The following publication M. Nawaz, J. K. -T. Tang, K. Bibi, S. Xiao, H. -P. Ho and W. Yuan, "Robust Cognitive Capability in Autonomous Driving Using Sensor Fusion Techniques: A Survey," in IEEE Transactions on Intelligent Transportation Systems, vol. 25, no. 5, pp. 3228-3243, May 2024 is available at https://doi.org/10.1109/TITS.2023.3327949.

Robust Cognitive Capability in Autonomous Driving using Sensor Fusion Techniques: A Survey

Mehmood Nawaz, Jeff Kai-Tai Tang *Senior Member, IEEE,* Khadija Bibi, Shunli Xiao, Ho-Pui Ho* *Senior Member, IEEE*, and Wu Yuan* *Senior Member, IEEE*

***Abstract*—Autonomous driving has become a prominent topic with the rise of intelligent urban vision in communities. Advancements in automated driving technology play a significant role in the intelligent transportation system. Autonomous vehicles (AVs) rely heavily on sensor technologies as they are responsible for navigating safely through their environment and avoiding obstacles. This paper aims to outline the vital role of sensor fusion in intelligent transportation systems. Sensor fusion is the process of combining data from multiple sensors to obtain more comprehensive measurements and greater cognitive abilities than a single sensor could achieve. By merging data from different sensors, it ensures that driving decisions are based on reliable data, with improved accuracy, reliability, and robustness in AVs. This paper provides a comprehensive review of AV capacity, impacts, planning, technological challenges, and omitted concerns. We used state-of-the-art evaluation tools to check the performance of different sensor fusion algorithms in AVs. This paper will help us to determine our position, direction, the impacts of AVs on society, the need for smart city mobility outcomes, and the way to solve the auto industry challenges in the future. The analysis of AV systems from the perspective of sensor fusion in this research is expected to be beneficial to current and future researchers.**

***Index Terms*—Sensor fusion, Autonomous vehicles, RGB cameras, LiDAR points, Radar points, Object detection, and Object tracking.**

I. Introduction

AUTONOMOUS driving is a significant disruptive innovation for the future. It is expected to have a substantial societal influence in a digital transportation system. A summary of AV technology and development is provided in this section to fulfill the customers requirements. According to an expert (Marlon G. Boarnet) of transportation and urban growth [1] at the University of Southern California says; “roughly every two generations, we should reinvent the transportation infrastructure in our cities in ways that affect the viability of our communities, our economy, our society, our culture and the settlement patterns in our cities and countryside.” We believe AV can significantly change everyone’s life.

AV system has the potential to have a remarkable influence on environmental benefits, such as lower fuel consumption (platoon driving might save fuel consumption up to 30%) [1], the highway improvement [2], [3], and a 15% decrease in the number of automobiles needed [1]. However, according to John Leonard [4], an MIT professor says: present technology relies on highly accurate prior maps, which should be kept up to date instead of changing over time. Another research scientist at MIT’s Age Lab (Bryan Reimer) believes that the most significant limiting variables in autonomous driving are ”factors connected to the human experience.” In reality, driving simulators have been used for full automation trials with prospective users [1].

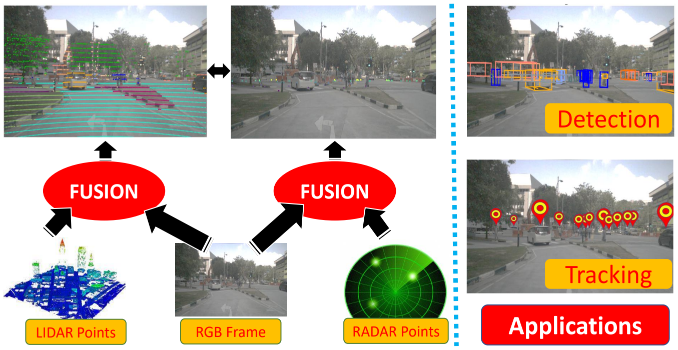


Fig. 1: Example of sensors fusion that used Lidar points, RGB, and radar data. The right column images show the sensor fusion applications.

According to the US traffic department [5], 22% of vehicle crashes occur due to adverse weather (i.e., rain, snow, fog, sleet, severe crosswinds, etc.). Some Asian countries are also facing smog problems in winter (Pakistan, India, Bangladesh, etc.). In adverse weather, AVs need different sensors for navigation and detection purposes. These sensors interact with their surroundings like human senses (hearing, vision, taste, smell, and touch). AVs can benefit from the strengths of several sensors and can make a comprehensive sensor system in backup (in case one sensor fails, another will operate). The combination of different sensor domains in AV is known as sensor fusion, which enables automated vehicles to detect and recognize the surrounding objects in real time, as shown in Fig. 1. Sensor fusion can help self-driving vehicles better perceive and respond to their environment.

Several different approaches have been proposed for sensor fusion in AVs. These approaches produced encouraging results on commonly available data sets with favorable weather conditions. A Bayesian-based sensor fusion method was proposed by Ratheesh et al [6]. This research used camera, LiDAR,

Mehmood Nawaz, Khadija Bibi, Wu Yuan, and Ho-Pui Ho are with the Department of Biomedical Engineering, The Chines University of Hong Kong, Hong Kong SAR, China. Email: mehmoodnawaz@cuhk.edu.hk.
Jeff K.T. Tang and Shunli Xiao are with the Automotive Platforms and Application Systems (APAS) RD Centre, Hong Kong Productivity Council, Hong Kong SAR, China.
Corresponding authors: aaron.ho@cuhk.edu.hk (Ho-Pui Ho) and wyuan@cuhk.edu.hk (Wu Yuan).

and RADAR sensors using Bayesian neural network (CLR-BNN) to enhance detection accuracy and minimize uncertainty in AV. Fuling et al. [7] proposed a Multi-sensor fusion method based on the estimated motion of nearby objects. This technique adopted the influence of optical flow tracking, the complementarity of front and rear visual-inertial odometry, and a bidirectional loopback strategy to boost the AV system resilience and adaptability.

Alfred et al. [8] developed a fully convolutional neural network for LIDAR-camera fusion in AVs for pedestrian detection. It combines Lidar data with several camera images to deliver the best pedestrian detection solution. Nouar et al. [9] developed a space object recognition approach that combines RGB and Depth maps using the CoAtNets network. It used the lite version of CoAtNet, which is known as CoAtNet0. This approach had previously been trained on the ImageNet data set and all of the parameters of all the layers were adjusted using the SPARK data set.

A. *Levels of autonomous vehicles*

There are six categories of self-driving vehicles, ranging from driver assistance to fully AVs. The Society of Automotive Engineers (SAE) established different levels of AVs, which are known as the SAE Levels of Automation Driving [10]. These levels vary according to the level of human engagement in the act of driving. Fig. 2 shows the six levels of the driving system (level-0, level-1, level-2, level-3, level-4, and level-5). The detail of each level is given below.

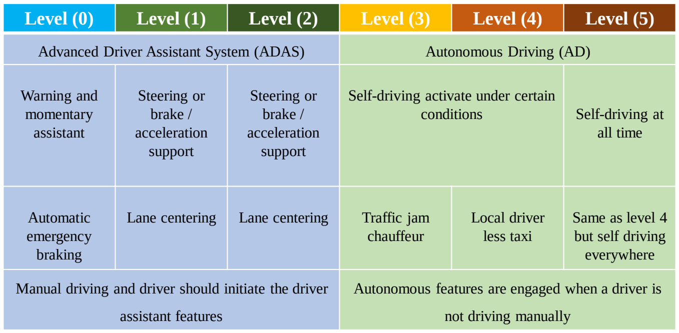
1) Level-0: This is a basic level in the driving system. It is managed by a human and has no automation function. In some vehicles, it has a warning and momentary assistance, which controls the emergency braking and blind spot warning features.

2) Level-1: At this level, a driving automation system in the vehicle supports the steering and acceleration, but not at the same time. A human driver controls all elements of vehicle operation at this level, including accelerating, steering, braking, and keeping an eye on the surroundings.

3) Level-2: At this level, the automation system helps the driver with steering and acceleration at the same time, but the majority of protection tasks and environmental monitoring are still in the hands of the driver. Level 2 AVs are now the most frequent on the roads.

4) Level-3: At this level, a simple vehicle uses AV sensors to monitor the surroundings and perform more dynamic driving tasks, such as automatic braking due to unexpected incidents that occur while driving. The warning alarm, front vehicle speed, and vehicle distance are also included in this level. The human driver must be prepared to react at this level.

5) Level-4: It shows a high level of automation, where the vehicle can complete a journey without the driver's help, even in difficult conditions. Furthermore, there are certain restrictions: the driver may only utilize this mode if the system determines that traffic conditions are safe and no traffic bottlenecks exist. Recently, the International Standardization Organization (ISO) introduced the new standard (ISO 22737:2021) that devised performance requirements, system requirements, and performance test procedures for low-speed automated driving (LSAD) systems for predefined routes [1]. This standard provides achievable requirements for deploying level 4 AVs on real roads.



Level (0)	Level (1)	Level (2)	Level (3)	Level (4)	Level (5)
Advanced Driver Assistant System (ADAS)			Autonomous Driving (AD)		
Warning and momentary assistant	Steering or brake / acceleration support	Steering or brake / acceleration support	Self-driving activate under certain conditions		Self-driving at all time
Automatic emergency braking	Lane centering	Lane centering	Traffic jam chauffeur	Local driver less taxi	Same as level 4 but self driving everywhere
Manual driving and driver should initiate the driver assistant features			Autonomous features are engaged when a driver is not driving manually		

Fig. 2: Six levels of autonomous vehicles devised in the SAE J3016 standard [10].

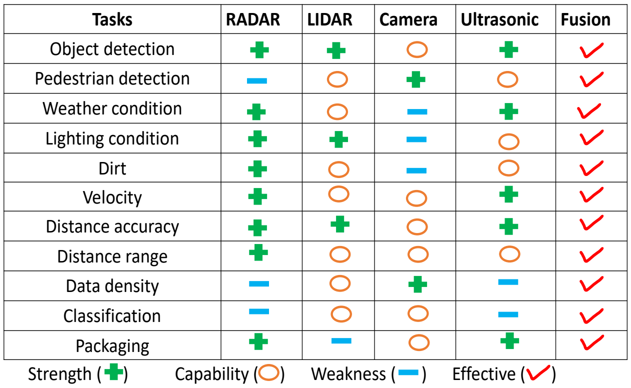
6) Level-5: This is a fully independent vehicle level in which the driver selects a destination, and the vehicle assumes all control and accountability for all driving modes. It does not exist yet, but manufacturers are working toward achieving level 5 AVs. All human controls, including steering wheels and pedals, will be removed on level-5 vehicles. This level has beautiful futures to attract AV lovers to the market. According to public statements made by automakers such as Ford, Honda, Toyota, Volvo, and others AVs are likely to hit the market around 2022–2023 [11]. Currently, partially autonomous levels 2 & 3 vehicles dominating the market. The overall registration share of AVs is expected to reach 12% globally by 2030 [1].

B. *Contributions*

The main contributions of this article are:

- **A fundamental analysis of sensors and sensors fusion:** A sensor fusion importance in AV and current challenges in sensor fusion.
- **A complete analysis of recent achievements of sensor fusion in the light of technical evaluations:** It includes the evaluation of different sensors on the base of advantages and disadvantages in AVs.
- **Challenges and in-depth assessment of important autonomous vehicle technologies:** It includes deep learning and traditional approaches for surrounding object detection, object recognition, and other vehicles movement detection using different sensors.
- **Discussions on present research status and prospective directions for future researcher:** It includes results comparison of state-of-the-art sensor fusion techniques. It also includes state-of-the-art assessment matrices to compare quantitative and qualitative aspects of different techniques.

The majority of the paper is structured as follows: Section II highlights the usage of several sensors in AVs. Section III explains sensor fusion types and object detection. Section IV highlights the state-of-the-art sensor fusion networks. Section V contains the evaluation metrics. Section VI shows the



Tasks	RADAR	LIDAR	Camera	Ultrasonic	Fusion
Object detection	✚	✚	○	✚	✓
Pedestrian detection	−	○	✚	○	✓
Weather condition	✚	○	−	✚	✓
Lighting condition	✚	✚	−	○	✓
Dirt	✚	○	−	○	✓
Velocity	✚	○	○	✚	✓
Distance accuracy	✚	✚	○	✚	✓
Distance range	✚	○	○	○	✓
Data density	−	○	✚	−	✓
Classification	−	○	○	−	✓
Packaging	✚	−	○	✚	✓

Strength (✚) Capability (○) Weakness (−) Effective (✓)

Fig. 3: Capability comparison of different sensors in autonomous vehicles.

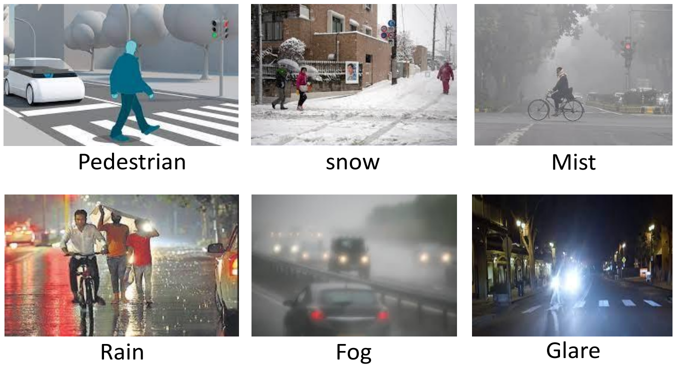


Fig. 4: Illustration of various challenging scenarios for Autonomous Vehicle cognitive system, including interaction with pedestrians, snow, mist, rain, fog, and glare.

performance comparison on different benchmarks. Section VII explains the discussion and future perspective. Limitations and conclusions are discussed in Section VIII.

II. Sensors used in Autonomous Vehicles

AVs use different types of sensors to navigate safely on the real road. They behave like people to interact with their surroundings. According to autonomous car technology engineers (ACTE), [10], self-driving vehicles have high-tech sensor systems that enable them to see roughly analogous. Some sensors in self-driving vehicles have overlapping and redundant functionality, as shown in Fig. 3.

A. Types of sensors:

In this section, we discussed four commonly used sensors in AV systems, delving into their respective strengths and weaknesses.

1) RGB camera: The RGB camera is used by different vehicles to sense their surrounding environment. In order to capture the entire 360° environment, it is necessary to deploy multiple cameras on AV. The RGB camera can record the internal and external views of the vehicles, which is very helpful for security purposes. The RGB camera images can be used by self-driving vehicles to see and understand the objects in their surroundings. The RGB cameras do not operate well in weather situations such as snowfall, fog, heavy rain, etc. The RGB cameras only record information that is visible to the human eye.

2) Ultrasonic sensor: This sensor uses high-frequency sound waves (like a bat) to estimate distances between obstacles. Ultrasonic sensors can be paired with LiDAR, radar, and RGB cameras to provide a comprehensive view of a vehicle's surroundings, as shown in Fig. 5. They are particularly adept at locations of poor visibility (e.g. bad weather). For AV systems, proximity, and moderate speeds are essential. Ultrasonic sensors have been used in vehicle parking for many years. The parking data is used to train deep-learning models for auto parking systems. The data gathered from millions of vehicles is a critical component of the automotive Internet of Things.

3) LiDAR: It is abbreviated for "light detection and ranging." It uses pulsed laser beams to measure the distance between objects. A large number of self-driving vehicles use it to navigate their surroundings. LiDAR's advantage lies in its capability of depth perception, which allows it to determine the distance between objects from up to 60 meters to within a few meters. It's also suitable for 3D mapping in self-driving technology to navigate the environment safely. Another advantage of LiDAR is that it covers a large field of view in 360°. In addition, it is cheaper than solid-state sensors and can measure an object's velocity and position in 3D space. The 3D clouds of points from LiDAR are significantly better at measuring distances than cameras on texture-less surfaces. For the same purpose, cameras require extensive computer algorithms to combine several cameras' views to determine the distance between objects, such as complicated neural networks [12].

4) Radar: It's a popular and essential perceptive sensor for AVs [13], as shown in Fig. 3. It is low cost and has a wide measurement range, an advanced capability on target recognition, and superior adaptability. It maintains the stability, security, and reliability of a vehicle. Compared to image sensors and LiDAR [14], radar is able to estimate the relative velocity of objects up to 250m away with a resolution of 0.1 m/s using the Doppler effect. It is a key element of safe autonomous driving and advanced driving-assistance systems (ADAS), due to its high performance and low cost. Furthermore, it works well in harsh environments, such as fog, smoke, and dust. It is highly adaptive to different illumination and weather conditions.

B. Importance of sensors in Autonomous Vehicles

The idea of a $'driverless'$ vehicle on the road has provoked the interest of a wide range of people. In autonomous driving, one sensor failure could be catastrophic. An AV is categorized into 6 levels, with the driver controlling all parts (brakes, throttle, steering, etc.) at the most basic level, and an ADAS system controlling the functions at a higher level. Hundreds of sensors and actuators are located in various parts of the vehicle and controlled by a complex system, which is divided into three parts.

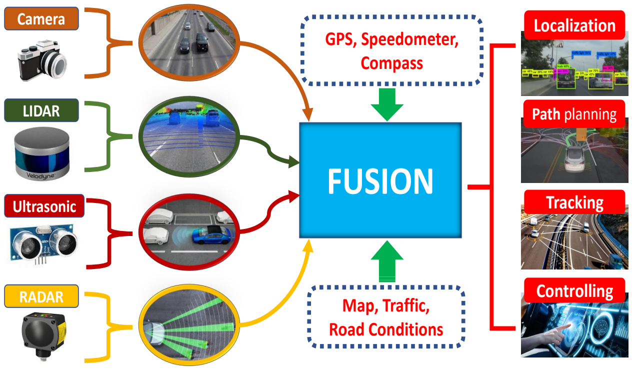


Fig. 5: Fusion of four different sensors like Camera, LiDAR, Radar, and Ultrasonic. The right side shows the application of sensor fusion.

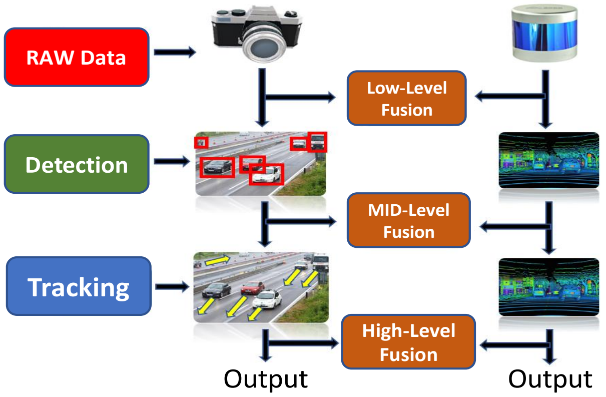


Fig. 6: Shows the sensor fusion by Abstraction Level. It has three sub-types (1) Low-level, (2) Middle-level, and High-level fusion.

(1) *Navigation and guidance*: The system that decides where you are, where you want to go, and how you get there is known as navigation and guidance. In this area, we use different instruments and software like Compass, Sextant, LORAN radio location, and dead reckoning to calculate the degrees of precision, consistency, and availability, as shown in Fig. 5.

(2) *Driving and Safety*: The safety of people is the main concern in AVs. The AV should behave appropriately in all worst conditions. The AV should be able to see the 360° view using different sensors. We can put an array of sensors; where one sensor determines the lane on the road and others detect the objects in 360° FOV.

(3) *Performance*: A large amount of the design of an AV and power management are common challenges. A conventional vehicle is converted into an autonomous functions vehicle with the help of specific applications and sub-systems. We need a proper reliable power system to regulate power management, total power consumption, and thermal dissipation in AV.

C. Challenges in Autonomous Vehicles

AV systems face several difficulties in devising comprehensive algorithms for sensor fusion in self-driving, as shown in Fig. 4. Research indicates self-driving vehicles have an increased risk of motion sickness [15], and passengers who have never driven feel uncomfortable at lower acceleration speeds due to the fear of hardware failure. AV technology brings social implications, e.g. mass employment losses and changes to health insurance companies and public transportation systems structures [16]. To tackle the various challenges, scientists, engineers, and problem solvers are working on light, and compatible AV algorithms [17]. Over the last decade, numerous research centers and industries have made substantial progress toward smarter autonomous driving systems [18].

(1) *Road conditions*: The condition of the roads might be unpredictable and vary drastically from one region to the next region. Autonomous driving systems require smooth and well-marked large highways [17].

(2) *Weather conditions*: Harsh weather poses a major challenge for AV driving systems, ranging from sunny days to rain, fog, smoke, and snowfall. Self-driving vehicles must be able to perform optimally in these conditions, with no margin for error.

(3) *Traffic conditions*: Obstacles and too many cars are common causes of traffic jams [16]. AVs should be constantly up-to-date on traffic conditions and be able to navigate varied traffic jams while sharing the road with other self-driving cars and interacting with many people. The traffic flow might be carefully managed and self-regulating. The velocity speed of AV makes a difference in crowded regions.

(4) *Accident liability*: In the case of an accident, AV liability is the most important factor to consider [17]. Who is accountable for incidents involving self-driving vehicles? In the case of self-driving vehicles, the software will be the major component that controls the vehicle and makes all of the important decisions. In this case, the suggestion is that a human should physically sit behind the steering wheel. Unfortunately, Google's following prototypes do not have a dashboard and a steering wheel. It is difficult to handle the vehicle in the case of accident, If there is no steering wheel, brake pedal, or accelerator pedal. How will the passengers feel relaxed by using the features of AVs or paying careful attention to road conditions?

(5) *Sensors interference*: For navigation, self-driving vehicles employ different types of sensors [18] like Lidar, RGB camera, radar, ultrasonic, etc. The lidars are mounted on the vehicle's roof, while the other sensors are mounted on the vehicle's front end and back end. Radar works by detecting radio wave that reflects from the surrounding objects. The reflection time from AV to object is used to compute the distance between the vehicles and the object. As a result, the AV driving system takes appropriate action based on the sensor data.

III. Sensor fusion types and object detection

A. Sensor fusion types

Sensor fusion is a crucial aspect of self-driving vehicles, which merges data from multiple sensors to detect, recognize, and locate objects [19]. It is widely used in the robotic industry, and automated industries for object detection, localization,

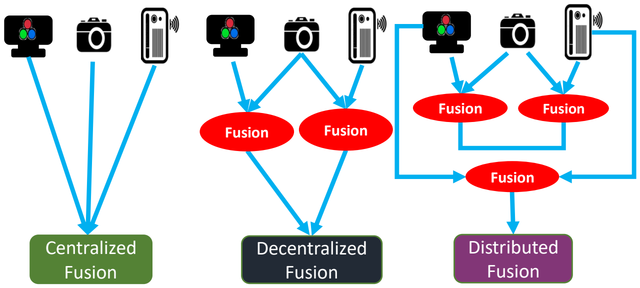


Fig. 7: Shows sensor fusion by Centralization Level. It has three sub-types (1) Centralized, (2) Decentralized, and (3) Distributed fusion.

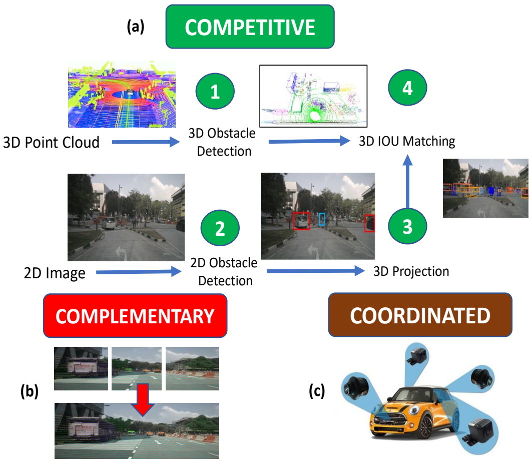


Fig. 8: Shows the sensor fusion by Competition Level. It has three sub-levels (a) Competitive, (b) Complementary, and (c) Coordinated Level.

path-planning, and tracking (Fig. 5). There are three main types of sensor fusion:

I - Sensor Fusion by Abstraction Level: It has three sub-types: low-level, middle-level, and high-level fusion, as shown in Fig. 6 [20]. Low-level fusion fuses raw data from multiple sensors (e.g. LiDARs and cameras) and has the potential for the future fast fusion of hundreds of thousands of points and pixels in milliseconds. Middle-level fusion combines independently recognized objects (e.g. obstacles detected by camera and radar) to get the best estimate of their position, class, and velocity. High-level fusion merges objects and their trajectories, but it may lose data if tracking is incorrect.

II - Sensor Fusion by Centralization Level: It has three subtypes: (1) centralized (one central unit handles fusion at a low level), (2) decentralized (each sensor fuses and forwards data), and (3) distributed (each sensor processes data locally and sends it to the next unit in late fusion), as shown in Fig. 7. Aptiv plugs many sensors and fuses them together into one central unit that handles the intelligence system called the Active Safety Domain Controller (ASDC) [20].

III - Sensor Fusion by Competition Level: It has three sub types; competitive, complementary, and coordinated fusion. (1) *Competitive Fusion*: When sensors are designed for the same goal, as shown in Fig. 8 (a). For example, when we use radar and LiDAR to detect the presence of a pedestrian. The data fusion process is referred to as redundancy in this case. (2) *Complementary Fusion*: It is the process of employing separate sensors to look at diverse situations in order to gather information that we wouldn't have been able to obtain otherwise. A panorama feature with many cameras is an example of complimentary fusion, as shown in Fig. 8 (b). We use the term $'complementary'$ because these sensors complement each other. (3) *Coordinated Fusion*: It combines data from two or more sensors to create a new scene that focuses on the same object, as shown in Fig. 8 (c). Executing 3D reconstruction or 3D scanning using 2D sensors is an example of coordinated fusion.

B. Sensor Fusion for 2D and 3D Object Detection

*1) **Sensor fusion for 2D object detection in AV:*** Sensor fusion has an important role in object detection [21, 22], object classification, object analysis, etc. The researcher used 2D cameras and 2D LiDAR points to detect objects on the roads. Hsu et al. [23] proposed a sensor fusion strategy for a collision avoidance system that combined a camera and a 2D LiDAR to achieve high object recognition accuracy using the pixel analysis technique. The empty area in front of cars is limited by objects on practically vertical surfaces, therefore pixel analysis is done with a single camera. Ryu et al.[24] presented a 2D object detection algorithm to recognize the object on the road. In this technique, two sensors (camera & LiDAR) are fused to detect objects. A camera sensor can work in low-light conditions, and a LiDAR sensor has great near-field dependability and can determine object location precisely.

Deng et al. [25] combined LiDAR and Camera sensors for multi-scale object identification, as shown in Fig. 9. Most of the object detection applications consist of cameras and LiDAR fusion. Both cameras and LiDAR have certain inherent flaws. As a result, integrating LiDAR and a camera is a logical way to overcome each sensory modality's inherent weaknesses. Multi-sensor fusion approaches use deep-learning methods, which have achieved outstanding detection results on large-scale objects like cars and buses. This is due to the advancement of deep learning-based approaches. Chellappa et al. [26] proposed a 2D technique to fuse the video and auditory sensors for vehicle detection and tracking. In this technique, an approximate estimate of the target direction-of-arrival (DOA) is used in the detection phase, which is based on a beam-forming algorithm. The approximate target position in the video is designated by this initial DOA estimate. The DOA is adjusted using video data and moving target detection given to the original target position. A Markov chain Monte Carlo method [27] is applied for integrated audio-visual tracking in this technique.

Silva et al. [28] proposed a sensor fusion technique that improves the navigation of visually impaired persons using everyday mobile devices, sensors, and cloud computing resources. In this technique, object detection and recognition are based on sensor data, including sonar, camera, LiDAR, and inertial sensors. This method includes grid-based obstacle localization using sonar sensors and accurate obstacle recogni-

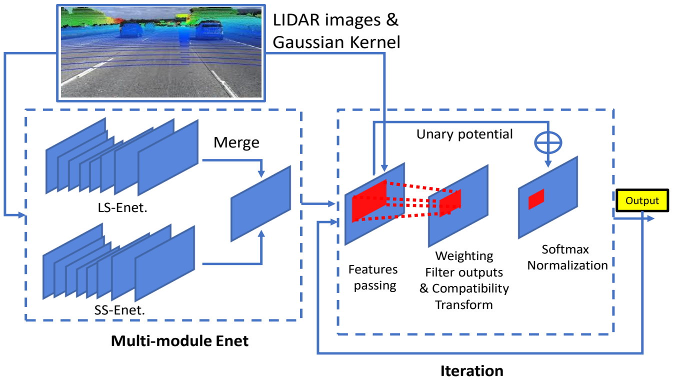


Fig. 9: Shows LiDAR and Camera fusion for multi-scale object recognition [25]. It has two basic parts (1) multi-module processing, and (2) integration of different features.

tion with camera sensors. Song et al. [29] proposed a reliable Vision-Based Relative-Localization technique using a LiDAR sensor and RGB-depth camera. A LiDAR and an RGB-D camera sensor are used to measure the three-dimensional (3-D) and two-dimensional (2-D) position information of a target in the proposed method. A low-level vision-LiDAR fusion algorithm, visual tracking approach, and depth information from the structured light sensor are employed to find a target object.

*2) **Sensor fusion for 3D object detection in AV:*** 3D object detection modality is increasing in the vehicle industry. The existing methods applied fusion methods on recognition data [30] and also successfully applied using advanced LiDAR device to detect the 3D object in AVs. The sensor fusion system proposed by Kim et al. [31] can distinguish numerous 3D objects using 2D projection images and tactile data. This system aims to increase the rate of object recognition. A tactile sensor is a device that captures data by interacting with its surroundings directly. Tactile sensors are often inspired by the biological sense of cutaneous touch, which can detect mechanical stimulation, temperature, and pain, among other things (although pain sensing is not common in artificial device sensors). Tactile sensors are utilized in hardware, security systems, robotics, and computers. Tactile sensors are widely utilized in touchscreen devices like phones and PCs. The LiDAR sensor is widely used in 3D object detection in different vehicle industries, as shown in Fig. 13. Wen et al. [32] proposed a single shared voxel-based backbone for fast and accurate 3D object detection in LiDAR-Camera based AVs. First, this study offers an early-fusion method for quick 3D object recognition using only one backbone, which achieves a good balance of accuracy and efficiency. It comes with a new point feature fusion module that extracts point-wise features from raw RGB images and fuses them with their matching point cloud without using a backbone.

A multi-stage 3D object detection fusion technique was proposed by Jiarong et al. [33]. This is an end-to-end learnable architecture that takes LiDAR point clouds and RGB images as inputs and produces high-precision oriented 3D bounding box prediction using a second-stage detector and 3D region proposal subnet, as illustrated in Fig. 11.

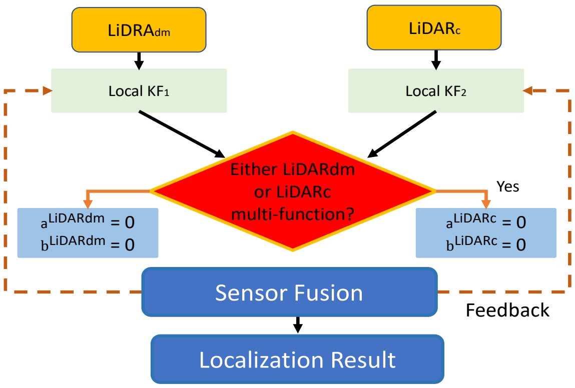


Fig. 10: Shows LiDAR fusion for multi-scale object tracking [29].

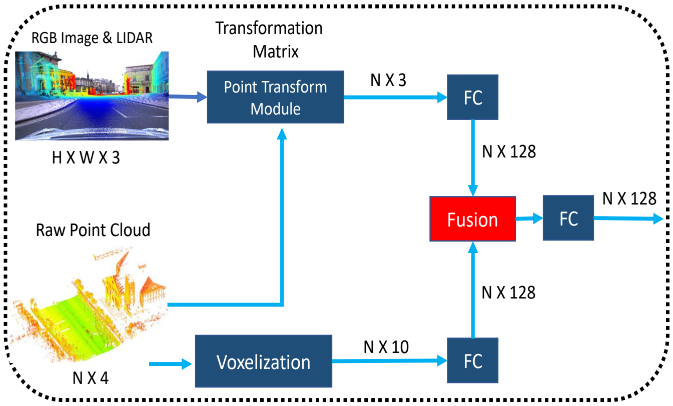


Fig. 11: Shows LiDAR and Camera fusion for multi-scale object recognition [33].

Hechuang et al. [34] use intelligent machine learning visual features to make a multi-sensor fusion module for the detection of an object. This approach investigates multi-sensor fusion approach characteristics in an AV system. According to the needs of the study, an enhanced efficient semantic segmentation network model (Enet-CRF) is built with the help of the efficient neural network (Enet) model. This network architecture combines the original Enet with a CRF-RNN back-end optimization network, which boosts the classification performance by limiting the position relationship between image pixels and RGB data. Experiments show that the developed Enet-CRF improves the obstacle classification performance of pedestrians and bicycles. It fuses the higher-order features of LiDAR and vision sensors using a deep learning network. This method augments the original Enet-CRF network model with high-precision radar information.

C. Autonomous Vehicles Data sets

This section describes the public data set of traffic for object detection and tracking, as shown in Fig. 10. Where the number of new 3D detectors utilize fully supervised models with object bounding boxes learned from labeled data.

1) nuScenes: It is a basic data set that contains visible and 3D Lidar data. It covers the 360° view of the vehicle by

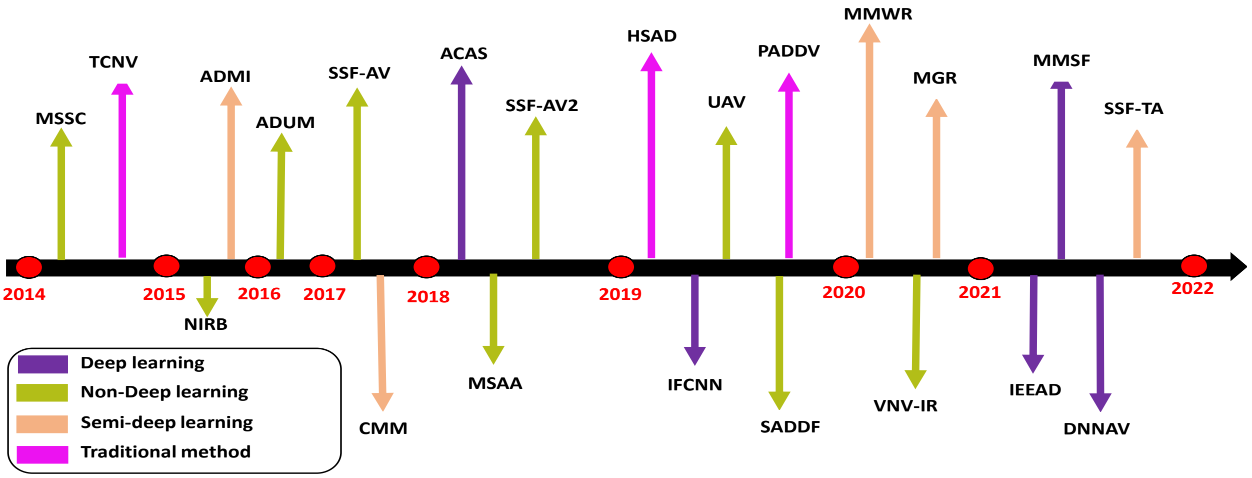


Fig. 12: The 2D detection approaches in Autonomous Vehicles; MSSC [35], TCNV [36], NIRB [37], ADMI [1], ADUM [38], SSF-AV [39], CMM [40], ACAS [41], MSAA [42], SSF-AV2 [39], HSAD [43], IFCNN [44], UAV [45], PADDV [46], SADDF [47], MMWR [48], VNV-IR [49], MGR [50], IEEAD [51], MMSF [52], DNNAV [53], and SSF-TA [54].

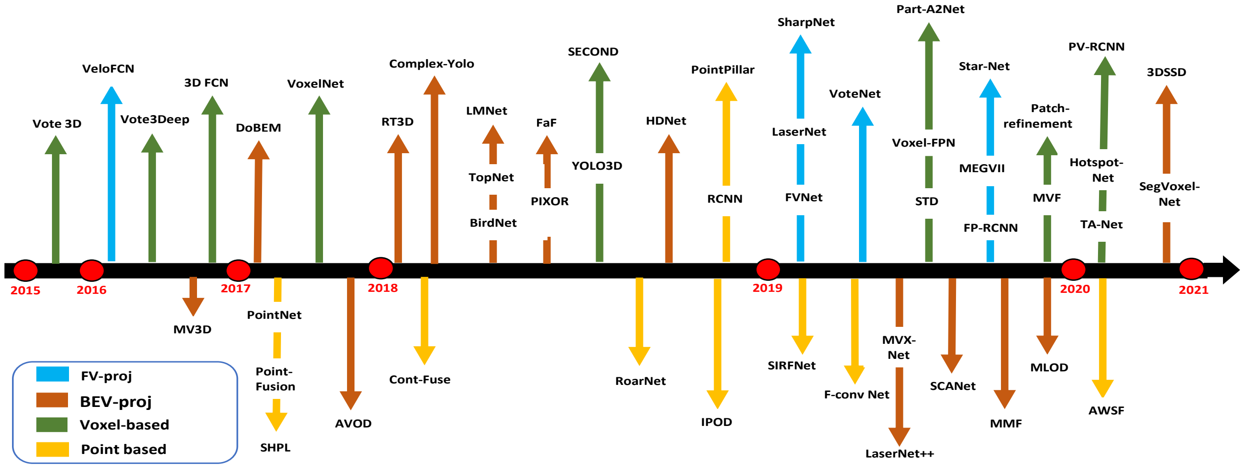


Fig. 13: The 3D detection approaches in Autonomous Vehicles; Vote3D [55], VeloFCN [56], Vote3Deep [57], 3D FCN [12], MV3D [58], DoBEM [59], VoxelNet [60], F-Pointnets [61], PointFusion [62], SHPL [63], AVOD [64], RT3D [65], ComplexYolo [66], BirdNet [67], TopNet [68], LMNet [69], PIXOR [70], FAF [71], Yolo3D [72], SECOND [73], ContFuse [74], HDNET [75], RoarNet [76], Ipod [77], Point RCNN [78], PointPillars [79], SIRFNet [80], FConvNet [81], FVNet [82], LaserNet [83], SHARPNet [84], VoteNet [85], MVXNet [86], LaserNet+ [87], SCANNet [88], MMF [89], STD [90], Voxel-FPN [91], PartA [92], FastPointRCNN [37], MEGV [93], StarNet [94], MLOD [95], AWSF [96], MVF [97], Patch [98], TANet [99], HoptspotNet [100], PVRCNN [101], SegVoxel [102], 3DSSD [103], BirdNet+ [104].

placing different sensors on the vehicle [105]. It has 23 labeled categories. This data set has been collected from over 1000 scenes with various object locations across houses, greenery, road signs, vehicles, and traffic lanes in various illumination conditions. It collects point cloud data using a Velodyne HDL-32 LiDAR.

2) ApolloScape: This data set is related to autonomous driving specifically [106]. It includes a trajectory data set from a 3D perception LiDAR detection and tracker. A mobile LiDAR scanning sensor was used to collect point cloud data, which enables an accurate and solid point than Velodyne LiDAR. About 100,000 images and 80,000 LiDAR point cloud images are included in this data set. For 3D point cloud labeling, ApolloScape has used a semi-automatic strategy. These annotations are manually improved after utilizing PointNet [107] to pre-label the over-segmented point cloud clusters. It is important to note that ApolloScape's marked 3D box only confines the observable individual object, not the entire object.

3) Waymo: It consists of 1,000 scenes that are used for training and validation purposes. For testing purposes, it contains 150 separate scenes that are not included in 1000 scenes [108]. Waymo is the first company to offer a range of images that synchronize the camera and the LiDAR data points. Waymo is working on a LiDAR Honeycomb with a wide field of view for acquiring point clouds. Elongation is one of the sensor features that Honeycomb could provide. It boasts a 360° horizontal FOV and a 95° vertical field-of-view (FOV), as well as a zero-degree minimum range, which makes it simpler to identify and avoid objects in close proximity.

4) FLIR Data: This data set consists of two sensor domains; one is RGB camera, and the second is thermal. It can detect infrared rays or heat reflected from objects. The autonomous driving assistant system (ADAS) outperforms conventional sensor technologies such as visible cameras, LiDAR, and radars. This data set provides both complimentary and unique advantages. FLIR has used thermal sensors in over 500,000 vehicles for driver alert systems. This data set contains 14,000 images, including brief video clips and random image samples. FLIR thermal sensors can identify and categorize people, bikers, animals, and vehicles in tough environments such as complete darkness, haze, smoke, bad weather, and intense light. This data set is open-sourced for training purposes and available at (www.flir.eu).

5) KITTI: Karlsruhe Institute of Technology and Toyota Inc. (KITTI) developed the first and most popular 3D object detection data set [109]. It significantly increases the use of recognition systems in robotics applications. The autonomous driving platform collects actual urban scene data using LiDAR, visible cameras, and a localization sensor system. There are 15 vehicles, 30 pedestrians, and varying degrees of obstruction in each frame, including 2D image annotations and 3D point cloud annotations. In 2017, KITTI released 3D object detection benchmarks, which included 3D and bird's eye view testing. Hundreds of submissions have been made to the KITTI 3D benchmark, which makes it the most popular benchmark.

IV. State-of-the-art SenFusion Networks

In the last decade, many 2D and 3D sensor fusion networks have been proposed. Figures **??** show the state-of-the-art 2D and 3D algorithms, respectively. The related work of 2D and 3D networks has been shown in this timeline (2014 to 2022). For 3D networks, We divided all algorithms into three groups based on point cloud featurization methodology: (1) projection-based approaches, (2) voxel-based approaches, and (3) point-based approaches, as shown in Fig. 13. For the 2D network, we divided all algorithms into four groups (1) deep learning, (2) non-deep learning, (3) semi-supervised, and (4) traditional methods as shown in Fig. 12. The 2D networks are explained in section III (A-I). The methodology, features, and limits of each approach in 2D and 3D networks are summarized in Table I. The following subsections examine the major challenges and technical evaluation of 2D and 3D fusion networks.

A. Front View Projection Based Networks(FVPBN)

These networks used front-view data of different sensor domains. Recently, Dario et al. [110] proposed a 2D sensor fusion technique for an autonomous navigation system. The vehicular Synthetic Aperture Radar (SAR) is used to improve radar imaging capacity and vehicle motion, providing two-dimensional (2D) images of the front view.

Jing et al. [111] proposed a commodity sensors-based network for navigation among pedestrians (NAP) in AVs. A modified velocity-obstacle (VO) algorithm computes velocities and navigates a robot to a target using probabilistic partial observations of the environment. In this technique, an optical flow estimates approach is used to detect the object and sensor fusion in AV. The NAP system benefit is that it uses common visual sensors, such as a mono-camera and a 2D LiDAR, to forecast the velocities and positions of the front view obstacles explicitly.

Hand-generated Front View (FV) Features: The 2D FV map is like a cylindrical image with several feature channels in each pixel. To completely retain the point cloud attributes, researchers established various statistical FV features, which can improve object localization and classification accuracy. The most widely used FV features are shown in Table II. Li et al. demonstrated the VeloFCN [56], a one-stage vehicle detector that uses a CNN for 3D vehicle detection in a LiDAR FV map. They used a 2D fully convolutional network (FCN) to determine object score and position using a 2-channel FV map with length and height information [12].

According to Minemura et al. [82], the FV map can store the 3D information of the vehicle. They suggested LMNet [112] and created a 5-channel input FV map. This technique uses optical flow estimates approach parameters. On the CPU, LMNet does a multi-class detection of cars, pedestrians, and bikes at a rate of 10 fps (frames per second). It is an appropriate technique for low-power robotic devices such as flying drones and housework robots. To reduce data loss during projection, Zhou et al. proposed the FVNet [82], which is called a two-stage detector. They make a three-channel FV map using height, width, and brightness data, then scale it up through closest neighbor interpolation to provide more detail.

B. Bird Eye View Based (BEVB) Networks

The BEVB projection map is created by projecting the whole point cloud onto the RGB map. The ground object length and width are preserved by the BEVB map, which simplifies the calculation of the object's yaw angle. Ground plane projection solves object size and occlusion difficulties in ground-based outside-scene applications such as autonomous driving and object location.

Fusing Multi-View Features: The BEVB maps are frequently combined with data from other sensors. It uses late and deep fusion schemes for fusion frameworks. Chen and his colleagues were the first to use the BEVB map to recognize 3D objects in the MV3D [58] network. Where, a Region Proposal Network (RPN) [84] locates object proposals in BEVB CNN feature maps, as shown in Fig. 14. The obtained multi-sensor characteristics are then merged to enhance the 3D boxes using a deep fusion approach. A point density feature is sliced into 12 height features, and a reflecting feature map is sliced into 14 channels in the input BEVB map. The maximum height of the grid points in the current slice is calculated by each map.

BEVB is extrapolated to 3D boxes using various height priors, which are then projected onto the camera image view and LiDAR front view map to crop the associated proposal region. In the second step, a deep fusion network with element-wise means hierarchically fuses the fixed-length pooled proposal

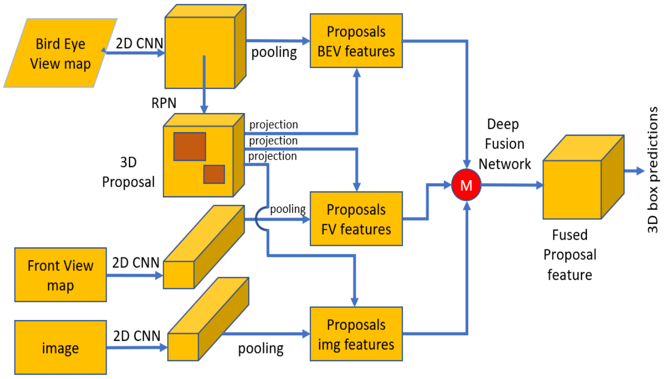


Fig. 14: Shows Bird Eye View Projection based network for 3D object detection MV3D [58]. It takes three inputs (LiDAR bird view, LiDAR front view, and RGB image). The approach begins by generating 3D object proposals from a LiDAR bird's-eye-view map and then projecting them over three views. A deep fusion network is then employed to combine ROI-pooled region-wise features from each view. The fused features are used to predict the object class with 3D-oriented boxes.

features of three perspectives. After that, these fused features are utilized to anticipate object orientation and enhance 3D boxes. Three multi-view 3D detection networks have been discussed in the literature: AVOD [64], MLOD [95], and SCANET [88]. For multi-view purposes, they employ the MV3D [58] detection architecture.

AVOD [64] does a better job than MV3D [58] using the feature extractor to design the BEVB map resolution and down-sampling duration to produce a high-resolution feature map with small object features. In order to build highly expressive feature maps, bottom-up cross-layer connectivity is applied in the feature extraction layers, as inspired by FPN [61]. The AVOD only examines the BEVB map and conducts RPN on both the BEVB and image map at the same time. Before region-wise fusion, a 1×1 convolution [84] is used to filter the outside scene and minimize the dimension of proposal features to improve RPN calculation performance. Deng et al. [95] proposed a multi-sensor network to overcome the problem of low image feature usage.

C. Voxel-Based (VB) Networks

The purpose of voxel-based techniques is to turn irregular point cloud data into a uniform matrix that can be used in an AV system for 3D object detection. The three-dimensional space is partitioned into voxel grids of varying sizes. Each voxel has unstructured points. the Voxelization is three-dimensional, which preserves the underlying point cloud data in a 3D structure. However, there are many empty voxels due to the sparse LiDAR point cloud. The 3D spatial computation rises exponentially as the voxel resolution gets better. The efficient parsing of sparse voxels is difficult for voxel-based algorithms.

Object Localization in Large 3D Voxel Space: Object localization in 3D voxel space is a big challenge. The most widely used paradigm for object localization in 2D images is the sliding window. Due to the increased computing burden caused by the additional dimension of searching space, we use a 3D voxel technique. Wang and Posner [55] proposed a voting schema to make use of the sparsity inherent in point cloud data. They only evaluate the points in each sliding window position that aren't empty when calculating the object classification score. They start by vocalizing the whole point cloud and assigning a feature vector to every occupied voxel point.

Engelcke et al. proposed the Vote3D [55] technique that uses sparse convolutional layers to reproduce the voting schema. They use the Rectified Linear Unit (ReLU) and the l_1 norm to urge the intermediate layer in CNN. Nonlinear CNN is more complicated and expressive than the linear model, and it greatly improves detection accuracy. The same authors describe a 3D FCN network [12] that automatically recognizes objects on the whole output map without the need for an object selection phase by expanding the prior 2D [12] that is performed in a 2D FV map.

Combining Image Features: Two primary networks are proposed to merge voxel and camera image characteristics. The first is the vertical voxel features that are obtained and stored in each BEVB pixel position, and the second is to fuse the pseudo-BEVB map network.

D. Point-Based (PB) Networks

Both voxel-based and projection-based approaches regularize the raw point cloud into image grids or 3D voxel points to perform dense convolution for feature extraction and object recognition. The regularity procedure is complicated in this scenario, and natural point geometry is lost. To minimize information loss, it is appealing to model the raw point cloud feature directly. Since 2017, the point segmentation approach PointNet [61] has created a unified deep network architecture that collects local and global point characteristics by directly consuming unordered point sets. The idea of directly processing point clouds has quickly extended to 3D networks. As a result, 3D object detection is accessible and independent of point cloud type. In this section, we first go over some of the most important related works on point-wise feature extraction.

For point segmentation: PointNet network uses 3D point cloud coordinates and extracts point features and object class labels. After extracting the class label, the saliency-guided transformer network (STN), and multiple layer perceptions (MLPs) network are employed to learn a high-dimensional local geometry of each point. For subsequent point-wise semantic segmentation label prediction, the global semantic features are merged with the local geometry of each point as a point feature. Pointnet models outperform the voxel-based approaches. In a hierarchical approach, the authors enhance PointNet [61] to PointNet++ [107] to incorporate more local neighbor attributes for data points. PointNet++ uses the Furthest Point Sampling (FPS) approach to create local regions to aggregate neighboring points in center points.

Point Features for Image Refining: PointNet and PointNet++ were designed for pixel-wise segmentation and single object detection applications, respectively. These methods faced difficulty to detect the 3D object in 3D space. The image detectors in PointNet were used to trim objects from camera images. The object boxes are then refined using the cropped region attributes. Each 2D bounding box predicted by the image detector F-point network (FPN) [61] is transformed to a 3D frustum using the camera projection matrix. Shin et al. create a variety of networks utilizing RoarNet [76] to reduce the search space for the 3D object. They begin by adapting the Faster RCNN [37] 2D detector to generate 2D proposal boxes with 3D postures. To generate 3D box predictions, the points in the vertical cylinders centered on the centroid of each box are passed into a simple PointNet [61]. Xu et al. introduced PointFusion [62], which uses the PointNet feature and a simple image feature to predict 3D bounding boxes from 2D image crops. They present a unique dense fusion architecture that predicts dense 3D boxes using each 3D point of suggestion. By extracting point-wise information from PointNet-based networks, PV-RCNN [101] enhances 3D voxel CNN proposals. The FPS method encodes a point-wise feature from a small sample of features selected from the whole point cloud.

V. Evaluation metrics

A. *mAP with a Threshold on IoU:*

The mAP value with 3D IoU is used to analyze an object's overall detection performance. The average precision (AP) is used to determine the mAP score of all classes C.

$$mAP = \frac{1}{|C|}\sum_{c\in C} AP_c \tag{1}$$

where C is the set of classes. A series of ranked detection is used to determine the AP. Ranking detection is usually assigned iteratively in most circumstances. The intersection over union technique is used to calculate the 3D IoU, which employs a 3D prediction box and 3D ground truth values. The IoU is used to calculate the precision of object position [109, 113]. For overall performance, we used the TP (true positive) and FP (false positive) values for each object. We apply a certain threshold on the IoU score to calculate the TP and FP scores. It's a TP if the threshold matches the IoU threshold value; else, it's an FP. The KITTI benchmark has a 3D IoU threshold of 0.5 for motorcycles and 0.7 for automobiles. The proportion of all correct detection (TP) in all detection N_{all} is known as precision, which represents the total number of detection that is calculated as:

$$p = \frac{N_{TP}}{N_{all}}, \tag{2}$$

The recall is defined as the percentage of all accurate detections (TP) in ground truth divided by the total number of ground truths:

$$r = \frac{N_{TP}}{N_{all-GR}} \tag{3}$$

Precision and recall are one-sided descriptions of model performance. The P-R (Precision-Recall) curve might be parameterized by increasing the number of ranked detection used in computing precision and recall. The model's accuracy and performance were assessed using the P-R curve. AUC represents the area of the P-R curve. In reality, there are primarily two AP approximations: (1) the N-point interpolated AP metric used in the KITTI and ApolloScape benchmarks, and (2) the AUC-based AP metric used in the Waymo benchmark [108].

1) N-Point Interpolated AP Metric: The average precision (AP) is a measure of the mean precision at a set of N equally-spaced recall levels $[q_0, q_0+\frac{q_1-q_0}{N-1}, q_0+\frac{2(q_1-q_0)}{N-1}, ..., q_1]$, which is denoted as S. This begins at the recall point q_0 and finishes at the recall point q_1:

$$AP = \frac{1}{N}\sum_{r\in S} P_{inter(r)} \tag{4}$$

The precision $P_{inter(r)}$ at each recall level r is interpolated by the highest precision at a score larger or equal to r to maintain the monotonicity of the P-R curve, which is described as:

$$P_{inter(r)} = \max_{\hat{r}:\hat{r}\geq r} p(\hat{r}) \tag{5}$$

where $\hat{r}$ is the highest precision score at recall level r. The 11-Points Interpolated AP measure is a subset of 11 recall levels $S_{11} = [0, 0.1, 0.2, ...1]$. It is used in the early edition of the KITTI benchmark. Simonelli et al. [114] proposed a 40-point Interpolated AP metric to use additional information to approximate the P-R curve and eliminate precise computation of the 0 recall location for properly judging the quality of detection algorithms. $S40 = [1/40, 2/40, 3/40, ...1]$ denotes the subset of 40 tested recall levels. All object recognition task analyses in the KITTI benchmark will use the 40-Point interpolated AP measure started in 2019.

In some cases, the AP measure is incapable of distinguishing between the heads and tails of the objects. The orientation of the 3D box must be evaluated to determine detection quality. KITTI presents an average orientation similarity (AOS) metric that determines the mean orientation similarity s at recall r to quantify the orientation prediction. AOS is a term used to define a group of persons who share some features.

$$AOS = \frac{1}{N}\sum_{r\in S} \max_{x:x\geq r} s(r) \tag{6}$$

The normalized cosine of the orientation similarity $s[0, 1]$ is defined as:

$$s(r) = \frac{1}{|D(r)|}\sum_{i\in D(r)} \frac{1+cos\nabla_\theta}{2}\delta_i, \delta_i = \left\{\begin{matrix} 0, i\in TP \\ 1, i\in FP \end{matrix}\right\} \tag{7}$$

Where $D(r)$ is the set of all object detection with a recall rate of r and δ_i is the angle variation between the recognition i and the ground truth. δ_i determines whether the detection i has been assigned to a ground truth or not.

2) Area Under Curve (AUC) Based AP Metric: AP is calculated by the AUC curve of PR in large-scale Waymo benchmark [108]. It is defined as:

$$AP = 100\int_0^1 \max\{p(r)|r \geq r\}\, dr \tag{8}$$

To avoid excessive AP caused by excessively sparse P-R curve sampling, the maximum gap between two consecutive points in recall value is kept to a pre-set threshold (set to 0.05). If this is not done, further points are added between the two consecutive points. To measure the orientation predictions, the average precision heading (APH) metric is proposed, which is defined as:

$$APH = 100 \int_0^1 \max\left\{h(r) | r \geq r\right\} dr \tag{9}$$

where $h(r)$ is calculated in the same way as $p(r)$, but each TP is weighted by its heading accuracy $min(|\theta|, 2\pi|\theta|)/\pi$s.

B. nuScenes 3D Detection Metric

The nuScenes benchmark builds a TP by measuring the 2D center distance d on the ground plane between detection and ground truth instead of utilizing IoU. We can differentiate detection performance from object size and orientation when we apply the IoU measure to minimize the mismatch of TP objects with small areas. The AUC of a normalized P-R curve is used to generate AP, which excludes points with a recall or accuracy of less than 10% to eliminate noise in low precision and recall regions. Averaging the AP across the D threshold generates the mAP.

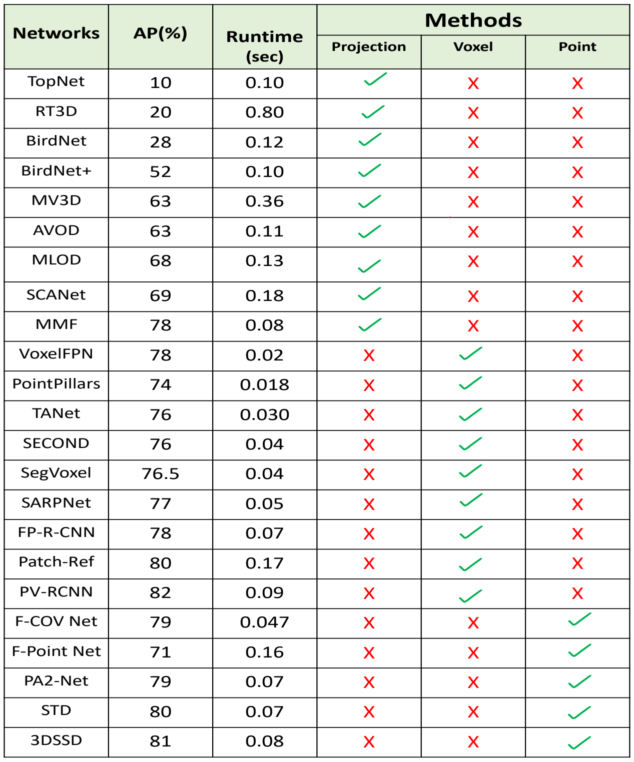
$$mAP = \frac{1}{|C||D|} \sum_{c \in C} \sum_{d \in D} AP_{c,d} \tag{10}$$

A set of true positives (TP) are quantified in terms of size, orientation, characteristics, box position, and velocity, in addition to the mAP metric.

VI. Performance comparison on different benchmarks

This section contains the comparisons of algorithm performance in various scenarios, such as average precision, detection accuracy, LiDAR point cloud sparsity, object densities, etc. Tables I, II, and III compare the algorithms' performance in three public benchmarks: KITTI [109], nuScenes [105], and Waymo [108]. The distribution of data in these three data sets varies. Based on the object height, occlusion, and truncation rate, KITTI has three primary object classes (vehicle, pedestrian, and bike) and shows three categories (Easy, Moderate, and Hard). The nuScenes benchmark comprises eleven object classes that are imbalanced.

The nuScenes benchmark has seven times more annotations and fifty times more scenes as compared to the KITTI benchmark. It has a few LiDAR points as compared to other benchmarks, which are explained in Section IV, and it employs the nuScenes detection score (NDS) as the ultimate assessment metric. There are four object classes in the Waymo benchmark. It contains the most LiDAR points per frame and 3D box annotations and collects training data from a variety of weather scenarios. We compared algorithm accuracy to calculate the algorithm speed in the most recently submitted benchmark KITTI, as illustrated in Table I. Early project-based techniques have a fast detection speed because they compress 3D input to 2D and employ mature 2D CNNs to detect objects. Despite its ease of use, the discretization process diminishes accuracy and leads to information loss. The detector's accuracy can be improved by fusing it with a camera image, but its speed can suffer as a result. PointNet's performance highlights research that directly helps the 3D geometry from raw point cloud data. The detection accuracy of point-based approaches is good. On the other hand, the processing of point-by-point features takes time. Voxel-based techniques employ a PointNet-based module to extract local point characteristics in each voxel after downsampling the z-axis, which is followed by an efficient 2D detector in an organized voxel grid. These methods improve the speed and accuracy of detection.



Networks	AP(%)	Runtime (sec)	Methods		
			Projection	Voxel	Point
TopNet	10	0.10	✓	✗	✗
RT3D	20	0.80	✓	✗	✗
BirdNet	28	0.12	✓	✗	✗
BirdNet+	52	0.10	✓	✗	✗
MV3D	63	0.36	✓	✗	✗
AVOD	63	0.11	✓	✗	✗
MLOD	68	0.13	✓	✗	✗
SCANet	69	0.18	✓	✗	✗
MMF	78	0.08	✓	✗	✗
VoxelFPN	78	0.02	✗	✓	✗
PointPillars	74	0.018	✗	✓	✗
TANet	76	0.030	✗	✓	✗
SECOND	76	0.04	✗	✓	✗
SegVoxel	76.5	0.04	✗	✓	✗
SARPNet	77	0.05	✗	✓	✗
FP-R-CNN	78	0.07	✗	✓	✗
Patch-Ref	80	0.17	✗	✓	✗
PV-RCNN	82	0.09	✗	✓	✗
F-COV Net	79	0.047	✗	✗	✓
F-Point Net	71	0.16	✗	✗	✓
PA2-Net	79	0.07	✗	✗	✓
STD	80	0.07	✗	✗	✓
3DSSD	81	0.08	✗	✗	✓

Fig. 15: Shows the accuracy and run time of different algorithms in autonomous vehicles on KITTI benchmarks.

Tables I, II, and III clearly indicate that detection performance varies with item type. The average AP of a car is 30% greater than the AP of a pedestal and 25% higher than the AP of a bike in the KITTI benchmark. The difficulty is increased by the nuScenes data set, which has more categories and uneven training samples. The nuScenes with the biggest AP difference from the same detection algorithm are 70%. This issue might be explained in three distinct ways. First, the LiDAR points are more susceptible to occlusion, and small-size object classes. This also shows that average detection precision (AP) in the KITTI benchmark is lower in the hard and moderate difficulty levels as compared to the simple difficulty level. Furthermore, due to fewer points, a vehicle of distant objects performs worse in the Waymo benchmark as the distance increases. Second, the training on an unbalanced sample generates uneven detection results across object classes. The sample has a significant amount of vehicle data

TABLE I: Object detection algorithms evaluation on KITTI data set: Vote3D [55], VeloFCN [56], Vote3Deep [57], 3D FCN [12], MV3D [58], DoBEM [59], VoxelNet [60], F-Pointnets [61], PointFusion [62], SHPL [63], AVOD [64], RT3D [65], ComplexYolo [66], BirdNet [67], TopNet [68], LMNet [69], PIXOR [70], FAF [71], Yolo3D [72], SECOND [73], ContFuse [74], HDNET [75], RoarNet [76], Ipod [77], Point RCNN [78], PointPillars [79], SIRFNet [80], FConvNet [81], FVNet [82], LaserNet [83], SHARPNet [84], VoteNet [85], MVXNet [86], LaserNet+ [87], SCANNet [88], MMF [89], STD [90], VoxelFPN [91], PartA [92], FastPointRCNN [37], MEGV [93], StarNet [94], MLOD [95], AWSF [96], MVF [97], Patch [98], TANet [99], HoptspotNet [100], PVRCNN [101], SegVoxel [102], 3DSSD [103], BirdNet+ [104].

Method	Lidar Data	Camera	Car			Pedestrian			Cyclist			mAP	Time
			Easy	*Mod*	*Hard*	*Easy*	*Mod*	*Hard*	*Easy*	*Mod*	*Hard*		
TopNet	BEV	No	12.67	9.28	7.95	10.40	6.92	6.63	2.49	1.67	1.88	5.96	0.101
RT3D	BEV	No	23.74	19.14	18.86	--	--	--	--	--	--	--	0.09
BirdNet	BEV	No	40.99	27.26	25.32	2.04	17.08	15.82	43.98	30.25	27.02	24.86	0.11
BirdNet++	BEV	No	70.14	51.58	50.03	37.99	31.46	29.46	67.38	47.72	42.89	43.68	0.100
MV3D	BEV	Yes	74.97	63.63	54.00	--	--	--	--	--	--	--	0.36
MLOD	BEV	Yes	77.24	67.76	62.05	47.58	37.47	35.07	68.81	49.43	42.84	51.55	0.120
SCANet	BEV	Yes	78.65	68.12	61.44	48.41	37.93	34.10	68.71	53.38	47.59	53.14	0.17
F-PointNet	Point	Yes	82.19	69.79	60.59	50.53	42.15	38.08	72.27	56.12	49.02	56.02	0.17
AVOD-FPN	BEV	Yes	77.63	63.78	55.89	50.46	42.27	39.04	63.76	50.55	44.93	56.02	0.100
PointPillars	Voxel	No	82.58	74.31	68.99	51.45	41.92	38.89	77.10	58.65	44.93	52.2	0.1
PointRCNN	Point	No	86.96	75.64	70.70	47.98	39.37	36.01	74.96	58.82	52.53	57.94	0.1
TANet	Voxel	No	84.39	75.94	68.82	53.72	44.34	40.49	75.70	59.44	52.53	59.91	0.035
SECOND	Voxel	No	84.65	75.96	68.71	--	--	--	--	--	--	--	0.04
SegVoxelNet	Voxel	No	86.04	76.13	70.76	--	--	--	--	--	--	--	0.04
F-CNET	Point	Yes	87.36	76.39	66.69	52.16	43.38	38.80	81.98	65.07	56.54	61.61	0.47
SARPNET	Voxel	No	85.63	76.64	71.31	--	--	--	--	--	--	--	0.05
VoxelFPN	Voxel	No	85.64	76.70	69.44	--	--	--	--	--	--	--	0.02
Patch R	Voxel	No	89.84	78.41	73.15	--	--	--	--	--	--	--	0.15
FP RCNN	Point	No	85.29	77.40	70.24	--	--	--	--	--	--	--	0.06
MMF	BEV	Yes	88.40	77.43	70.22	--	--	--	--	--	--	--	0.08
Part ANet	Point	No	87.81	78.49	73.51	53.10	43.35	40.00	79.17	63.52	56.93	61.82	0.08
3DSSD	Point	No	88.36	79.57	74.55	54.64	44.27	40.23	82.48	64.10	56.90	62.65	0.04
STD	Point	No	87.95	79.71	75.09	53.29	42.47	38.35	78.69	61.59	55.30	61.26	0.08
PV-RCNN	Point	No	90.25	76.82	52.17	52.17	43.29	40.29	78.60	63.71	57.65	62.81	0.08

TABLE II: Object detection algorithms evaluation on nuScene data set: MEGVII [93], SHARPNet [84], PointPillars [79].

Method	Lidar Data	Camera	Car	Truck	Bus	Troller	Vehicle	Person	Motor	Bike	Cone	Barrier	mAP	NDS	Time (s)
MEGVII	Voxel	No	81.10	48.50	54.90	42.90	10.50	80.10	51.50	22.30	70.90	65.70	52.80	63.60	--
SHARPNET	Voxel	No	59.90	18.70	19.40	18.00	11.60	69.40	29.80	14.20	44.60	38.30	32.40	48.40	0.070
PointPillar	Voxel	No	68.40	23.00	28.20	23.40	4.10	59.70	27.40	1.100	30.80	38.90	30.50	45.30	0.020

TABLE III: Object detection algorithms evaluation on Waymo data set: PointPillars [79], MVF [97], PVRCNN [101], Point RCNN [78].

Difficulty	Method	Lidar Data	Camera	3D mAP (IoU=0.7)				3D mAPH (IoU=0.7)			
		Representation	*w/o*	*0-30 m*	*30-50 m*	*50- inf*	*Overall*	*0-30 m*	*30-50 m*	*50- inf*	*Overall*
Level-1	PointPillar	Voxel	No	81.01	51.75	27.94	56.62	--	--	---	--
	MVF	Voxel	No	86.30	60.02	36.02	62.93	--	--	--	--
Level-2	PV-RCNN	Voxel	No	91.92	69.21	42.17	70.30	91.34	68.53	41.31	69.69
	RV-RCNN	Voxel	No	91.58	65.13	36.46	65.36	91.00	64.49	35.70	64.79

due to the high frequency of vehicles on real roads. Robust object detection may be learned by using learning-based methods from a large number of training examples. However, developing a model for object classes with small training data is difficult. Third, the detection network needs to modify the convolution kernels and hyperparameter sizes for fine-tuning. In the nuScenes benchmark, MEGVII [93] exhibits a balanced detection performance across distinct object classes. They improve multi-class detection in a particular way by using class-balanced data augmentation and training data.

VII. Discussion and future perspective

The integration of various sensors domain in autonomous vehicles provides much-needed support and assistance to the vehicle during difficult conditions such as harsh weather, stop signs, traffic congestion, roadside markers, checkpoints, speed, etc. However, the accurate detection of objects for commercial purposes is still a constant dilemma. In this section, we provide a comprehensive comparison of different sensor fusion frameworks.

A. Sparsity of Raw Data

As we raised the resolution of LiDAR data in 3D space, the cost of the device increased. Some projection-based techniques downscale high-resolution LiDAR points to collect high-level deep CNN features [115, 116], but the small-size object shrinks or vanishes. Critical point sampling exacerbates the sparsity problem in point-based methods. The synthesis of 3D objects and complete shape geometry information from a partial point cloud are used to compensate for sparse LiDAR points.

B. Whether to Fuse Multi-Sensor Data

The camera can record the colors and textures of an object, which can then be used with LiDAR data to derive higher-level information [8]. Depending on the use case, there are various fusion techniques that can be applied. Early 3D detection systems [31] relied on image detection for sensor fusion, while current systems leverage deep fusion schemes that add region-wise detail to objects for more accurate bounding box regression, as shown in Fig. 15. Additionally, recent advances in 3D feature extraction have enabled the evaluation of 3D boxes solely via point clouds, enabling a high recall rate. The 3D data from LiDAR can also be leveraged to enhance sensors such as infrared cameras and radars, allowing them to perform under a variety of conditions. Testing these algorithms in a laboratory setting is the first step to take on a path toward real-world applications. However, the evaluation should also factor in the value of the diverse sensor collaborations, in addition to the timestamps and calibration matching.

C. Performance Verification in Complex Scenarios

In rainy and foggy conditions, the RGB camera and LiDAR did not work well. Rain and fog particles are easily reflected by LiDAR in the real environment, causing noise in the LiDAR data. Similarly, the RGB camera produces visual information of poor quality in foggy conditions. On the other hand, thermal and radar technologies are proven to have positive results in such inclement situations. So, potential research is necessary to make robust object detection algorithms that can work with noisy data. It is also expected that multi-class object detection [21, 30, 117] methods will be in high demand in the future. In addition, the intra-class scale difference in the 2D images induced by the camera viewpoint should have the same 3D size [103]. The majority of anchor-based algorithms predict the object bounding box against a predefined 3D box based on data set statistics [12, 55]. Furthermore, algorithms must be capable of handling the learning of disparate classes. As a result, model construction is a promising research issue for unsupervised and semi-supervised learning of small sample data.

D. Perspective of Large-Scale Application of Object Detection

We can't easily modify network parameters for different input sensors or application settings since detection performance is heavily dependent on the training data set. Transfer learning is a simple method for retraining an existing model using a small amount of additional data. It's also normal to combine known and self-learned features to compensate for CNN [44, 118]. The detecting speed is also a matter of concern. Point-wise feature extraction and 3D convolutions should be less expensive to improve performance even further. Existing algorithms are largely intended for autonomous driving in congested areas. When LiDAR and 3D detection technology improves, we expect our research to be useful in a variety of fields such as agriculture, medicine, education, etc.

E. Promising Future Directions and Tasks

Multi-sensor fusion is a major advantage for many fields, allowing robots to become more productive and flexible in industrial applications like material handling, component production, inspection, and assembly [17, 18, 22]. In the past few years, there have been tremendous advances in robotics, especially around multi-robot cooperative systems, under-actuated and powered systems, robot-environment interaction, teleoperation, and visual servicing [17]. All of these are being realised through multi-sensor integration and fusion to improve system capabilities and dependability.

Mobile robotics is one of the most important applications for multi-sensor fusion and integration. Mobile robots can achieve quick perception for navigation and obstacle avoidance when working in unknown or unfamiliar dynamic circumstances by integrating and fusing information from multiple sensors.

VIII. CONCLUSION

Sensors are the important components of autonomous vehicles. The AV technology has been rapidly growing since the last decade. This paper explains the importance of sensors in the autonomous industry, sensor fusion challenges in AV, and the future potential for the new researcher. Sensor fusion techniques enable autonomous vehicles to accurately perceive the operating environment, enabling them to make decisions and control their motions in a safe and reliable manner. The combination of various sensors, such as cameras, LiDAR, RADAR, and GPS, enable autonomous vehicles to accurately identify their surroundings and make decisions that are safe and reliable. Sensor fusion techniques are a crucial part of autonomous driving and are likely to continue to be developed in the future. With further research and development, it is likely that autonomous vehicles will be able to safely navigate complex and changing environments with robust cognitive capability. In the future, we can investigate deeper into the methods of fusing different sensors including visual cameras, thermal (IR) cameras, Radar, LiDAR, Ultra-sonic, etc. In the near future, several directions can be targeted, such as increasing levels of realism achieved by AV simulation software, the recent development of data augmentation techniques (influence various climatological conditions on collected data during driver training), or particular algorithmic suggestions designed to improve the efficiency of AV models to hidden environments tasks.

IX. Acknowledgment

This work is supported by the Innovation and Technology Fund (ITF) of Hong Kong SAR (ITP/015/20AP and ITS/240/21), the Shun Hing Institute of Advanced Engineering at the Chinese University of Hong Kong, and the Research Grants Council (RGC) of Hong Kong SAR (ECS24211020, GRF14203821, GRF14216222), and the Science, Technology and Innovation Commission (STIC) of Shenzhen Municipality (SGDX20220530111005039).

References

[1] J. Rosenzweig and M. Bartl, "A review and analysis of literature on autonomous driving," *E-Journal Making-of Innovation*, pp. 1–57, 2015.

[2] T. Luettel, M. Himmelsbach, and H.-J. Wuensche, "Autonomous ground vehicles—concepts and a path to the future," *Proceedings of the IEEE*, vol. 100, no. Special Centennial Issue, pp. 1831–1839, 2012.

[3] W. Payre, J. Cestac, and P. Delhomme, "Intention to use a fully automated car: Attitudes and a priori acceptability," *Transportation research part F: traffic psychology and behaviour*, vol. 27, pp. 252–263, 2014.

[4] W. Knight, "Driverless cars are further away than you think," *Technology Review*, vol. 116, p. 44, 2013.

[5] National Highway Traffic Safety Administration and U. D. of Transportation, *Traffic Safety Facts 1998-a Compilation of Motor Vehicle Crash Data from the Fatal Accident Reporting System and the General Estimates System*. Createspace Independent Publishing Platform, 1999.

[6] R. Ravindran, M. J. Santora, and M. M. Jamali, "Camera, lidar, and radar sensor fusion based on bayesian neural network (clr-bnn)," *IEEE Sensors Journal*, vol. 22, no. 7, pp. 6964–6974, 2022 .

[7] F. Li, L. Wei, J. Chen, X. Huang, and K. Wang, "Ms-fusion: Multi-layer sensor fusion based robust motion estimation," *IEEE Sensors Letters*, 2023.

[8] J. Alfred Daniel, C. Chandru Vignesh, B. A. Muthu, R. Senthil Kumar, C. Sivaparthipan, and C. E. M. Marin, "Fully convolutional neural networks for lidar–camera fusion for pedestrian detection in autonomous vehicle," *Multimedia Tools and Applications*, pp. 1–24, 2023.

[9] N. Aldahoul, H. A. Karim, M. A. Momo, F. I. F. Escobara, and M. J. T. Tan, "Space object recognition with stacking of coatnets using fusion of rgb and depth images," *IEEE Access*, vol. 11, pp. 5089–5109, 2023.

[10] On-Road Automated Driving (ORAD) committee. (2014) Sae j3016: Taxonomy and definitions for terms related to driving automation systems for on-road motor vehicles. [Online]. Available: https://doi.org/10.4271/J3016_201401

[11] M. Du, "Autonomous vehicle industrialization," in *Autonomous Vehicle Technology*. Springer, 2023, pp. 233–262.

[12] B. Li, "3d fully convolutional network for vehicle detection in point cloud," in *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2017, pp. 1513–1518.

[13] T. Zhou, M. Yang, K. Jiang, H. Wong, and D. Yang, "Mmw radar-based technologies in autonomous driving: A review," *Sensors*, vol. 20, no. 24, p. 7283, 2020.

[14] D. Yang, K. Jiang, D. Zhao, C. Yu, Z. Cao, S. Xie, Z. Xiao, X. Jiao, S. Wang, and K. Zhang, "Intelligent and connected vehicles: Current status and future perspectives," *Science China Technological Sciences*, vol. 61, no. 10, pp. 1446–1471, 2018.

[15] M. Sivak and B. Schoettle, "Road safety with self-driving vehicles: General limitations and road sharing with conventional vehicles," University of Michigan, Ann Arbor, Transportation Research Institute, Tech. Rep., 2015.

[16] A. Alessandrini, A. Campagna, P. Delle Site, F. Filippi, and L. Persia, "Automated vehicles and the rethinking of mobility and cities," *Transportation Research Procedia*, vol. 5, pp. 145–160, 2015.

[17] K. Muhammad, A. Ullah, J. Lloret, J. Del Ser, and V. H. C. de Albuquerque, "Deep learning for safe autonomous driving: Current challenges and future directions," *IEEE Transactions on Intelligent Transportation Systems*, vol. 22, no. 7, pp. 4316–4336, 2020.

[18] V. R. Kumar, C. Eising, C. Witt, and S. Yogamani, "Surround-view fisheye camera perception for automated driving: Overview, survey & challenges," *IEEE Transactions on Intelligent Transportation Systems*, 2023.

[19] S. Kim, H. Kim, W. Yoo, and K. Huh, "Sensor fusion algorithm design in detecting vehicles using laser scanner and stereo vision," *IEEE Transactions on Intelligent Transportation Systems*, vol. 17, no. 4, pp. 1072–1084, 2015.

[20] K. P. Clark, "Aptiv becoming a more sustainable business," *Journal of Applied Corporate Finance*, vol. 31, no. 2, pp. 15–21, 2019.

[21] M. Nawaz, R. Qureshi, M. A. Teevno, and A. R. Shahid, "Object detection and segmentation by composition of fast fuzzy c-mean clustering based maps," *Journal of Ambient Intelligence and Humanized Computing*, pp. 1–16, 2022.

[22] M. Nawaz, S. Khan, J. Cao, R. Qureshi, and H. Yan, "Saliency detection by using blended membership maps of fast fuzzy-c-mean clustering," in *Eleventh International Conference on machine vision (icmv 2018)*, vol. 11041. SPIE, 2019, pp. 565–572.

[23] P.-M. Hsu, M.-H. Li, and Y.-F. Su, "Object detection and recognition by using sensor fusion," in *11th IEEE International Conference on Control & Automation (ICCA)*. IEEE, 2014, pp. 56–60.

[24] H. Ryu, I. Wee, T. Kim, and D. H. Shim, "Heterogeneous sensor fusion based omnidirectional object detection," in *2020 20th International Conference on Control, Automation and Systems (ICCAS)*. IEEE, 2020, pp. 924–927.

[25] Q. Deng, X. Li, P. Ni, H. Li, and Z. Zheng, “Enet-crf-lidar: Lidar and camera fusion for multi-scale object recognition,” *IEEE Access*, vol. 7, pp. 174 335–174 344, 2019.

[26] R. Chellappa, G. Qian, and Q. Zheng, “Vehicle detection and tracking using acoustic and video sensors,” in *2004 IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 3. IEEE, 2004, pp. iii–793.

[27] S. Brooks, “Markov chain monte carlo method and its application,” *Journal of the royal statistical society: series D (the Statistician)*, vol. 47, no. 1, pp. 69–100, 1998.

[28] C. S. Silva and P. Wimalaratne, “Towards a grid based sensor fusion for visually impaired navigation using sonar and vision measurements,” in *2017 IEEE Region 10 Humanitarian Technology Conference (R10-HTC)*. IEEE, 2017, pp. 784–787.

[29] H. Song, W. Choi, and H. Kim, “Robust vision-based relative-localization approach using an rgb-depth camera and lidar sensor fusion,” *IEEE Transactions on Industrial Electronics*, vol. 63, no. 6, pp. 3725–3736, 2016.

[30] M. Nawaz and H. Yan, “Saliency detection using deep features and affinity-based robust background subtraction,” *IEEE Transactions on Multimedia*, vol. 23, pp. 2902–2916, 2020.

[31] J. K. Kim, J. W. Wee, and C. H. Lee, “Sensor fusion system for improving the recognition of 3d object,” in *IEEE Conference on Cybernetics and Intelligent Systems, 2004.*, vol. 2. IEEE, 2004, pp. 1207–1212.

[32] L.-H. Wen and K.-H. Jo, “Fast and accurate 3d object detection for lidar-camera-based autonomous vehicles using one shared voxel-based backbone,” *IEEE Access*, vol. 9, pp. 22 080–22 089, 2021.

[33] J. Wang, M. Zhu, D. Sun, B. Wang, W. Gao, and H. Wei, “Mcf3d: multi-stage complementary fusion for multi-sensor 3d object detection,” *IEEE Access*, vol. 7, pp. 90 801–90 814, 2019.

[34] H. Wang, “Multi-sensor fusion module for perceptual target recognition for intelligent machine learning visual feature extraction,” *IEEE Sensors Journal*, 2021.

[35] R. O. Chavez-Garcia and O. Aycard, “Multiple sensor fusion and classification for moving object detection and tracking,” *IEEE Transactions on Intelligent Transportation Systems*, vol. 17, no. 2, pp. 525–534, 2015.

[36] J. M. Kriesel and N. Gat, “True-color night vision (tcnv) fusion system using a vnir emccd and a lwir microbolometer camera,” in *Signal Processing, Sensor Fusion, and Target Recognition XIX*, vol. 7697. International Society for Optics and Photonics, 2010, p. 76970Z.

[37] Z. Guo, Y. Huang, X. Hu, H. Wei, and B. Zhao, “A survey on deep learning based approaches for scene understanding in autonomous driving,” *Electronics*, vol. 10, no. 4, p. 471, 2021.

[38] I. Shim, J. Choi, S. Shin, T.-H. Oh, U. Lee, B. Ahn, D.-G. Choi, D. H. Shim, and I.-S. Kweon, “An autonomous driving system for unknown environments using a unified map,” *IEEE transactions on intelligent transportation systems*, vol. 16, no. 4, pp. 1999–2013, 2015.

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

[40] Y. Zheng, “An overview of night vision colorization techniques using multispectral images: From color fusion to color mapping,” in *2012 International Conference on Audio, Language and Image Processing*. IEEE, 2012, pp. 134–143.

[41] D. Perić, B. Livada, and M. Perić, “Analysis and selection of components for active swir/nir vision systems,” *Target*, vol. 8, p. 14μm, 2017.

[42] Y. Choi, N. Kim, S. Hwang, K. Park, J. S. Yoon, K. An, and I. S. Kweon, “Kaist multi-spectral day/night data set for autonomous and assisted driving,” *IEEE Transactions on Intelligent Transportation Systems*, vol. 19, no. 3, pp. 934–948, 2018.

[43] J.-P. Giacalone, L. Bourgeois, and A. Ancora, “Challenges in aggregation of heterogeneous sensors for autonomous driving systems,” in *2019 IEEE sensors applications symposium (SAS)*. IEEE, 2019, pp. 1–5.

[44] Y. Zhang, Y. Liu, P. Sun, H. Yan, X. Zhao, and L. Zhang, “Ifcnn: A general image fusion framework based on convolutional neural network,” *Information Fusion*, vol. 54, pp. 99–118, 2020.

[45] J. Weiss, R. Hledik, R. Lueken, T. Lee, and W. Gorman, “The electrification accelerator: Understanding the implications of autonomous vehicles for electric utilities,” *The Electricity Journal*, vol. 30, no. 10, pp. 50–57, 2017.

[46] K. Ren, Q. Wang, C. Wang, Z. Qin, and X. Lin, “The security of autonomous driving: Threats, defenses, and future directions,” *Proceedings of the IEEE*, vol. 108, no. 2, pp. 357–372, 2019.

[47] Y. Kang, H. Yin, and C. Berger, “Test your self-driving algorithm: An overview of publicly available driving datasets and virtual testing environments,” *IEEE Transactions on Intelligent Vehicles*, vol. 4, no. 2, pp. 171–185, 2019.

[48] T. Zhou, M. Yang, K. Jiang, H. Wong, and D. Yang, “Mmw radar-based technologies in autonomous driving: A review,” *Sensors*, vol. 20, no. 24, p. 7283, 2020.

[49] L. Volfson, “Visible, night vision and ir sensor fusion,” in *2006 9th International Conference on Information Fusion*. IEEE, 2006, pp. 1–4.

[50] S. Yu, C. Fu, A. K. Gostar, and M. Hu, “A review on map-merging methods for typical map types in multiple-ground-robot slam solutions,” *Sensors*, vol. 20, no. 23, p. 6988, 2020.

[51] M. A. Khan, “Intelligent environment enabling autonomous driving,” *IEEE Access*, vol. 9, pp. 32 997–33 017, 2021.

[52] Z. Huang, C. Lv, Y. Xing, and J. Wu, “Multi-modal sensor fusion-based deep neural network for end-to-end

autonomous driving with scene understanding," *IEEE Sensors Journal*, vol. 21, no. 10, pp. 11 781–11 790, 2020.

[53] R. Ravindran, M. J. Santora, and M. M. Jamali, "Multi-object detection and tracking, based on dnn, for autonomous vehicles: A review," *IEEE Sensors Journal*, vol. 21, no. 5, pp. 5668–5677, 2020.

[54] D. J. Yeong, G. Velasco-Hernandez, J. Barry, J. Walsh *et al.*, "Sensor and sensor fusion technology in autonomous vehicles: A review," *Sensors*, vol. 21, no. 6, p. 2140, 2021.

[55] D. Z. Wang and I. Posner, "Voting for voting in online point cloud object detection." in *Robotics: Science and Systems*, vol. 1, no. 3. Rome, Italy, 2015, pp. 10–15.

[56] B. Li, T. Zhang, and T. Xia, "Vehicle detection from 3d lidar using fully convolutional network," *arXiv preprint arXiv:1608.07916*, 2016.

[57] M. Engelcke, D. Rao, D. Z. Wang, C. H. Tong, and I. Posner, "Vote3deep: Fast object detection in 3d point clouds using efficient convolutional neural networks," in *2017 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2017, pp. 1355–1361.

[58] X. Chen, H. Ma, J. Wan, B. Li, and T. Xia, "Multi-view 3d object detection network for autonomous driving," in *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 2017, pp. 1907–1915.

[59] S.-L. Yu, T. Westfechtel, R. Hamada, K. Ohno, and S. Tadokoro, "Vehicle detection and localization on bird's eye view elevation images using convolutional neural network," in *2017 IEEE International Symposium on Safety, Security and Rescue Robotics (SSRR)*. IEEE, 2017, pp. 102–109.

[60] H. Yi, S. Shi, M. Ding, J. Sun, K. Xu, H. Zhou, Z. Wang, S. Li, and G. Wang, "Segvoxelnet: Exploring semantic context and depth-aware features for 3d vehicle detection from point cloud," in *2020 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2020, pp. 2274–2280.

[61] C. R. Qi, O. Litany, K. He, and L. J. Guibas, "Deep hough voting for 3d object detection in point clouds," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 9277–9286.

[62] 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*, 2018, pp. 244–253.

[63] Z. Wang, W. Zhan, and M. Tomizuka, "Fusing bird's eye view lidar point cloud and front view camera image for 3d object detection," in *2018 IEEE Intelligent Vehicles Symposium (IV)*. IEEE, 2018, pp. 1–6.

[64] J. Ku, M. Mozifian, J. Lee, A. Harakeh, and S. L. Waslander, "Joint 3d proposal generation and object detection from view aggregation," in *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2018, pp. 1–8.

[65] 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.

[66] M. Simony, S. Milzy, K. Amendey, and H.-M. Gross, "Complex-yolo: An euler-region-proposal for real-time 3d object detection on point clouds," in *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*, 2018, pp. 0–0.

[67] J. Beltrán, C. Guindel, F. M. Moreno, D. Cruzado, F. Garcia, and A. De La Escalera, "Birdnet: a 3d object detection framework from lidar information," in *2018 21st International Conference on Intelligent Transportation Systems (ITSC)*. IEEE, 2018, pp. 3517–3523.

[68] S. Wirges, T. Fischer, C. Stiller, and J. B. Frias, "Object detection and classification in occupancy grid maps using deep convolutional networks," in *2018 21st International Conference on Intelligent Transportation Systems (ITSC)*. IEEE, 2018, pp. 3530–3535.

[69] K. Minemura, H. Liau, A. Monrroy, and S. Kato, "Lmnet: Real-time multiclass object detection on cpu using 3d lidar," in *2018 3rd Asia-Pacific Conference on Intelligent Robot Systems (ACIRS)*. IEEE, 2018, pp. 28–34.

[70] B. Yang, W. Luo, and R. Urtasun, "Pixor: Real-time 3d object detection from point clouds," in *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 2018, pp. 7652–7660.

[71] W. Luo, B. Yang, and R. Urtasun, "Fast and furious: Real time end-to-end 3d detection, tracking and motion forecasting with a single convolutional net," in *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 2018, pp. 3569–3577.

[72] W. Ali, S. Abdelkarim, M. Zidan, M. Zahran, and A. El Sallab, "Yolo3d: End-to-end real-time 3d oriented object bounding box detection from lidar point cloud," in *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*, 2018, pp. 0–0.

[73] Y. Yan, Y. Mao, and B. Li, "Second: Sparsely embedded convolutional detection," *Sensors*, vol. 18, no. 10, p. 3337, 2018.

[74] S. Wang, S. Suo, W.-C. Ma, A. Pokrovsky, and R. Urtasun, "Deep parametric continuous convolutional neural networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 2589–2597.

[75] B. Yang, M. Liang, and R. Urtasun, "Hdnet: Exploiting hd maps for 3d object detection," in *Conference on Robot Learning*. PMLR, 2018, pp. 146–155.

[76] K. Shin, Y. P. Kwon, and M. Tomizuka, "Roarnet: A robust 3d object detection based on region approximation refinement," in *2019 IEEE Intelligent Vehicles Symposium (IV)*. IEEE, 2019, pp. 2510–2515.

[77] Z. Yang, Y. Sun, S. Liu, X. Shen, and J. Jia, "Ipod: Intensive point-based object detector for point cloud," *arXiv preprint arXiv:1812.05276*, 2018.

[78] S. Shi, X. Wang, and H. Li, "Pointrcnn: 3d object proposal generation and detection from point cloud," in *Proceedings of the IEEE/CVF conference on computer*

vision and pattern recognition, 2019, pp. 770–779.

[79] A. H. Lang, S. Vora, H. Caesar, L. Zhou, J. Yang, and O. Beijbom, "Pointpillars: Fast encoders for object detection from point clouds," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 12 697–12 705.

[80] X. Zhao, Z. Liu, R. Hu, and K. Huang, "3d object detection using scale invariant and feature reweighting networks," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, no. 01, 2019, pp. 9267–9274.

[81] Z. Wang and K. Jia, "Frustum convnet: Sliding frustums to aggregate local point-wise features for amodal 3d object detection," in *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2019, pp. 1742–1749.

[82] J. Zhou, X. Tan, Z. Shao, and L. Ma, "Fvnet: 3d front-view proposal generation for real-time object detection from point clouds," in *2019 12th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI)*. IEEE, 2019, pp. 1–8.

[83] G. P. Meyer, A. Laddha, E. Kee, C. Vallespi-Gonzalez, and C. K. Wellington, "Lasernet: An efficient probabilistic 3d object detector for autonomous driving," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 12 677–12 686.

[84] Y. Ye, H. Chen, C. Zhang, X. Hao, and Z. Zhang, "Sarpnet: Shape attention regional proposal network for lidar-based 3d object detection," *Neurocomputing*, vol. 379, pp. 53–63, 2020.

[85] C. R. Qi, O. Litany, K. He, and L. J. Guibas, "Deep hough voting for 3d object detection in point clouds," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 9277–9286.

[86] V. A. Sindagi, Y. Zhou, and O. Tuzel, "Mvx-net: Multimodal voxelnet for 3d object detection," in *2019 International Conference on Robotics and Automation (ICRA)*. IEEE, 2019, pp. 7276–7282.

[87] G. P. Meyer, A. Laddha, E. Kee, C. Vallespi-Gonzalez, and C. K. Wellington, "Lasernet: An efficient probabilistic 3d object detector for autonomous driving," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 12 677–12 686.

[88] H. Lu, X. Chen, G. Zhang, Q. Zhou, Y. Ma, and Y. Zhao, "Scanet: Spatial-channel attention network for 3d object detection," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 1992–1996.

[89] M. Liang, B. Yang, Y. Chen, R. Hu, and R. Urtasun, "Multi-task multi-sensor fusion for 3d object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 7345–7353.

[90] Z. Yang, Y. Sun, S. Liu, X. Shen, and J. Jia, "Std: Sparse-to-dense 3d object detector for point cloud," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 1951–1960.

[91] H. Kuang, B. Wang, J. An, M. Zhang, and Z. Zhang, "Voxel-fpn: Multi-scale voxel feature aggregation for 3d object detection from lidar point clouds," *Sensors*, vol. 20, no. 3, p. 704, 2020.

[92] S. Shi, Z. Wang, J. Shi, X. Wang, and H. Li, "From points to parts: 3d object detection from point cloud with part-aware and part-aggregation network," *IEEE transactions on pattern analysis and machine intelligence*, 2020.

[93] B. Zhu, Z. Jiang, X. Zhou, Z. Li, and G. Yu, "Class-balanced grouping and sampling for point cloud 3d object detection," *arXiv preprint arXiv:1908.09492*, 2019.

[94] J. Ngiam, B. Caine, W. Han, B. Yang, Y. Chai, P. Sun, Y. Zhou, X. Yi, O. Alsharif, P. Nguyen *et al.*, "Starnet: Targeted computation for object detection in point clouds," *arXiv preprint arXiv:1908.11069*, 2019.

[95] J. Deng and K. Czarnecki, "Mlod: A multi-view 3d object detection based on robust feature fusion method," in *2019 IEEE Intelligent Transportation Systems Conference (ITSC)*. IEEE, 2019, pp. 279–284.

[96] Y. Tian, K. Wang, Y. Wang, Y. Tian, Z. Wang, and F.-Y. Wang, "Adaptive and azimuth-aware fusion network of multimodal local features for 3d object detection," *Neurocomputing*, vol. 411, pp. 32–44, 2020.

[97] J. Zhou, X. Tan, Z. Shao, and L. Ma, "Fvnet: 3d front-view proposal generation for real-time object detection from point clouds," in *2019 12th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI)*. IEEE, 2019, pp. 1–8.

[98] J. Lehner, A. Mitterecker, T. Adler, M. Hofmarcher, B. Nessler, and S. Hochreiter, "Patch refinement–localized 3d object detection," *arXiv preprint arXiv:1910.04093*, 2019.

[99] Z. Liu, X. Zhao, T. Huang, R. Hu, Y. Zhou, and X. Bai, "Tanet: Robust 3d object detection from point clouds with triple attention," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 07, 2020, pp. 11 677–11 684.

[100] Q. Chen, L. Sun, Z. Wang, K. Jia, and A. Yuille, "Object as hotspots: An anchor-free 3d object detection approach via firing of hotspots," in *European Conference on Computer Vision*. Springer, 2020, pp. 68–84.

[101] S. Shi, C. Guo, L. Jiang, Z. Wang, J. Shi, X. Wang, and H. Li, "Pv-rcnn: Point-voxel feature set abstraction for 3d object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 10 529–10 538.

[102] H. Yi, S. Shi, M. Ding, J. Sun, K. Xu, H. Zhou, Z. Wang, S. Li, and G. Wang, "Segvoxelnet: Exploring semantic context and depth-aware features for 3d vehicle detection from point cloud," in *2020 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2020, pp. 2274–2280.

[103] Z. Yang, Y. Sun, S. Liu, and J. Jia, "3dssd: Point-based 3d single stage object detector," in *Proceedings of the*

IEEE/CVF conference on computer vision and pattern recognition, 2020, pp. 11 040–11 048.

[104] A. Barrera, C. Guindel, J. Beltrán, and F. García, “Birdnet+: End-to-end 3d object detection in lidar bird’s eye view,” in *2020 IEEE 23rd International Conference on Intelligent Transportation Systems (ITSC)*. IEEE, 2020, pp. 1–6.

[105] 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.

[106] X. Huang, X. Cheng, Q. Geng, B. Cao, D. Zhou, P. Wang, Y. Lin, and R. Yang, “The apolloscape dataset for autonomous driving,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2018, pp. 954–960.

[107] C. R. Qi, L. Yi, H. Su, and L. J. Guibas, “Pointnet++: Deep hierarchical feature learning on point sets in a metric space,” *arXiv preprint arXiv:1706.02413*, 2017.

[108] P. Sun, H. Kretzschmar, X. Dotiwalla, A. Chouard, V. Patnaik, P. Tsui, J. Guo, Y. Zhou, Y. Chai, B. Caine *et al.*, “Scalability in perception for autonomous driving: Waymo open dataset,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 2446–2454.

[109] A. Geiger, P. Lenz, and R. Urtasun, “Are we ready for autonomous driving? the kitti vision benchmark suite,” in *2012 IEEE conference on computer vision and pattern recognition*. IEEE, 2012, pp. 3354–3361.

[110] D. Tagliaferri, M. Rizzi, M. Nicoli, S. Tebaldini, I. Russo, A. V. Monti-Guarnieri, C. M. Prati, and U. Spagnolini, “Navigation-aided automotive sar for high-resolution imaging of driving environments,” *IEEE Access*, vol. 9, pp. 35 599–35 615, 2021.

[111] J. Liang, Y.-L. Qiao, T. Guan, and D. Manocha, “Of-vo: Efficient navigation among pedestrians using commodity sensors,” *IEEE Robotics and Automation Letters*, 2021.

[112] P. Mandikal, K. Navaneet, M. Agarwal, and R. V. Babu, “3d-lmnet: Latent embedding matching for accurate and diverse 3d point cloud reconstruction from a single image,” *arXiv preprint arXiv:1807.07796*, 2018.

[113] M. Nawaz and H. Yan, “Saliency detection via multiple-morphological and superpixel based fast fuzzy c-mean clustering network,” *Expert Systems with Applications*, vol. 161, p. 113654, 2020.

[114] A. Simonelli, S. R. Bulo, L. Porzi, M. López-Antequera, and P. Kontschieder, “Disentangling monocular 3d object detection,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 1991–1999.

[115] B. Yu, H. Liu, J. Wu, Y. Hu, and L. Zhang, “Automated derivation of urban building density information using airborne lidar data and object-based method,” *Landscape and Urban Planning*, vol. 98, no. 3-4, pp. 210–219, 2010.

[116] M. Nawaz, A. Uvaliyev, K. Bibi, H. Wei, S. M. D. Abaxi, A. Masood, P. Shi, H.-P. Ho, and W. Yuan, “Unravelling the complexity of optical coherence tomography image segmentation using machine and deep learning techniques: A review,” *Computerized Medical Imaging and Graphics*, p. 102269, 2023.

[117] M. Nawaz, “Saliency detection and feature matching for object segmentation in digital images,” Ph.D. dissertation, City University of Hong Kong, 2020.

[118] M. Nawaz, R. W. Chan, A. Malik, T. Khan, and P. Cao, “Hand gestures classification using electrical impedance tomography images,” *IEEE Sensors Journal*, vol. 22, no. 19, pp. 18 922–18 932, 2022.

Mehmood Nawaz received the Ph.D. degree from the City University of Hong Kong and the MSc Degree from Shanghai Jiaotong University, China. He is currently a Postdoc fellow in the Department of Biomedical Engineering at the Chinese University of Hong Kong. His research interests include sensor fusion, image segmentation, object detection, object tracking, pattern recognition, etc.

Jeff Kai-Tai Tang received the Ph.D. and B.Eng (Hons.) degrees from the City University of Hong Kong in 2012 and 2003 respectively, and M.Sc. degree in information technology from HKUST in 2005. He is currently a faculty member at the Hong Kong Polytechnic University and he was a senior researcher at the Automotive Platforms and Application Systems (APAS) R&D Centre.

Khadija Bibi received the B.Sc degree in Bio-Technology from Baha-uddin-Zakariya (B.Z.U) University, Multan, Pakistan. She is currently working on Medical Images Segmentation using deep-learning models at the Chinese University of Hong Kong. Her research interests include medical image segmentation, image reconstruction, shape matching, pattern recognition, etc.

Shunli Xiao received his Ph.D. in 2013. He is currently the deputy head cum Senior R&D Manager (specializing in Autonomous Driving Technology) at the Automotive Platforms and Application Systems (APAS) R&D Centre at the Hong Kong Productivity Council. His research interests include smart mobility, robotics, and autonomous driving technology.

Wu Yuan joined the Department of Biomedical Engineering at the Chinese University of Hong Kong as an Assistant Professor in January 2020. He received his Ph.D. degree in Electronic Engineering from the Chinese University of Hong Kong. He was a research associate at the Department of Biomedical Engineering, Johns Hopkins University, with a joint appointment to the Department of Neurologic Surgery, Mayo Clinic.

Ho-Pui Ho received his BEng and PhD. in Electrical and Electronic Engineering at the University of Nottingham in 1986 and 1990 respectively. During 1990-1992, he was a post-doctoral research fellow at the University of Leeds. In 1994, he joined the Fiber Optics Components Operation of Hewlett-Packard as a senior process engineer. He joined The Chinese University of Hong Kong in 2002.