A Literature Survey of Safety Critical Multi-Modal-Object-Detection in Autonomous System

Om Tiwari,
Master's informatic(of Aff.),
Technical University Kaiserslautern (of Aff.),
Kaiserslautern, Germany,
tiwari@rhrk.uni-kl.de

***Abstract*—In recent years, there has been an increasing use of autonomous systems in various applications that has highlighted the need for ensuring safety in these systems. Object detection, the task of identifying and locating objects in an environment, is a crucial component in ensuring the safety of autonomous systems. Multi-modal object detection, which involves using multiple sensors or modalities, has emerged as a promising approach to improve the reliability and robustness of object detection in these systems. This literature survey provides an overview of the recent advances in safety-critical multi-modal object detection in autonomous systems. The paper will focus on the various modalities used for object detection and their advantages, sensor fusion techniques, data association methods and performance evaluation techniques. The paper will also highlight the open research problems and future directions in the field.**

Overall, this literature survey aims to demonstrate the potential of multi-modal object detection in improving the safety and reliability of autonomous systems

***Index Terms*—Autonomous System, Multi-modal,object detection, sensor fusion.**

I. Introduction

Autonomous systems, such as self-driving cars and drones, have the potential to revolutionize transportation and other industries. However, a key challenge in the development of these systems is ensuring their safety. One important aspect of safety in autonomous systems is the ability to accurately detect and classify objects in the environment. A literature survey of safety-critical multi-modal object detection in autonomous systems is necessary to understand the current state of the art and identify potential areas for improvement. This paper aims to provide a comprehensive review of the literature on this topic, covering both the technical challenges and the approaches used to address them. By understanding the current state of the field and the key challenges that remain, we can better guide the development of future autonomous systems and ensure their safe operation.

Autonomous vehicles, for example, self-driving cars, etc are designed to operate without human input to control their movement. They use a combination of sensors and artificial intelligence (AI) to navigate and make decisions about their environment. The level of autonomy in a self-driving car can range from no automation at all, to fully autonomous operation under all driving conditions [2], [3]. Fig. 1 [1, Fig 1] illustrates the use of sensors on the vehicle that collect data from the surrounding and then information is processed for Object detection, Decision making, etc

Fig. 1. An autonomous system with different Sensors. [1, Fig. 1]

According to the Society of Automotive Engineers(SAE), the level of autonomy can be classified into 6 levels as illustrated in Fig.2 [2, Fig. 2]. "The transfer of total control from humans to machines is classified by the Society of Automotive Engineers (SAE) as a stepwise process on a scale from 0 to 5, where 0 involves no automation and 5 means full-time performance by an automated driving system of all driving aspects, under all roadway and environmental conditions.", according to SAE's definition of autonomy [2], [3].

The five levels of autonomy defined by SAE are Level 0, Level 1, Level 2, Level 3, Level 4, and Level 5. Level 0 refers to no automation, where the driver is in complete control of the vehicle and all its systems at all times. Level 1 refers to driver assistance, where the vehicle has one or more systems that assist the driver with specific functions, such as braking or steering, but the driver is still responsible for monitoring the environment and controlling the vehicle. Level 2 refers to partial automation, where the vehicle has multiple systems that work together to assist the driver with multiple functions, such as braking, steering, and acceleration, and the driver is still responsible for monitoring the environment and is expected to take over control of the vehicle when necessary. Level 3 refers to conditional automation, where the vehicle is capable of performing all driving functions under certain conditions, such as on a highway, but the driver must still be prepared to take

over control of the vehicle when necessary. Level 4 refers to high automation, where the vehicle is capable of performing all driving functions and can handle all situations that may arise, but the driver may still have the option to take over control if they choose to do so. Level 5 refers to full automation, where the vehicle is capable of performing all driving functions and can handle all situations that may arise without any human interaction [2], [3]. In the next sections, The author will talk about sensors, that are vital for autonomous navigation, object detection, and avoidance.

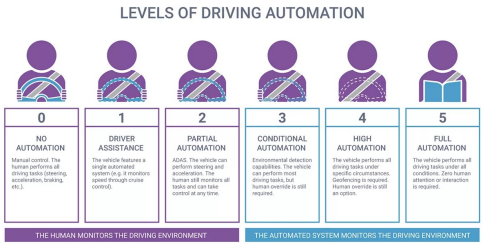


Fig. 2. Level of autonomy in Self-driving cars by SAE [2], [3, Fig. 2]

II. Modalities used in Autonomous Vehicles

Autonomous vehicles use a variety of sensors and modalities to navigate and understand their environment. These can include lidar, radar, cameras, ultrasonic sensors, and GPS.

A. Lidar

Lidar (Light Detection and Ranging) is a powerful sensing technology that has gained significant attention in recent years, particularly in the field of autonomous systems. Lidar sensors emit laser beams and measure the time it takes for the light to reflect back to the sensor, providing high-resolution range information. In this section, the author aims to provide an overview of the different types of lidar sensors, their benefits, and their usability in various applications [4]. An image representation of lidar can be seen in Fig.3 [2, Fig. 3].

One of the most commonly used types of lidar sensors is mechanical lidar. This type of lidar uses a spinning mirror to scan the environment and is capable of capturing a high-resolution 3D point cloud. Mechanical lidar sensors are known for their high accuracy and long range, making them well-suited for applications that require high precision such as self-driving cars, industrial robotics, and mapping [4].

Another type of lidar sensor is solid-state lidar. This type of lidar uses a static mirror and a detector array to scan the environment and is capable of capturing a high-resolution 3D point cloud. Solid-state lidar sensors are known for their compact size and low power consumption, making them well-suited for applications where size and power consumption are important, such as in drones and mobile robots [4].

Flash lidar is another type of lidar sensor, it uses a high-powered laser to illuminate a large area of the environment at once, and is capable of capturing a lower-resolution 3D point cloud. Flash lidar sensors are known for their wide field of view and fast measurement rate, making them well suited for applications that require a wide field of view and fast measurement rates such as obstacle detection and avoidance [4].

Frequency-Modulated Continuous-Wave (FMCW) Lidar is another type of lidar sensor, it uses a laser that is modulated in frequency to measure the distance and velocity of objects. They are relatively low cost and can operate at long ranges and in high-light conditions, making them well suited for outdoor monitoring and surveillance, and Velocimetry [4], [5].

Overall, the Lidar sensor is a powerful sensing technology that provides high-resolution range information, it is useful in many applications including autonomous navigation, environment perception, mapping, and object recognition. The different types of lidar sensors, such as mechanical, solid-state, flash and FMCW lidar, each have their own unique advantages and disadvantages and are suitable for different applications. The choice of the appropriate lidar sensor for a specific application will depend on the requirements of the system in terms of accuracy, range, the field of view, power consumption, and cost [4], [5].

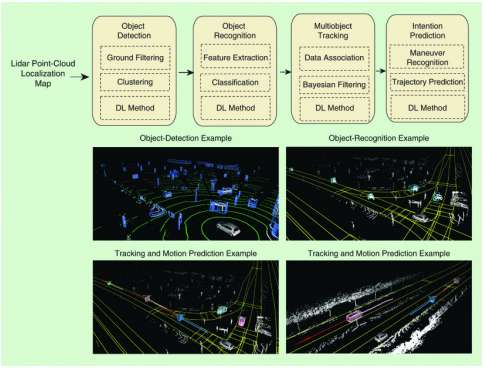


Fig. 3. Mechanical lidar and it's processing [6, Fig. 3]

B. Radar

Radar (Radio Detection and Ranging) is a sensing technology that uses radio waves to detect and locate objects, and to measure the distance, speed, and other characteristics of that objects [4]. An illustration of a Radar can be seen in Fig.4 [5, Fig. 4]. RADAR in autonomous vehicles operates at the frequencies of 24, 74, 77, and 79 GHz, corresponding to short-range radars (SRR), medium-range radars (MRR), and long-range radars (LRR), respectively. They each have slightly different functions:

SRR technology enables blind-spot monitoring, lane-keeping assistance, and parking assistance in autonomous vehicles [4], [5]. MRR sensors are used when obstacle detection is in the range of 100-150 meters with a beam angle varying between 30° to 160° [4], [5]. The automatic distance control and brake assistance are supported by LRR radar sensors [4], [5]. RADAR technology in autonomous vehicles operates with millimeter waves and offers millimeter precision. The utilization of millimeter waves in autonomous vehicular RADAR ensures high resolution in obstacle detection and centimeter accuracy in position and movement determination. Compared to other sensor technologies in autonomous vehicles, RADAR

Fig. 4. Radar and it's processing [5, Fig. 4]

works reliably under low visibility conditions such as cloudy weather, snow, rain, and fog [4], [5].

There are two types of RADAR used in autonomous vehicles.

Impulse RADAR - In impulse RADAR, one pulse is emitted from the device and the frequency of the signal remains constant throughout the operation [4], [5].

Frequency-modulated continuous wave (FMCW) RADAR - In FMCW RADAR, pulses are emitted continually. Pulses are modulated over the entire operation and the frequency varies over the transmission time [4], [5].

In general, radar is a powerful sensing technology that provides range and velocity information, it is useful in many applications including autonomous navigation, environment perception, mapping, and object recognition. The ability of radar to sense objects in 3D and in any weather conditions make it a valuable sensor for object detection, tracking, and classification in different environments.

Fig. 5. Camera and it's processing [7, Fig. 5]

C. Camera

In an autonomous vehicle, a camera is a sensor used to gather visual information about the vehicle's environment. This information is then processed by the vehicle's onboard computer to make decisions and control the vehicle's movements. Autonomous vehicles typically use multiple cameras, each with different capabilities, to provide a comprehensive view of the environment [4], [7]. An illustration of a camera can be seen in Fig. 5 [7, Fig. 5].

Some examples of cameras that may be used in autonomous vehicles include:

Stereo cameras: These cameras have two lenses that capture images at slightly different angles, providing depth information similar to human vision. They are commonly used in autonomous vehicles for object detection and obstacle avoidance [4], [7].

Thermal cameras: These cameras detect infrared radiation and can be used in low-light or no-light conditions. They are commonly used in autonomous vehicles for night vision and in industrial applications for temperature sensing [4], [7].

Monocular cameras: These cameras capture a single image and are often used in combination with other sensors, such as LIDAR, to provide visual input for the autonomous system [4], [7].

RGB-D cameras: These cameras combine the traditional RGB image capture with a depth sensor, providing both color and depth information in a single image [4], [7].

These cameras are mounted in various positions on the vehicle, such as on the front, sides, and rear, to provide a comprehensive view of the vehicle's surroundings. The information from the cameras is used by the vehicle's control system to make decisions, such as steering, braking, and accelerating, and avoid obstacles and other vehicles on the road.

In the next section, The authors discuss several techniques and algorithms used in autonomous systems to enable the vehicle to sense and understand its environment, make decisions, and control its movement.

III. Technique and Algorithms used in Autonomous Vehicles.

Autonomous systems use a combination of techniques and algorithms to sense, understand, and respond to their environment.

A. Fusion

There are two main approaches to sensor fusion: late fusion and early fusion. Late fusion involves combining the data from multiple sensors after it has been processed separately, while early fusion involves combining the data before it is processed. Both approaches have their own strengths and weaknesses, and the decision of which one to use depends on the specific needs of the autonomous system.

B. Late Fusion

In late fusion, the data from each sensor is processed separately before being combined as shown in Fig. 6 [8, Fig. 6]. This allows for the use of specialized algorithms that are tailored to each individual sensor, which can lead to improved accuracy. However, it can also be more complex and time-consuming to implement, as the data from each sensor must be processed and analyzed separately [8]. For example, if a

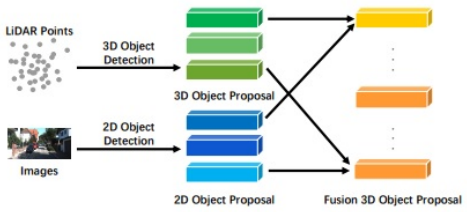


Fig. 6. An Example of Late fusion [8, Fig. 6]

LIDAR sensor, a camera sensor, and a radar sensor are used for late fusion, each sensor can be processed using the most appropriate algorithm for that sensor type. This can lead to improved performance, as each sensor can be processed using algorithms that are optimized for that specific sensor. Another benefit of late fusion is that it allows for the handling of different measurement errors and noise models of the sensors, as the sensor data is processed individually before being combined. This can lead to improved accuracy and robustness of the fused data.

Additionally, Late Fusion can also lead to improved performance in certain situations, for example, if the sensor data is fused at the decision level, it can lead to an improved decision-making process, as the sensor data is fused based on the specific decision needs.

C. Early Fusion

In contrast to late fusion, early fusion involves combining the data from multiple sensors before it is processed as shown in Fig. 7 [8, Fig. 7]. This can be faster and more efficient than late fusion, as the data does not need to be processed separately for each sensor. However, it also requires the use of more general-purpose algorithms that are not tailored to specific sensors, which can lead to reduced accuracy [8]. One of the

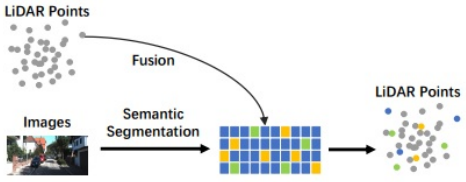


Fig. 7. An Example of Early fusion [8, Fig. 7]

main advantages of early fusion is that it allows for faster decision-making in time-sensitive situations. By combining the data from multiple sensors at once, an autonomous system can make quicker and more informed decisions about its environment. This can be especially useful in situations where speed is of the utmost importance, such as when a self-driving car needs to avoid a collision [8].

One of the main drawbacks of early fusion is that it can lead to loss of information and reduced accuracy if the sensor data is not properly aligned or if the sensors have different resolutions or scales. Additionally, early fusion can be computationally intensive and can lead to problems if the sensors have different measurement errors such as systematic errors, bias or noise that are not properly handled. [8]. For example, if a LIDAR sensor and a camera sensor are used for early fusion, the LIDAR sensor may provide a high-resolution, 3D point cloud of the environment, while the camera sensor may provide a lower resolution 2D image of the same environment. If the two sensor data sets are not properly aligned, the information from the LIDAR sensor may not be accurately represented in the final fused data, leading to a loss of information and reduced accuracy [8].

Next section talks about different datasets used for object detection tasks and also how many images were collected, what modality is used for the collection of data, etc.

IV. Datasets used for the research on Autonomous Vehicle

Table 1 explores different datasets used in the related research work that has been done in the field of Object detection and 3D depth estimation for Autonomous systems. The author have compared various parameters such as Dataset, Number of Images, Modality used, Annotations, etc. The table is intended to give a general idea of the number of images and modality, technical detail, annotation, resolution, data format that each dataset provides, and the benchmark that each dataset is used for. The information in the table is intended to help researchers understand the characteristics of different datasets and help them choose the dataset that best suits their needs for their research.

V. Overview

Table 2 examines the related research work that has been done in the field of Object detection and 3D depth estimation for Autonomous systems. The authors have discussed various parameters such as Dataset, Accuracy, Evaluation methods, Fusion, etc, and tried to show how each paper has taken an approach. The paper [16] uses the commonly used dataset that is nuScenes, Kitti data, and also synthetic data which are obtained from simulation. The modalities which are taken into consideration are Camera and Radar. The accuracy and evaluation metric used are Absolute and Squared relative Differences. The paper [17] used only synthetic data and Radar , camera for its modalities. The loss is used as evaluation metric. The paper [18] used nuScenes and modalities such as camera and radar. The evaluation metric used Average precision. The papers from [16] to [19] in the table II uses state of the art Deep learning techniques in their paper and most of the paper in the table II also used Early fusion techniques. This show how is the trend in the papers. The author by using this table II intends to inform researchers about recent trends into multi object detection about the datasets,

TABLE I
DATASETS USED FOR THE RESEARCH ON AUTONOMOUS VEHICLE

Dataset	No. of Images	Modality Used	Annnotation	Resolution of Data	Data Format	Benchmark
KITTI [9]	7481	LIDAR,RGB,& Camera	3D object detection on real image , semantic segmentation, and stereo estimation	LIDAR: 64x1024, Camera: 1242x375	.bin, .png, .txt	Object Detection, Scene Flow, and Semantic Segmentation
nuScenes [10]	1 million	LIDAR,Camera, Radar,RGB, GPS,& IMU	3D object detection on real image, Semantic segmentation, object tracking	LIDAR: 64x1024, Camera: 1242x375	.bin, .png, .txt	Semantic Segmentation and Object Detection
DDAD [11]	10K	LIDAR,RGB, Camera.	Semantic segmentation,3D object detection on a real image and stereo estimation	LIDAR: 64x1024, Camera: 1242x375	.bin, .png, .txt	Object Detection, Scene Flow, and Semantic Segmentation
ApolloScape [12]	n/a	LIDAR,Camera, Radar, RGB, HD map	Semantic segmentation, instance segmentation, lane detection	LIDAR:64x1024, Camera:1920x1080	.bin, .png, .json	Object Detection, Semantic Segmentation, Lane Detection
Berkeley DeepDrive [13]	100K	RGB,& Camera	Object Detection, semantic segmentation, instance segmentation	Camera: 1920x1080	.jpg, .xml	Object Detection, Semantic Segmentation, Instance Segmentation
Google-Landmarks [14]	5 Million	RGB,& Camera	Landmark recognition	Camera: various resolution	.jpg	Landmark Recognition
Waymo Open [15]	n/a	LIDAR, Camera, Radar, RGB, GPS, IMU	3D object detection, semantic segmentation, object tracking	LIDAR: 64x1024, Camera: 1920x1080	.tfrecord	Scene Understanding, Object Detection, and Object Tracking

techniques, and results which can be used by other researchers to improve on the existing papers.

VI. LITERATURE REVIEW

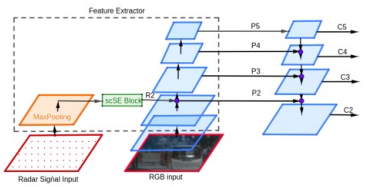


Fig. 8. Radar+RGB feature map fusion. [18, Fig.8]

The two papers [18] and [21] uses Radar and camera to perform object detection in an autonomous vehicle. The authors argue that while radar sensors provide robust object detection in adverse weather conditions and at long ranges, they often produce inaccurate object detections when the object is occluded or has a low radar cross-section. On the other hand, RGB cameras provide accurate object detections but are sensitive to lighting and weather conditions [18].

To address this issue, the proposed method for robust object detection in autonomous vehicles using a fusion of radar and RGB data works by using two parallel networks to extract features from the radar and RGB modalities, respectively. The features from both modalities are then passed through an attention mechanism that uses a probabilistic model to estimate the reliability of the features from the radar and RGB modalities.

The probabilistic model takes into account various factors such as the object's size, shape, radar cross-section, lighting, weather conditions, and object occlusion to estimate the reliability of the features from the radar and RGB modalities. These reliability estimates are then used to weight the features from the two modalities.

The weighted features from the radar and RGB modalities are then concatenated and passed through a fully connected layer. The fully connected layer is followed by a series of convolutional layers which act as the fusion module. The output from the final convolutional layer is passed through a

TABLE II
SALIENT CHARACTERISTIC OF RESEARCHED PAPERS

References	Dataset and and modalities used	Accuracy and Evaluation metric	Fusion method	Neural network based
[16]	nuScenes,KITTI&Synthetic ,Camera & Radar	Absolute Relative Distance **=0.102** , Squared Relative Difference $= 0.374$, Root Mean Square Error $= 2.41$	Early fusion	Yes
[17]	Synthetic, Camera & Radar	$8.2 * 10^{-}5$, Binary cross-entropy & a Dice loss.	Early Fusion	Yes
[18]	nuScenes , Camera and Radar	65.7% for resolution of 1024*1024 & 56.2% for resolution 512*512, Average Precision (AP) & mean Average Recall (AR).	Early Fusion	Yes
[20]	nuScenes, Camera and Radar	43.0 %, Average Precision (AP) and mean Average Recall (AR).	Early Fusion	Yes
[21]	nuScenes & Synthetic , Camera and Radar	55.9 %, Mean Average Precision(mAP).	Early Fusion	Yes
[19]	Synthetic, Camera, lidar & Radar	Radar+Camera 61.0%&Lidar+Camera 41.0%,Average Precision (AP)	Late Fusion	Yes
[22]	Synthetic, Motion stereo Camera & Radar	Accuracy for Stationary Vehicle in meters $= \pm 0.03$, Accuracy for Moving Vehicle in meters $= \pm 0.02$ Accuracy for Pedestrian in meters $= \pm 0.08$.	Early Fusion	No
[23]	Synthetic, Camera and Radar	n/a.	Early Fusion	No
[24]	Synthetic, Camera	Average precision= 93 % and Average Recall= 94 %.	No Fusion	No

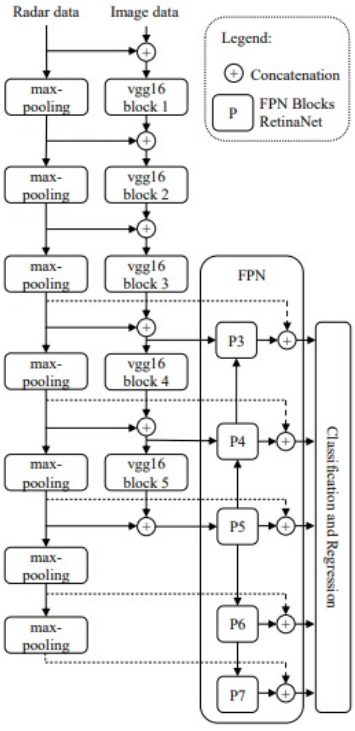


Fig. 9. Structure of CameraRadarFusionNet (CRFNet). [21, Fig. 9]

sigmoid activation function to produce a final object detection. detection that can be seen in Fig. 8 [18, Fig. 8]

Now for the second paper, the authors [21] propose a sensor fusion architecture that combines the strengths of both radar and camera data. The architecture consists of three main parts:

A radar data pre-processing module that converts the radar data into a format suitable for deep learning. A deep learning-based object detection module that processes the radar and camera data [21]. A fusion module that combines the object detections from the radar and camera data to produce a final object detection [21]. The radar data pre-processing module converts the radar data into a format suitable for deep learning by applying range compression and azimuth compression. The compressed radar data is then passed through a deep learning-based object detection module, which processes the radar data and camera data simultaneously [21].

The deep learning-based object detection module is a convolutional neural network (CNN) that processes the radar and camera data and produces object detections. The CNN is trained using a dataset that includes both radar and camera data [21].

The output of the object detection module is a set of object detections from both radar and camera data. The object detections are then passed through a fusion module, which combines them to produce a final object detection as shown

in Fig. 9 [21, Fig. 9] The radar and camera detections are fused by a decision-making algorithm, this is done with a Multi-Object Tracking (MOT) algorithm that assigns a weight to each detection based on the confidence of the detector and the spatial proximity of the detection to other detections. The weights are used to generate a final detection output that combines the information from both sensors.

A. Architecture used

The authors of the paper had experimented with radar and RGB data fusion at multiple levels of the feature extractor, the proposed architecture (BIRANet) shown in Fig. 11 [18, Fig. 11] proved to be better as compared to other configuration [18].

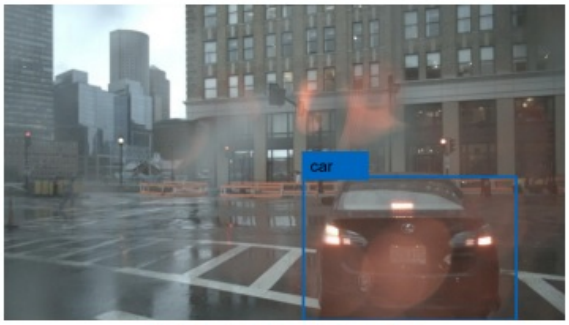
The paper [21] architecture builds on RetinaNet as implemented with a VGG backbone as shown in Fig. 10 [21, Fig. 10] The architecture automatically learns at which stage the fusion of the sensor data is most helpful for the detection, also BlackIn, a training strategy which takes inspiration from Dropout, which focuses the learning on a specific sensor type. Both the paper [18] and [21] have used the Early data fusion technique and Convolution Neural Network.

B. DataSet used and Implementation of the Network

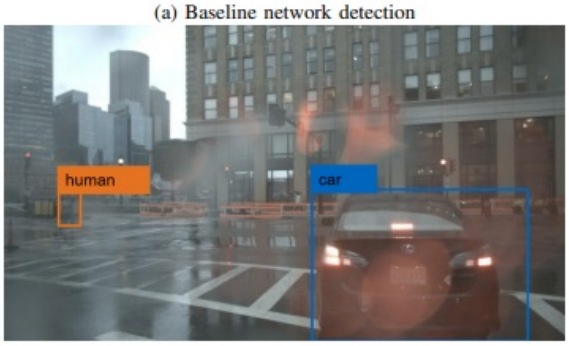
Both the paper [18] and [21] use the nuScenes dataset but [21] also used a synthetic dataset. The paper [18] is trained on one TITAN Pascal GPU with a batch size of two and the learning rate is set at 0.001. Experiments are conducted on images of sizes 1024x1024 and 512x512. The scale used for anchor generation is (16, 32, 64,128, 256) and (8, 16, 32, 64, 128) respectively. The total number of objects per image is set to 100. The paper [21] is trained and evaluations are performed with an Intel Xeon Silver 4112 CPU, 96GB RAM, and an NVIDIA Titan XP GPU.The nuScenes images at an input size of 360 x 640 pixels. The fish-eye images of the dataset are processed at 720 x 1280 pixel resolution.

C. Evaluation and Results

For evaluating results, both the papers [18]and [21] use Average Precision (AP) and mean Average Precision (mAP) respectively. The paper [18] gives an accuracy of 71.9 percent (AP) as shown in Fig. 11 [18, Fig. 11] on 1024x1024 input image size and an accuracy of 67.4 percent(AP) on 512x512 input image size. Whereas the paper [21] has an accuracy of 43.95 percent on nuScenes dataset as shown in the 10 [21, Fig. 10] and 57.50 percent on custom dataset.



(a) Baseline network detection



(b) CRF-Net detection

Fig. 10. Results from CRF-net. [21, Fig. 10]

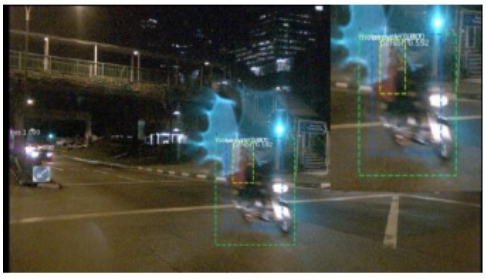


Fig. 11. Result from BiraNet. [18, Fig. 11]

VII. Open research problem and future direction

In this literature survey, we have reviewed the recent advances in safety-critical multi-modal object detection in autonomous systems. However, there are still several open research problems and future directions that need to be addressed in order to further improve the reliability and robustness of these systems.

One open research problem is the development of real-time and online multi-modal object detection algorithms that can handle the high volume and variety of data generated by autonomous systems. Another important direction is to develop scalable multi-modal object detection methods that can handle large and complex environments, such as urban environments and industrial settings. Improving the robustness of multi-modal object detection methods to deal with various challenges such as occlusions, changing lighting conditions, and different weather conditions is also an important research direction [25].

Another important future direction is the development of methods for making multi-modal object detection more ex-

plainable to human operators and other stakeholders. Additionally, more sophisticated and robust techniques for fusing data from different modalities, such as vision and lidar, to improve object detection performance is an important research direction. Furthermore, developing methods for formally verifying the safety of multi-modal object detection algorithms in autonomous systems is also a crucial future direction [25].

Finally, it is important to develop and test multi-modal object detection methods in different application domains such as self-driving cars, drones, and industrial robotics. Additionally, developing object detection methods that can run on resource-constrained systems, such as edge devices, to enable the deployment of autonomous systems in resource-constrained settings is also a crucial research direction [25].

Overall, these open research problems and future directions highlight the ongoing need for further research in the field of safety-critical multi-modal object detection in autonomous systems in order to improve the reliability and robustness of these systems.

VIII. Conclusion

In summary, multi-modal object detection is a crucial component of safety-critical autonomous systems, as it allows for improved accuracy and robustness in detecting and classifying objects in the environment. There have been numerous research efforts focused on developing multi-modal object detection systems using various combinations of sensors and modalities, including LiDAR, cameras, and radar. These approaches have demonstrated improved performance compared to using a single modality, and have the potential to enable the deployment of safer and more reliable autonomous systems.

References

[1] S. Kato, E. Takeuchi, Y. Ishiguro, Y. Ninomiya, K. Takeda, and T. Hamada, "An open approach to autonomous vehicles," *IEEE Micro*, vol. 35, no. 6, pp. 60–68, 2015.

[2] A. C. Serban, E. Poll, and J. Visser, "A standard driven software architecture for fully autonomous vehicles," in *2018 IEEE International Conference on Software Architecture Companion (ICSA-C)*, 2018, pp. 120–127.

[3] (no author), "Taxonomy and definitions for terms related to driving automation systems for on-road motor vehicles," https://www.sae.org/standards/content/j3016_202104, (Accessed: 2022-07-20).

[4] J. Kocić, N. Jovičić, and V. Drndarević, "Sensors and sensor fusion in autonomous vehicles," in *2018 26th Telecommunications Forum (TELFOR)*, 2018, pp. 420–425.

[5] L. Arnone and P. Vicari, "Simultaneous odometry, mapping and object tracking with a compact automotive radar," in *2019 AEIT International Conference of Electrical and Electronic Technologies for Automotive (AEIT AUTOMOTIVE)*, 2019, pp. 1–6.

[6] Y. Li and J. Ibanez-Guzman, "Lidar for autonomous driving: The principles, challenges, and trends for automotive lidar and perception systems," *IEEE Signal Processing Magazine*, vol. 37, no. 4, pp. 50–61, 2020.

[7] D. Zhu, Z. Xu, J. Dong, C. Ye, Y. Hu, H. Su, Z. Liu, and G. Chen, "Neuromorphic visual odometry system for intelligent vehicle application with bio-inspired vision sensor," in *2019 IEEE International Conference on Robotics and Biomimetics (ROBIO)*, 2019, pp. 2225–2232.

[8] K. Huang, B. Shi, X. Li, X. Li, S. Huang, and Y. Li, "Multi-modal sensor fusion for auto driving perception: A survey," *arXiv preprint arXiv:2202.02703*, 2022.

[9] A. Geiger, P. Lenz, C. Stiller, and R. Urtasun, "Vision meets robotics: The kitti dataset," *The International Journal of Robotics Research*, vol. 32, no. 11, pp. 1231–1237, 2013.

[10] 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 *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 11 621–11 631.

[11] V. Guizilini, R. Ambrus, S. Pillai, A. Raventos, and A. Gaidon, "3d packing for self-supervised monocular depth estimation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 2485–2494.

[12] X. Huang, P. Wang, X. Cheng, D. Zhou, Q. Geng, and R. Yang, "The ApolloScape open dataset for autonomous driving and its application," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 42, no. 10, pp. 2702–2719, oct 2020. [Online]. Available: https://doi.org/10.1109%2Ftpami.2019.2926463

[13] F. Wu, D. Wang, M. Hwang, C. Hao, J. Lu, J. Zhang, C. Chou, T. Darrell, and A. Bayen, "Decentralized vehicle coordination: The berkeley deepdrive drone dataset," 2022. [Online]. Available: https://arxiv.org/abs/2209.08763

[14] T. Weyand, A. Araujo, B. Cao, and J. Sim, "Google landmarks dataset v2 – a large-scale benchmark for instance-level recognition and retrieval," 2020. [Online]. Available: https://arxiv.org/abs/2004.01804

[15] P. Sun, H. Kretzschmar, X. Dotiwalla, A. Chouard, V. Patnaik, P. Tsui, J. Guo, Y. Zhou, Y. Chai, B. Caine, V. Vasudevan, W. Han, J. Ngiam, H. Zhao, A. Timofeev, S. Ettinger, M. Krivokon, A. Gao, A. Joshi, S. Zhao, S. Cheng, Y. Zhang, J. Shlens, Z. Chen, and D. Anguelov, "Scalability in perception for autonomous driving: Waymo open dataset," 2019. [Online]. Available: https://arxiv.org/abs/1912.04838

[16] S. A. Siddiqui, A. Vierling, and K. Berns, "Multi-modal depth estimation using convolutional neural networks," in *2020 IEEE International Symposium on Safety, Security, and Rescue Robotics (SSRR)*. IEEE, 2020, pp. 354–359.

[17] G. Zhang, H. Li, and F. Wenger, "Object detection and 3d estimation via an fmcw radar using a fully convolutional network," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 4487–4491.

[18] R. Yadav, A. Vierling, and K. Berns, "Radar+ rgb attentive fusion for robust object detection in autonomous vehicles," *arXiv preprint arXiv:2008.13642*, 2020.

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

[20] 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)*. IEEE, 2019, pp. 3093–3097.

[21] 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)*. IEEE, 2019, pp. 1–7.

[22] T. Kato, Y. Ninomiya, and I. Masaki, "An obstacle detection method by fusion of radar and motion stereo," *IEEE transactions on intelligent transportation systems*, vol. 3, no. 3, pp. 182–188, 2002.

[23] R. O. Chavez-Garcia, J. Burlet, T.-D. Vu, and O. Aycard, "Frontal object perception using radar and mono-vision," in *2012 IEEE Intelligent Vehicles Symposium*. IEEE, 2012, pp. 159–164.

[24] D. Neumann, T. Langner, F. Ulbrich, D. Spitta, and D. Goehring, "Online vehicle detection using haar-like, lbp and hog feature based image classifiers with stereo vision preselection," in *2017 IEEE Intelligent Vehicles Symposium (IV)*. IEEE, 2017, pp. 773–778.

[25] D. Feng, C. Haase-Schütz, L. Rosenbaum, H. Hertlein, C. Glaeser, F. Timm, W. Wiesbeck, and K. Dietmayer, "Deep multi-modal object detection and semantic segmentation for autonomous driving: Datasets, methods, and challenges," *IEEE Transactions on Intelligent Transportation Systems*, vol. 22, no. 3, pp. 1341–1360, 2020.