, F., López, A., Koltun, V.: CARLA: An open urban driving simulator. arXiv preprint arXiv:1711.03938 (2017)
32. Ullman, S.: Against direct perception. BBS **3**(3) (1980) 373–381
33. Chabot, F., Chaouch, M., Rabarisoa, J., Teulière, C., Chateau, T.: Deep MANTA: A Coarse-to-fine Many-Task Network for joint 2D and 3D vehicle analysis from monocular image. In: Proc. IEEE CVPR. (2017) 2040–2049
34. Li, B.: 3D fully convolutional network for vehicle detection in point cloud. arXiv preprint arXiv:1611.08069 (2016)
35. Li, B., Zhang, T., Xia, T.: Vehicle detection from 3d lidar using fully convolutional network. arXiv preprint arXiv:1608.07916 (2016)
36. Chen, X., Ma, H., Wan, J., Li, B., Xia, T.: Multi-view 3d object detection network for autonomous driving. In: IEEE CVPR. Volume 1. (2017) 3
37. Ren, S., He, K., Girshick, R., Sun, J.: Faster r-cnn: Towards real-time object detection with region proposal networks. In: Advances in NIPS. (2015) 91–99
38. Ku, J., Mozifian, M., Lee, J., Harakeh, A., Waslander, S.: Joint 3D Proposal Generation and Object Detection from View Aggregation. arXiv preprint arXiv:1712.02294 (2017)
39. Wang, Z., Zhan, W., Tomizuka, M.: Fusing Bird View LIDAR Point Cloud and Front View Camera Image for Deep Object Detection. arXiv preprint arXiv:1711.06703 (2017)
40. Bojarski, M., Del Testa, D., Dworakowski, D., Firner, B., Flepp, B., Goyal, P., Jackel, L.D., Monfort, M., Muller, U., Zhang, J., et al.: End to end learning for self-driving cars. arXiv preprint arXiv:1604.07316 (2016)
41. Xu, H., Gao, Y., Yu, F., Darrell, T.: End-to-end learning of driving models from large-scale video datasets. arXiv preprint (2017)
42. Yu, F., Koltun, V.: Multi-scale context aggregation by dilated convolutions. arXiv preprint arXiv:1511.07122 (2015)
43. Chi, L., Mu, Y.: Deep Steering: Learning End-to-End Driving Model from Spatial and Temporal Visual Cues. arXiv preprint arXiv:1708.03798 (2017)
44. Patel, N., Choromanska, A., Krishnamurthy, P., Khorrami, F.: Sensor Modality Fusion with CNNs for UGV Autonomous Driving in Indoor Environments. In: International Conference on Intelligent Robots and Systems (IROS). IEEE. (2017)