INTERNATIONAL JOURNAL OF AUTOMOTIVE AND MECHANICAL ENGINEERING
ISSN: 2229-8649 e-ISSN: 2180-1606
VOL. 22, ISSUE 1, 2025, 11965 – 11985
DOI: https://doi.org/10.15282/ijame.22.1.2025.1.0920

REVIEW ARTICLE

A Review of Automatic Driving Target Detection Based on Camera and Millimeter Wave Radar Fusion Technology

Tao Zhenhua[1,2], **Ngui Wai Keng**[1*]

[1]Faculty of Mechanical and Automotive Engineering Technology, Universiti Malaysia Pahang Al-Sultan Abdullah, 26600 Pekan, Pahang, Malaysia
[2]Xianyang Polytechnic Institute, Fengxi New Town, Xixian New District, Shaanxi, 71200, China

ABSTRACT – Autonomous driving relies heavily on precise target detection to ensure safety and efficiency in navigating complex environments. It typically utilizes multiple sensors to achieve comprehensive environmental perception. This review explores advancements in integrating these complementary sensors, focusing on state-of-the-art fusion methods, challenges, and applications. The combination of these sensors addresses the limitations of individual modalities: cameras excel in capturing detailed textures and colors, while millimeter wave radar provides reliable distance, velocity, and motion information under adverse weather conditions. Key findings reveal that the sparse radar data, lack of comprehensive multimodal datasets, and difficulties in correlating radar with image data pose significant hurdles. Future research should focus on developing comprehensive multimodal datasets, 4D millimeter-wave radar, and refining fusion algorithms for robustness in diverse environments. This review provides a comprehensive understanding of the current state and challenges in target detection, serving as a foundation for future innovation in autonomous driving technology.

ARTICLE HISTORY

Received	:	10th July 2024
Revised	:	24th Dec. 2024
Accepted	:	05th Jan. 2025
Published	:	20th Feb. 2025

KEYWORDS
Sensor fusion
Target detection
Automatic driving
Camera
MMW radar

1. INTRODUCTION

Driving safety and reducing traffic congestion could both be improved by autonomous driving. Environmental sensing is the primary source of information for autonomous driving, which serves as the basis for route planning and obstacle avoidance [1],[2]. Autonomous cars utilize various sensors, including cameras, lidar, radar, global positioning system (GPS), and inertial measurement units (IMU), to achieve the highest level of accuracy and reliability in sensing [3]. These sensors receive redundant and complementary data [4],[5]. In order to provide more precise information for self-driving cars, the current challenges lie in selecting and combining the data from these sensors. Figure 1 illustrates the typical object detection of automatic driving.

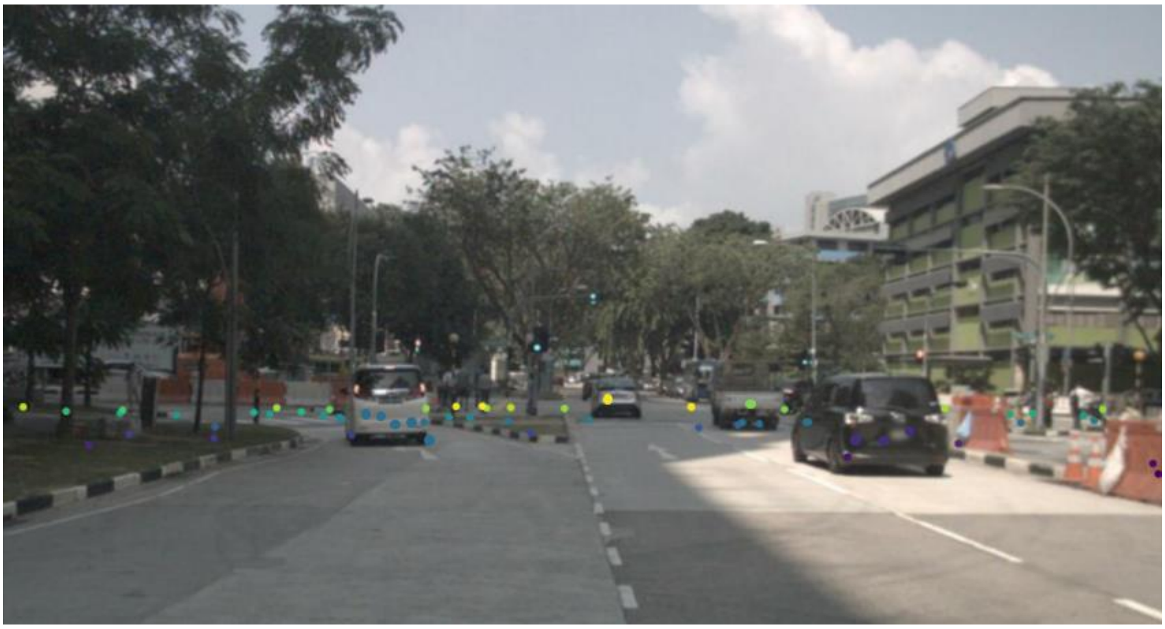


Figure 1. Typical object detection scenarios in autonomous driving. The dots display each radar point's position, while the boxes indicate the detection results. The dot's darkness indicates how close it is to the self-vehicle. These images were generated from the nuScenes [6] dataset

***CORRESPONDING AUTHOR | W.K. Ngui** | ✉ wkngui@umpsa.edu.my

The sensor most frequently used in autonomous driving is the camera, which is primarily applied for tracking, target identification, and segmentation. Additionally, lidar is commonly used to determine the spatial location of targets. The object's contour becomes more distinct as lidar emits more beams [7]. The data picked up by the lidar and the camera are complementary. As a result, the combination of lidar with cameras has recently gained popularity and has shown a respectable level of accuracy for both 2D and 3D target detection [8], [9], [10], [11]. However, lidar and camera fusion identification performance can be significantly reduced in unfavorable weather like fog, rain, snow, and bright light [12]. Furthermore, the broad usage of lidar is hindered by its high cost [13]. Millimeter wave (MMW) radar performs well in all weather conditions except for heavy rain and can penetrate fog, smoke, and dust better than lidar [14], [15]. Additionally, MMW radar can accurately determine the velocity of any object it detects using the Doppler effect with no temporal data [16]. In the application of advanced driver assistance systems (ADAS), MMW radar is widely used, like adaptive cruise control (ACC), Collision Avoidance (CAS), automatic emergency braking (AEB), and lane change assist (LCA).

Currently, there is an increasing number of MMW radar applications for self-driving cars. However, there are two key issues that explain why there are not as many investigations on the integration of MMW radar and cameras. Firstly, MMW radar data collection produces low-resolution images, resulting in sparse point clouds and a lack of height information. One problem is the lack of autonomous driving datasets, including MMW radar and cameras, which hinders researchers from conducting more in-depth analyses. Most algorithms for MMW radar currently rely on processing lidar data. However, there is a significant difference between the point cloud of lidar and MMW radar, with the latter being much sparser. As a result, applying algorithms and processing the point cloud often yields suboptimal results. While some researchers have attempted to improve point cloud density by using multiple frames of radar data, this approach also increases system latency.

Surveys related to sensor fusion for autonomous driving have primarily focused on fusing cameras and lidar, as well as other areas of sensor fusion such as lidar, MMV radar, cameras, or other sensors. For instance, Huang et al. [3] provided a comprehensive review of autonomous driving sensors and their fusion techniques. However, although fusion algorithms and evaluations are briefly discussed, this investigation primarily focuses on deep-learning fusion techniques using cameras and MMW radar. Because there has been limited previous research on integrating camera and radar sensors in autonomous vehicles, it is difficult for researchers to present a comprehensive outline of this area. This article aims to close the gap by thoroughly examining the integration of cameras and radars in self-driving vehicles. Our review highlights the following contributions:

- The study focuses exclusively on the combination of MMW radar and cameras for target detection in autonomous driving. We specifically limit ourselves to analyzing the use of datasets to assess how fusion algorithms are utilized.
- We provide an overview of automated driving datasets and algorithms from 2020 to 2023 and offer insights into fusion methods.
- Our analysis identifies key challenges and issues in camera and radar fusion and suggests possible orientations for follow-up research.

The rest of this article is structured as below: Section 2 describes the working principle and the application of the two sensors, camera and MMW radar, in target detection for species on self-driving vehicles and analyses the complementary features of MMW radar and camera in target detection. Then, Section 3 describes the autonomous driving dataset containing both MMW radar and camera data. Afterward, Section 4 presents various target detection methodologies on the grounds of integration of MMW radar and camera data. Section 5 outlines the pertinent evaluation metrics for target detection algorithms. Section 6 examines the integration process of MMW radar and camera for data annotation, data processing, selection of fusion methods, and construction of the fusion framework. Finally, Section 7 provides a conclusion.

2. TARGET DETECTION SENSORS AND APPLICATIONS (CAMERA AND MILLIMETER WAVE RADAR)

This section aims to outline the background on MMW radar and radar-camera fusion in self-driving vehicles. Firstly, we outline the fundamental concepts and signal types of both camera and MMW radar. Subsequently, we discuss the pros and cons of each sensor, as well as the corresponding target identification algorithms for autonomous driving. Our primary objective is to compare the properties of these two sensors with the aim of highlighting the significance of combining MMW radar and camera.

2.1 Sensors

This section explains the basic principles of cameras, millimeter-wave radar, and common signal formats. It also discusses the applications of these sensors in object detection and compares their advantages and limitations.

2.1.1 Camera

A camera captures light from the environment through a lens positioned in front of the sensor, which then projects the light onto a light-sensitive surface to create an image of the surrounding area [3], [17]. Figure 2 illustrates the process of imaging with a complementary metal oxide semiconductor (CMOS) camera. Typically, a camera is comprised of the lens, image sensor, input/output (I/O) interface, and an image signal processor (ISP) [18]. The lens gathers the light from the target and focuses it on the image sensor. The image sensor changes light waves into electric impulses, which are then converted into digital signals using an on-chip analog-to-digital converter (ADC). The ISP takes care of post-processing tasks like noise reduction and then transforms the digital signal into the RGB data format of the picture or video. Finally, the I/O interface is used to transmit and display the picture data.

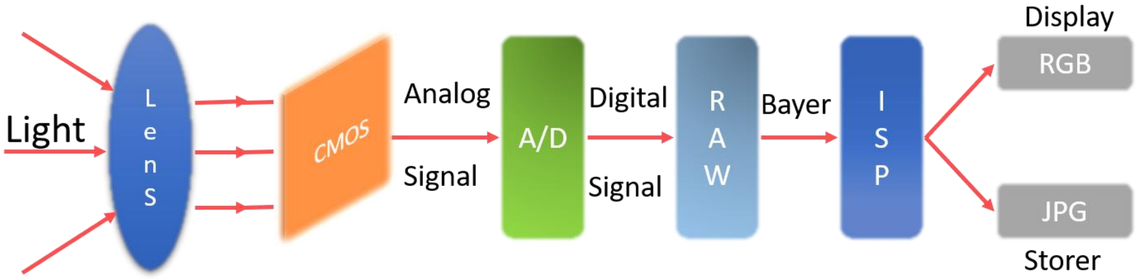


Figure 2. The process of imaging with a CMOS camera

Affordably priced cameras are equipped with software that can identify moving or stationary objects in their vision. High-resolution photos of the environment, complete with details on color and texture, can also be obtained from them. These features allow vehicle perception systems to recognize items such as other cars, road signs, traffic signals, lane markers, and barriers for moving vehicles. Lighting and unfavorable weather, such as snow, strong sun glare, heavy rain, and foggy days, can significantly affect the camera's image quality (resolution). Furthermore, the camera usually detects objects without distance information. An image is a two-dimensional grid of pixels, and each pixel stands for a different color at a specific point within the image. Color images typically use RGB (Red, Green, Blue) channels to represent the color of each pixel. Each pixel in a color image encompasses the values of three channels that denote blue, green, and red components.

2.1.2 MMW Radar

Millimeter wave typically refers to the 30 to 300 GHz band, which corresponds to a wavelength of 1 to 10 mm. Radar is, actually, the optimal sensor for determining radial velocity and distance. It was developed before the Second World War and is currently used in self-driving cars. There are two types of MMW radar used in self-driving cars: frequency-modulated continuous wave (FMCW). These radars acquire information by processing internally emitted signals and the reflected signals from objects [19]. Using the single-transmitter-receiver (1TIR) MMW radar system as a model, like Figure 3, the internal synthesizer generates a linear frequency modulation (FM) pulse, which is then delivered through a transmit (TX) antenna. The receive (RX) antenna receives the pulse after it has been reflected off the object. The mixer produces an intermediate frequency (IF) signal by combining the RX and TX signals [20].

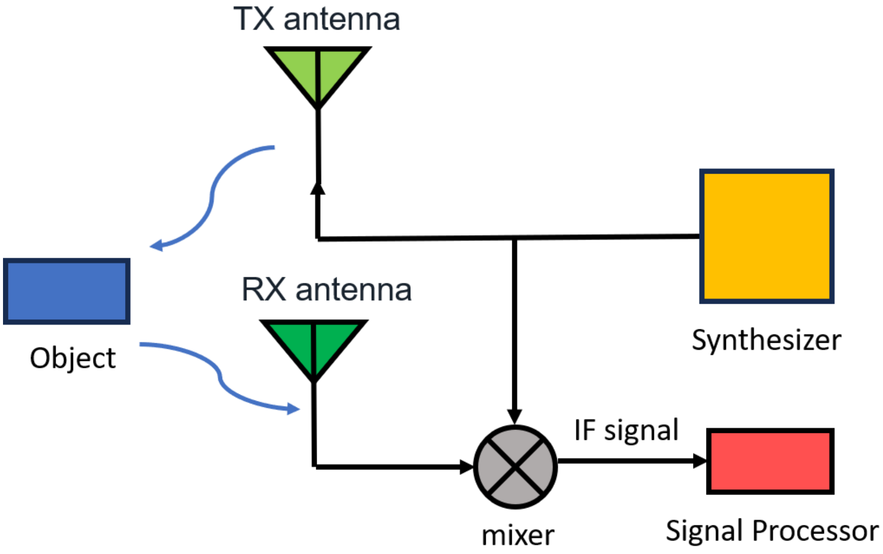


Figure 3. Working principle of single transceiver millimeter wave radar

The wavelength of millimeter waves falls between that of centimeter waves and light waves, giving millimeter waves the advantages of both microwave and photoelectric guidance. Millimeter wave guides are distinguished from centimeter-wave guides by their elevated spatial resolution, light weight, and small size. In addition, millimeter waves can penetrate smoke, dust, and fog, making them suitable for all weather conditions (except during heavy rain), unlike infrared, laser, television, and other optical guides. However, radar sensors have a drawback in that they may falsely detect metallic targets, like road guardrails or signs, in the surrounding environment. Furthermore, radar struggles to differentiate between static and stationary objects [21], and it is unable to discern colors, resulting in poor target classification [22]. MMW radar is a widely used and essential sensor for self-driving vehicles. It is known for its long detection range, affordability, and ability to detect moving targets. It can identify obstacles at a distance of 250 meters, which is important to the security of self-driving vehicles. With a precision of 0.1 m/s, MMW radar is capable of determining the corresponding speed of the target vehicle through the Doppler effect. This data is essential for self-driving vehicles to make informed decisions [23].

Depending on the various steps in the fast Fourier transform (FFT) based signal processing chain, there are multiple formats available for extracting raw data obtained by MMW radar [24]. These formats include the point cloud, the range-azimuth heat map, the range-azimuth-Doppler (RAD) tensor, along with the micro-Doppler spectrogram. Among these, the point cloud format becomes the most widespread.

- Range-Azimuth-Doppler

Three dimensions of the RAD data block undergo an FFT operation: angle, velocity, and distance. This operation results in the RAD data block, which characterizes distance-angle-velocity. The rotation angle in the horizontal direction represents the angle. Although the RAD tensor is large, it preserves strong radar features in the Doppler dimension and provides a high-resolution range. As a result, it enables the localization of the agent from an upward perspective using distance azimuth. In this format, target properties like location, 2D shape, and velocity can be inferred straightforwardly from the tensor using learned detection models.

- Point cloud

A sparse point cloud is generated through constant false alarm rate (CFAR) operations on RAD-dense data blocks. In addition to radar-specific characteristics such as radial velocity, signal-to-noise ratio, and measured radar cross section (RCS), each cloud point has 3D positional characteristics such as azimuth, distance, and elevation (if applicable). A point cloud is an intuitive spatial representation that works well for both visualization and analysis. However, it does not reliably transmit silhouette information [25], [26]. The point cloud format significantly reduces data dimensionality while maintaining important object and scene information, making it very useful for object recognition and classification algorithms. Nevertheless, weaker data from the agent may be filtered out by the point cloud format, which could be problematic for high-performance radar deep learning models.

- Micro-Doppler

A two-dimensional representation of Doppler frequency change over time is called a micro-Doppler spectrogram. In the radar signal processing chain, it is attained through implementing a short-time Fourier transform after a distance Fourier transform. The process of generating different MMW radar signals is shown in Figure 4. This approach effectively captures the motion characteristics of the agent. The micro-Doppler features of motion and twisting events for vehicles, bicycles, and walkers in Figure 5(d) demonstrate that these features are specific to distinct subjects and types of motion. Doppler spectrograms can be utilized to classify different objects and infer their actions. A study on the usage of this format for bird and drone detection was conducted in [27]. Since the format of the 2D spectrogram does not catch the spatial data, it is not suitable for use with target detection models. Nevertheless, the format highlights the crucial role of time-changing Doppler features in fulfilling accurate object detection in radar data.

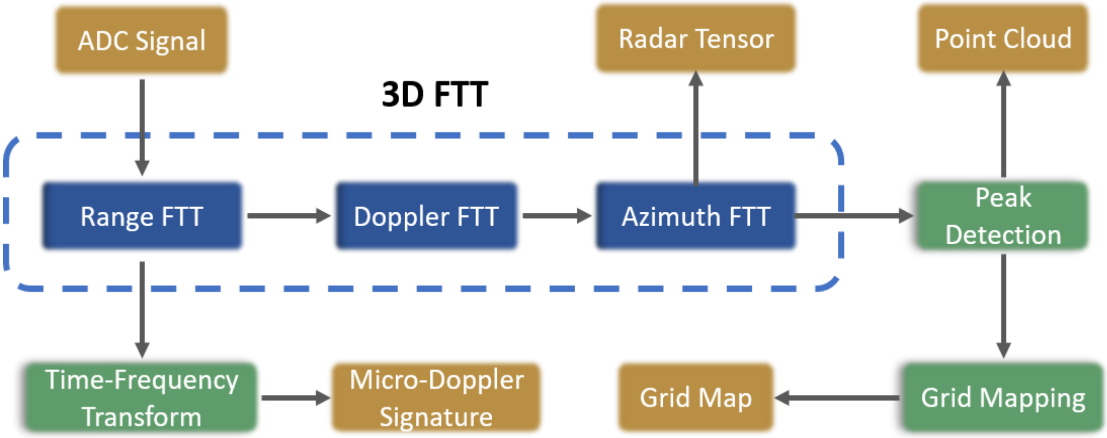
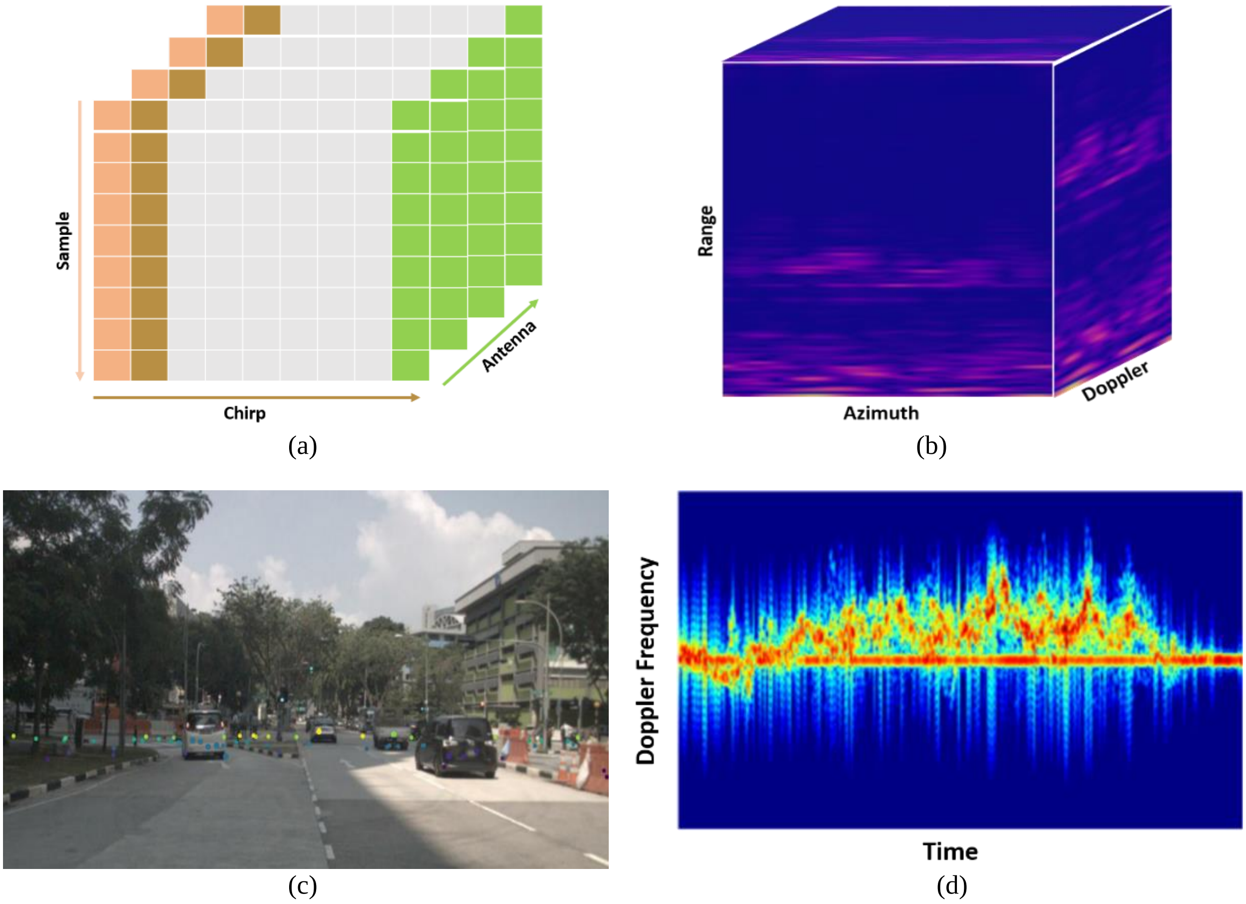


Figure 4. Generation process of different millimeter wave radar signals



(a) (b)

(c) (d)

Figure 5. Radar signal representation. (a) ADC signal in Simple-Chirp-Antenna tensor format. (b) Radar tensor represented by the 3D distance-azimuth Doppler tensor. The image was generated from the CARRADA [28] dataset. (c) Point cloud projected on the 2D image plane. The image was yielded from nuScence [6] dataset. (d) Micro-Doppler feature with pedestrian movement. The image was yielded from Open Radar dataset [29]

2.2 Application in Target Detection

Target detection involves identifying the position and type of targets by analyzing data from a picture or radar detection. Typically, researchers utilize cubic or rectangular boundaries to represent the bounding box of the object, as described in Figure 6.

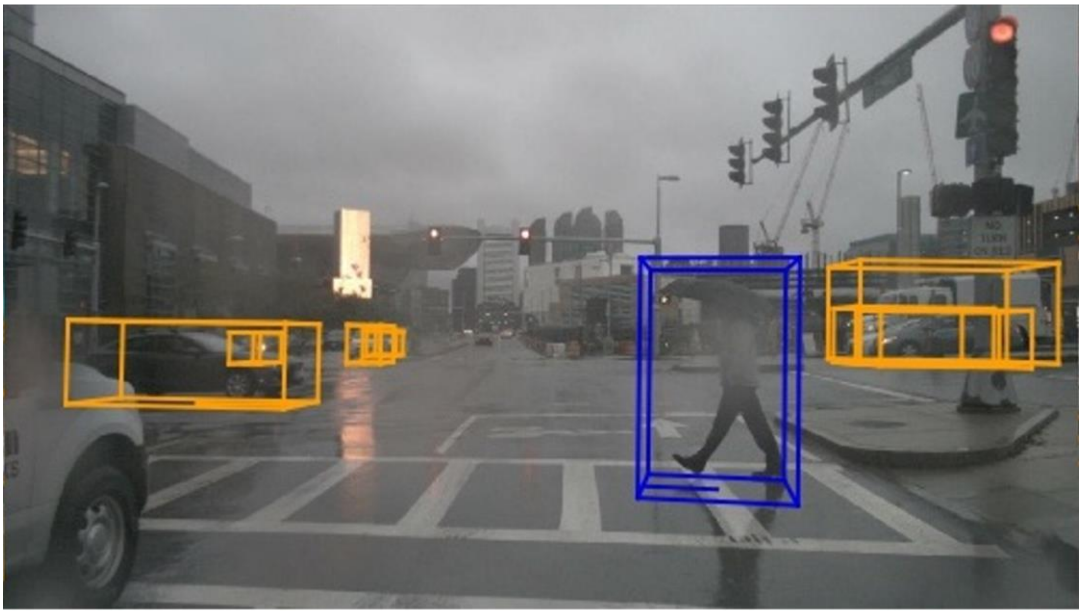


Figure 6. Target detection 3D box labeling from nuScenes [6]

2.2.1 Camera-Based Target Detection

Based on deep learning, target identification techniques have recently been increasingly popular. These techniques can be separated into two main types based on their approach to problem-solving. The first type consists of two-stage algorithms for detection, including R-CNN [30], Fast R-CNN [31], Faster R-CNN [32], and others. The algorithms extract target information from the candidate box of the object image and subsequently adopt the detection network to predict the object's location within the candidate box. They utilize heuristics such as the Selective Search Algorithm or convolutional neural networks (CNN) networks like region proposal network (RPN) networks to generate the Region Proposal. The algorithms then perform further target classification and location regression on the candidate frames. Among the two-stage detection algorithms, the R-CNN group, along with several enhanced versions derived from the R-CNN algorithm, is the most commonly used.

The second category consists of one-stage detection algorithms, like the SSD [33] and YOLO [34] series algorithms. These algorithms use a single CNN network for predictions, simplifying candidate area selection. When an image is inputted for detection, the network treats the detection process as a regression problem, which improves detection speed. In [35], Abdul Razak et al. employed an SSD-based TensorFlow Lite network to detect obstacles in images captured by a single camera. The results demonstrated that the corresponding obstacles could be identified with a probability of detection ranging from 50% to 80%, with enhanced performance observed during the daytime. However, compared to the R-CNN series, one-stage algorithms trade off accuracy for speed.

Unlike CNN-based detectors, transformer-based methods represent the most recent advancement in utilizing the self-attention mechanism to simulate contextual features and their relevance. This is achieved through the use of the self-attention mechanism. Representative transformer detectors include DETR [36], Deformable DETR [37], WB-DETR [38], and Swin [39]. Furthermore, several research efforts have been made to expedite conventional transformer modules by integrating self-attention and convolution, thereby combining the strengths of CNNs and transformers. Examples of such studies include Conformer [40], MobileVIT [41], and Visformer [42].

2.2.2 MMW Radar-Based Target Detection

Radar-based target detection methods are extensively applied to explore vehicles [43], [44], and pedestrians [45], [46]. Scientists commonly employ image-oriented networks, such as YOLOv3 [47] and Faster R-CNN [32], for identifying objects in different types of radar tensors: the 2D RA tensor [43], [48], the 2D RD tensor [49], [50], and the 3D RAD tensor [13], [51], [52]. However, compared with images, radar tensors are short of physical meaning, which complicates the conversion of characteristics acquired from image-oriented algorithms into radar information. Moreover, it is difficulty to use algorithms in radar tensors in real-time owing to their complex size, noise, interference, and clutter. Diverse point-based network models were employed to identify objects within radar data displayed in point cloud layouts. Point-by-point methods [53], [54], [55] directly operate on the original point cloud and utilize LIDAR-based algorithms, such as PointNet [56], PointNet++ [57], and Frustum PointNets [58], to classify the points into different object classes.

For mapping a 3D point cloud into a mesh-like structure, such as one 2D image plane or 3D voxel mesh, mesh-based methods [59], [60], [61] are used. Object detection algorithms, like YOLOv3 [47] and Voxel Net [62], are then utilized in the mesh presentation to identify targets. Grid-based methods have proven to be efficient in handling big datasets and are often applied to real-time applications. Graph-based approaches, such as Radar-Point GNN [63], employ Graph Neural Networks (GNN) to investigate targets in radar point clouds. These techniques consider points as nodes and their interconnections as edges within a graph. By employing graph algorithms and frameworks, the techniques catch spatial interconnections and contextual details in points efficiently, thus enhancing detection efficiency in contrast with conventional point-by-point approaches. However, creating graphs and extracting characteristics from point clouds present significant computational challenges, especially when handling extensive datasets.

2.3 Evaluating Radar and Camera

MMW radar is capable of measuring distance, speed, and azimuth [3]. Currently, vehicles equipped with driver support systems can identify distances up to 300 meters, providing a horizontal viewing angle of 140° and an angular resolution below 1° [51],[52]. Furthermore, the radar sensor's ability to withstand night and severe weather allows it to function during one day. On the other hand, cameras provide information on the color, texture, and shape of objects. In terms of classification, cameras outperform radar sensors. Both MMW radar and cameras are much cheaper than lidar for vehicle installation and are widely used. Although MMW radar and cameras have distinct advantages and disadvantages, they are irreplaceable; however, they can be combined to ensure sufficient access to information. Figure 7 shows the difference and connection between MMW radar and camera for object feature detection. Enhancing understanding of external data can be achieved by leveraging the complementary strengths derived from their individual characteristics. Moreover, if one sensor malfunctions, the other remains operational, thereby enhancing the credibility of the autonomous driving mechanism. Integrating camera and radar sensors is important to sustain the perceptual precision and stability of self-driving vehicles.

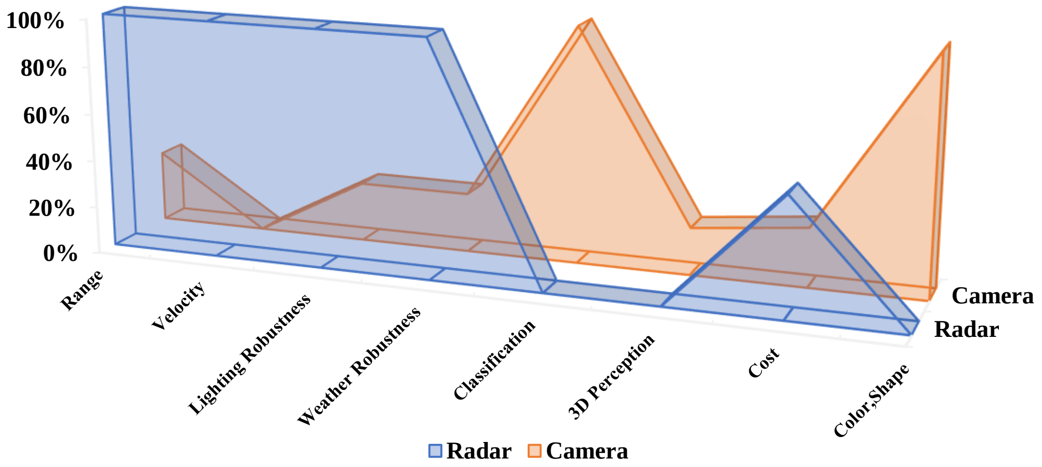


Figure 7. Complementary Diagram of Radar and Camera Information

3. DATASETS

Datasets are key players in object detection studies. Superior, extensive datasets are also key players in continuously enhancing and evolving algorithms for recognizing, detecting, and classifying objects, especially in deep learning. Furthermore, educating deep neural networks for intricate tasks, including auto driving, needs a substantial volume of training data. Consequently, to maintain the network's resilience and precision in intricate driving scenarios, a comprehensive, high-caliber, and annotated dataset from the real world is required. The dataset should encompass diverse driving scenarios, sizes of objects, and a range of benchmark calibration challenges. Table 1 lists some typical datasets and their basic information. This section concisely outlines a selection of typical datasets for automated driving.

Table 1. Autonomous driving datasets

Dataset	Release	Scale	Scenarios	Camera	Radar	Lidar	Classes	Use
KITTI [65]	2012	1.5h	ML	Y	N	Y	8	T
Cityscapes [66]	2016	16.7h	CT	Y	N	N	19	S
Apolloscape [67]	2018	16.7h	ML	Y	N	Y	35	S
Waymo [68]	2019	6.4h	CT	Y	N	Y	4	T
nuScenes [6]	2019	5.5h	ML	Y	Y	Y	23	T&S
CRUW [69]	2020	3.5h	ML	Y	Y	N	3	T
CARRADA [70]	2020	21min	EX	Y	Y	N	3	T
VOD [71]	2022	12min	ML	Y	Y	Y	3	T

In Table 1 "Y" indicates the presence of sensing data in the dataset, "N" indicates its absence. "ML" refers to city, town, highway, and other locations, "CT" refers to city ,"EX" refers to experiment. "D" indicates detection, "S" indicates segmentation.

There are several datasets containing high-quality MMW radars. This presents a challenge for researchers studying MMW radar and camera fusion. Hereafter, we describe in detail the three datasets that include MMW radar data: nuScenes, CRUW, CARRADA, and VOD.

3.1 nuScenes

The nuScenes dataset, developed by nuTonomy, is extensively identified as one of the most prominent publicly available datasets in autonomous driving. It is the largest collection of MMW radar signals for autonomous driving and includes 1,000 scenarios, each lasting 20 seconds. These scenarios feature multiple lanes, pedestrians, vehicles, and various road and traffic events. Notably, it is the first dataset to be equipped with sensors from fully self-driving vehicles. The dataset provides camera, lidar, and radar data, including radar point cloud data. The nuScenes 3D edge annotation consists of 23 classes and eight attributes, such as pedestrian posture and vehicle state.

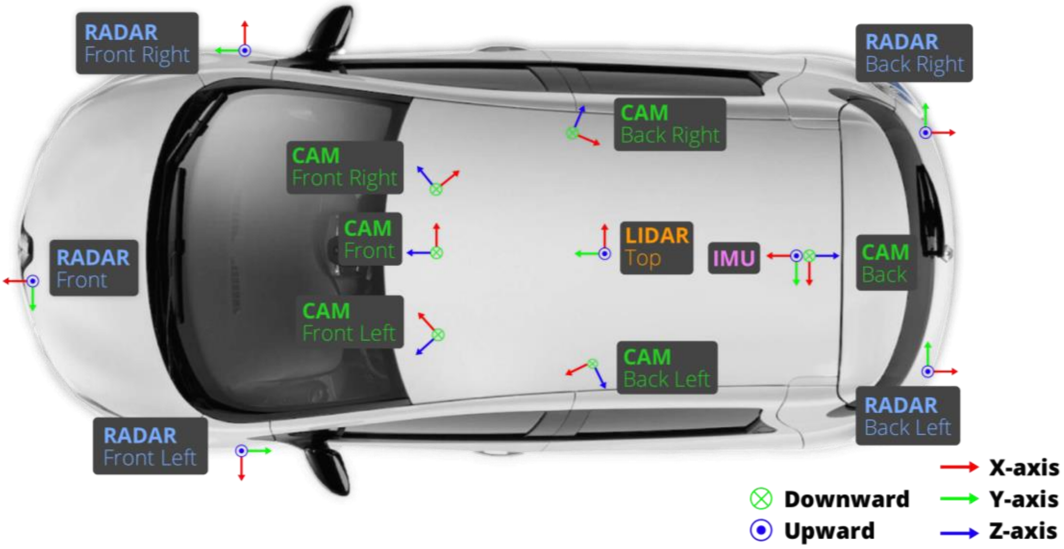


Figure 8. Sensor setup for nuScenes data collection platform [6]

Figure 8 displays the configuration of the nuScenes acquisition device. It includes a 32-line lidar (Velodyne HDL32E) operating at 20Hz and capturing 1.39M point clouds per second. Additionally, there are five 77GHz long-range MMW radars (Continental ARS 408-21) operating at 13Hz, six cameras (Basler acA1600-60gc) with a resolution of 1600x1200 at 12Hz, and one set of IMU and GPS.

As seen in the diagram, the sensor configuration of the nuScenes dataset closely resembles that of a production vehicle. It provides complete MMW radar data, including front and corner radar, which effectively reduces blind spots. This setup is ideal for researching MMW radar algorithms. Furthermore, nuScenes offers 360-degree ring-view data. The bird's eye view (BEV) sensing direction is becoming increasingly popular, and nuScenes' sensor configuration can provide matching data, resulting in excellent results based on the nuScenes dataset.

3.2 CRUW

The University of Washington researchers released the CRUW database in 2020. This database is relatively large-scale, containing 3.5 hours of 30 FPS (~400K frames) camera radar information. The data comes from various driving scenes, like urban streets, parking lots, highways, and campus roads. CRUW database provides object-level annotations, such as object location, size, and class. Additionally, it includes mask information on top of the image data. The radar data format used is the Range-Azimuth Map.

The CRUW Autonomous Driving dataset is an open-source dataset that uses radar frequency-domain imagery. It is primarily designed for target detection and tracking tasks in autonomous driving. This dataset is one of the largest international radar datasets available for autonomous driving scenarios. It features multiple scenarios, a large size, and realism. Figure 9 displays the configuration of the CRUW acquisition device. The sensor platform consists of two stereo cameras and two 77GHz FMCW radars. The FLIR BFS-U3-16S2C-CS cameras and TI AWR1843 + DCA1000 radars are used in this setup.

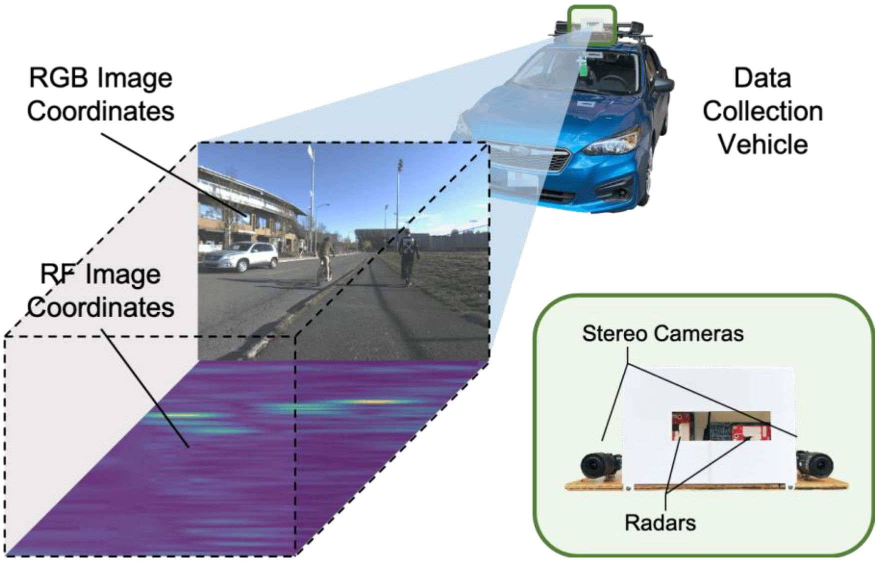


Figure 9. Sensor setup for CRUW data collection platform [69]

The CRUW dataset consists of radar data from disparate scenes, like cities, towns, and highway areas, which involves a series of weather and lighting conditions. Each radar scan in the dataset includes precise 3D annotations for vehicles, pedestrians, and road signs. This feature is beneficial for researchers as it allows for training and validating target detection and tracking algorithms. In comparison to other radar datasets, the CRUW dataset offers radar scans from various scenarios, making it suitable for training different target detection and tracking algorithms. The labeling accuracy is high, and each radar scan is meticulously labeled and calibrated to ensure accurate and reliable annotations.

3.3 CARRADA

The French researchers initially released this database in 2020, with a new version being released in 2021. The data collection setup includes one FMCW radar and one camera installed on a fixed vehicle. By utilizing a MIMO system setup with 2Tx and 4Rx, the radar system generates eight virtual antennas in total. The dataset consists of synchronized image and radar data for 30 sequences, comprising a total of 12,666 frames (equivalent to 21.1 minutes). Out of these frames, 7,193 frames contain labeled objects. To label the radar signals, the dataset employs a semi-automatic labeling method that relies on visual and physical information. This approach significantly reduces labeling time and cost. The labeled objects are separated into three types: cars, bikes, and pedestrians. The dataset offers three labeling formats: sparse point, bounding box, and dense mask. The radar data underlying CARRADA is presented in the Range-Angle-Doppler Tensor format. However, it is important to note that CARRADA documents the Canadian acquisition scenario on a test track, featuring one or two objects in the scenario simultaneously along different paths rather than real traffic road conditions. Therefore, its practicality may be somewhat affected.

3.4 VOD

The VOD [71] dataset uses 4D MMW imaging radar to provide height information, as well as distance, bearing, and Doppler velocity. It consists of 8,693 time-synchronized and calibrated frames of 64-line lidar, a binocular camera, and 4D radar data. The dataset includes 123,106 three-dimensional bounding boxes for both moving and stationary objects, with a total of 26,587 pedestrians, 10,800 cyclists, and 26,949 vehicles.

Figure 10 displays the configuration of the VOD acquisition device. The sensor configuration for the VOD data gathering system includes a ZF FRGen21 3+1D radar (see Table II for details, operating at approximately 13 Hz) installed behind the front bumper, a windshield-mounted stereo camera (1936×1216 px, operating at around 30 Hz), a roof-mounted Velodyne HDL-64 S3 lidar (operating at approximately 10 Hz) scanner, along with the ego vehicle's odometry (a filtered mix of wheel odometry, real-time kinematic (RTK) GPS, and IMU, operating at around 100 Hz).

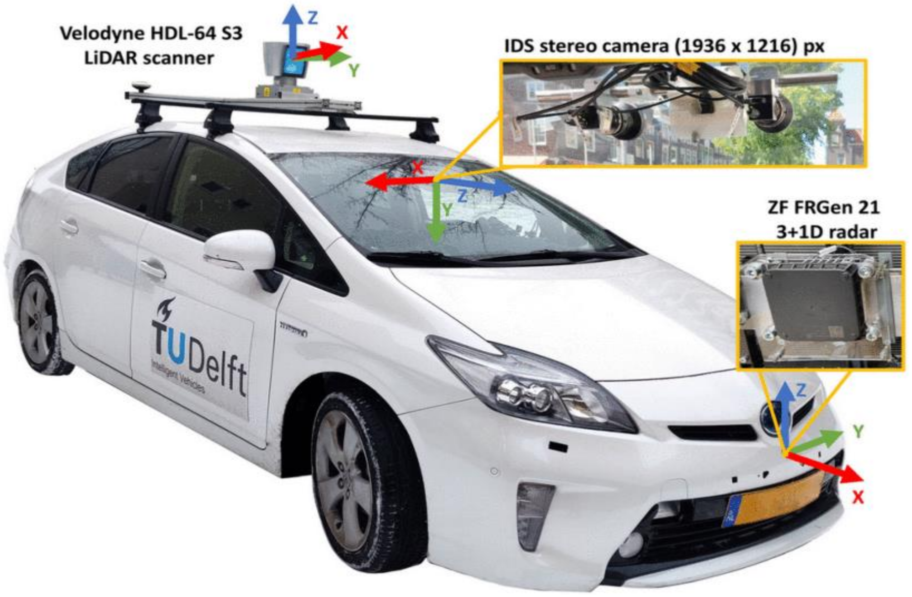


Figure 10. Sensor setup for VOD data collection platform [71]

At present, the VOD dataset is the mere publicly available autopilot dataset containing lidar, 4D MMW radar, and binocular camera information. Nevertheless, it owns relatively small quantities of data. High-quality datasets for millimeter wave radar-camera integration for target detection are limited in availability. Currently, most researchers rely on nuScenes dataset to verify and enhance fusion algorithms. The nuScenes dataset provides millimeter wave radar signals in the shape of point cloud signals, which has led to a focus on processing techniques for point cloud data. Developing better datasets is a future research direction for improving sensor fusion training data. However, obtaining comprehensive and larger datasets is a hard assignment due to the difficulties in labeling objects. Manual labeling is time-consuming, energy-intensive, and expensive, and the accuracy of labeling depends on human operation. Some datasets are starting to utilize automatic labeling methods, but their effectiveness is not as strong as manual labeling. As a result, constructing a dataset with a substantial amount of data is challenging.

4. MMW RADAR AND CAMERA FUSION ALGORITHM

Merging MMW radar with a camera aims to combine the benefits of each sensor, improving the accuracy of target detection. Integrating radar and camera sensors provides additional details about the object, such as color, shape, distance, velocity, and orientation. Additionally, the fusion of radar and camera allows self-driving cars to operate day and night, even in unfavorable weather conditions.

Multiple studies have shown that merging radar and cameras enhances the precision and resilience of algorithms in complex city traffic situations. Chadwick et al. [72] integrated a radar sensor with short-focus and long-focus cameras to improve the identification of distant objects. The radar sensor provides physical data about the movement of these objects, thereby enhancing camera detection efficiency. Major et al. [13] also demonstrated the velocity dimensions obtained from radar sensors can improve detection performance. Furthermore, Nabati et al. [43] used radar characteristics such as depth, rotation, and velocity to enhance image attributes, resulting in a more than 12% increase in the total nuclear scene detection score (NDS) compared to algorithms that rely solely on cutting-edge cameras, including OFT [73]. Other algorithms considered include MonoDIS [74] and CenterNet [75]. Yadav et al. [76] proposed the development of RANet and BIRANet for challenging weather scenarios like fog, dust, and rain and found that radar data is highly resistant to noise in detection. Incorporating radar information can improve efficiency in these demanding situations.

The primary difference between radar and visual fusion mechanisms lies in the fusion degree and the synchronization or asynchronous nature of their processing methods. Alessandretti et al. [77] categorized the fusion intensity into three types: high, medium, and low. Figure 11 illustrates the process of fusion at three different levels.

- The first level of data fusion is low-level fusion, which combines data detected by MMW radar and camera with minimal data loss and maximum reliability. This process generates new raw data by combining multiple raw data sources.
- As an intermediate level fusion, feature level fusion involves extracting feature details from radar and images, such as speed, distance, corners, edges, lines, position, and texture parameters, which are then integrated into a characteristic map. Subsequently, this map needs further treating.
- Advanced level fusion, or decision level fusion, involves making a decision from each input source and subsequently merging all these decisions. This results in a fusion of detection results.

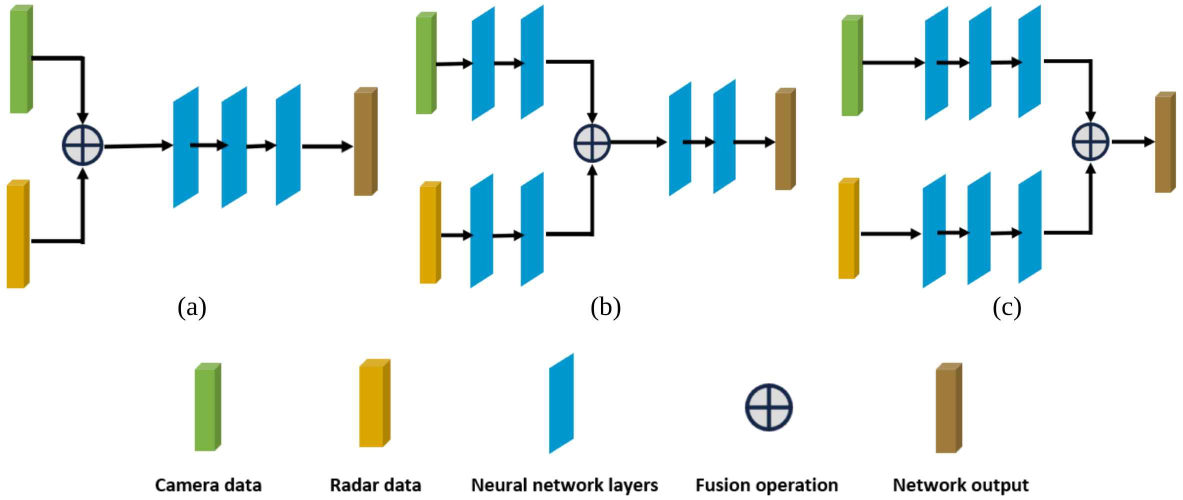


Figure 11. Classification of various levels of radar camera fusion. (a) Data level fusion, (b) Feature-level fusion, (c) Object level fusion

4.1 Data Level Fusion

Currently, data-level fusion is not a popular area of research. This method involves merging data from MMW radar and cameras, known as pixel-level fusion. In this process, radar point clouds and image pixels are directly combined without preprocessing. The goal of this technique is to generate new raw data that is enriched with valuable insights by integrating MMW radar and camera information. To achieve this, coordinates of radar information are mapped onto pixel images and aligned with these image pixels calibrated. Then, the resulting fused data is vulnerable to characteristic extraction and categorization.

In the initial phase of the deep learning model, either unprocessed or processed data from radar and camera sensors are merged for data-level integration. Since the merging of MMW radar and camera data usually happens asynchronously at the individual data level, the process of data fusion requires the use of filters to interpolate and calibrate the data over time. The data from multiple sensors are then correlated to associate them with different targets using various efficient classification algorithms. However, aligning radar tensor or point cloud with image pixels presents challenges due to differences in data representation and object form. Nobiset et al. [78] used integrated camera and radar point clouds while employing VGG [79] for feature extraction from the merged data, taking inspiration from the fusion of lidar and camera. While they achieved better detection results, the accuracy of detection was still affected by noisy radar raw data. Bansaet et al. [80] developed one semantic point grid (SPG) by integrating radar point cloud, radar BEV grid map, and camera semantic map. The approach uses SPG coding to derive semantic data from the camera, which helps identify radar locations associated with the targeted object. Long et al. [81] put forward the radar camera pixel depth association (RC-PDA) as a learning technique to improve and compress radar imagery. However, the camera is not effective in bad weather, and the fusion network lacks a system to integrate camera data.

4.2 Feature Level Fusion

Recently, the feature-level fusion method [72], [82] has become popular as a fusion technique. Typically, fusion methods at the feature level convert 3D radar data into 2D imagery. In the modified radar image, the radar points represent depths and velocities, which are recorded as pixel values. The system consists of multiple channels, each reflecting a distinct physical condition of the environment, as detected by the radar sensor. Consequently, in the same driving scenario, it is possible to capture two types of images: one radar-based and the other visual.

The strength of this approach is that it swiftly removes numerous regions without targets, significantly improving recognition speed. Additionally, the algorithm can swiftly eliminate false objects explored by the radar, therefore improving result credibility. Nonetheless, inaccuracies in target lateral distance and camera calibration errors cause the projection point of the MMW radar to diverge from the object. This deviation is more significant as the interest setup region encompasses multiple objects, leading to repeated target detection and confusion in target matching.

When creating a multichannel matrix, feature matrices are combined. In [72], ResNet [83] blocks are employed to produce the characteristics of the radar and camera branches, which are then integrated through concatenation and added operations. In literature [84], the tandem method is employed, while in literature [85], a new block for sensor feature fusion called spatial attention fusion (SAF) is proposed. Using the SAF block, an attention weighting matrix is developed, integrating both radar and visual elements. The researchers compared the SAF technique to three other approaches: addition at the element level, multiplication, and cascade. The findings suggest that the SAF technique outperforms its competitors. Furthermore, the research conducted experiments to generalize FAST R-CNN, with the SAF model enhancing its detection effectiveness. In 2021, Du et al. [86] proposed a 3D target detection algorithm. The algorithm begins by isolating image characteristics from an individual image. It then proceeds with the 3D enlargement of the radar point cloud and merges the algorithm's 3D data to align with the respective radar point cloud. The construction of radar characteristics utilizes the radar's depth and velocity data. Lastly, image features and radar features are fused in series to generate a precise 3D envelope of the object.

In 2019, Nabati et al. [43] suggested a method of intermediate fusion for detecting 3D targets using radar and camera information. The radar point cloud was processed using a similar approach to that of [86], which added height information missing from the radar to the point cloud. The authors used a CenterFusion network to detect the target by recognizing regions of interest (ROI) in the image as well as obtaining the target's 3D coordinates, depth, and rotation. They employed an approach derived from the frustum association technique to link radar detection data to the center of mass of the detection target. This linking technique incorporates the characteristics of the target identified by radar detection, including depth and velocity data, into a feature map. This technique enhances the visual characteristics and provides desired attributes such as depth, rotation, and speed.

In 2023, Li et al. [87] proposed the RCFN, which utilizes a CNN-based triple decoder architecture. For the image branch, the researchers modified the ResNet-34 network to satisfy the demands of characteristic extraction for further estimation. A dual-phase processing approach was employed for the radar segment. The process began by extracting features using a sparse invariant convolutional block, followed by additional feature extraction using the identical residual block as in the image branch but with a different count of convolutional channels. These derived features were then inputted into three separate decoders, each forecasting distinct categories of depth data. Compared to the standard single decoder model, the authors' forecasts showed a range of enhancements from slight to significant.

In 2023, Kurniawan et al. [88] proposed the ClusterFusion fusion network, which is known for its ability to implement cross-modal characteristic fusion within the image plane. This network directly extracts features onto the radar clusters as a point cloud, preserving the clusters' local spatial characteristics that would otherwise be obscured in the projection phase. Employing this method enhances the precision of feature extraction. ClusterFusion achieves attribute estimation comparable to advanced 3D object detection techniques using a monocular camera. In 2023, Pang et al. [89] developed TransCAR, a new camera-radar fusion network. This network is comprised of three main elements: a transformer decoder-based camera network, a radar system encoding radar point positions and deriving radar characteristics, and the TransCAR integration module utilizing three cross-attention decoders based on the transformer. The authors used transformer-based architectures to achieve this. The TransCAR architecture enables adaptive correlation between cameras and radar, as well as soft-associative learning of radar features with visually updated object queries, resulting in superior 3D object detection performance.

When it comes to characteristic-level integration, it has the potential to create suitable characteristic extraction networks tailored for each modality, considering their unique attributes. Neural networks have the ability to collectively capture characteristics across different modalities, making them mutually complementary. However, feature extraction and fusion are unable to address situations where the image sensor becomes unreliable [90].

4.3 Decision Level Fusion

Decision-level fusion is a widely used integration scheme in self-driving vehicles. The MMW radar and camera are capable of independently sensing and processing their data to obtain initial sensing results. The image detection target results are then effectively fused with the MMW radar detection results. Jha et al. [91] mapped the radar findings on the image plane through one transformation matrix and organized the separately identified objects from both sensors. Dong et al. [92] put forward the AssociationNet to acquire semantic representation data from both sensors, improving the precision of associations through the computation and reduction of Euclidean distances. Object-level fusion is pervasive in traditional radar and camera systems due to its adaptability and modularity, but it sacrifices rich feature information [90].

The fusion strategy offers several advantages, including flexibility in selecting sensor results, improved data reliability and fault tolerance, enhanced flexibility across diverse sensor sources, better immediate performance, and lower communication bandwidth requirements. However, the intense compression of information may reduce precision and require extensive preprocessing capacity. The accuracy of the final result depends heavily on the precision of the individual module outputs. For example, if the camera sensor is obstructed, the fusion at the decision level will solely rely on the ultimate object identified by the radar sensor. Additionally, mid-level characteristics are often overlooked due to flaws or inaccuracies in the sensors' detection techniques. Therefore, decision-level fusion methods are limited to the scant data derived from the detection outcomes.

Integrating MMW radar with camera fusion at the data layer offers multiple benefits. This method maintains data integrity, provides in-depth insights not achievable through other fusion layers, and demonstrates a strong correlation among the data. However, it fails to preprocess the initial data, resulting in excessive redundancy, subpar real-time performance, weak jamming resistance, elevated communication bandwidth requirements due to the large amount of original data, and increased demands for error rectification at the basic level of information merging.

Feature-level fusion utilizes neural networks to learn complementary properties between different sensors, adding richer information to the object. For example, in the fusion of MMW radar and camera features, fusion is capable of being performed based on the image, incorporating information such as speed, angle, and distance provided by the MMW radar into the image features. Fusion can also be performed based on the point cloud, where features such as the contour of the image, color, and texture are added to the radar point cloud for correlation. Objects have richer features that can be effectively recognized by neural networks. Feature-level fusion combines information from different sensors, retaining data information better and improving target recognition and tracking. Additionally, feature extraction and fusion can suppress noise and interference in sensor data, enhancing system robustness. However, the feature-level fusion process requires addressing differences in characteristics between the two sensors because of feature incompatibility. Moreover, the design of the feature fusion algorithm is relatively complex, requiring careful consideration of feature selection, extraction, and fusion methods.

5. TARGET DETECTION ALGORITHM EVALUATION METRICS

Radar camera fusion often uses evaluation metrics such as recall, precision, average precision (AP), average recall (AR), mean average precision (mAP), and mean intersection over union (mIoU). The measurements determine how accurate predictions are for a specific test dataset. More importantly, the measurements should be applied objectively, without any subjective judgments. Currently, the nuScenes [6] dataset is the main choice for evaluating algorithms in MMW radar and camera fusion, introducing the mean Average Precision (mAP), mean True Positive (mTP) metrics, and the nuScenes Detection Scores (NDS). For the object detection model, the assessment metrics include mAP and mTP. mAP is the average value of Average Precision (AP) for all classes at different center distance thresholds. On the other hand, mTP metrics consist of five sub-metrics: average translation error (ATE), average scale error (ASE), average orientation error (AOE), and average true positive (mTP). The NDS is calculated by weighting the sum of the five metrics: ATE, ASE, AOE, average velocity error (AVE), and average attribute error (AAE), along with mAP and mTP.

5.1 Average Precision (mAP)

AP calculation in object detection uses the center distance (CD) thresholding technique but not the widespread intersection over union (IOU) target detection matching algorithm. This approach separates the detection process from the object's dimensions and alignment, resulting in more accurate outcomes. The mean intersection over union (mIoU) is calculated as the average IoUs across all categories. It is important to note that for objects with minimal spatial presence, the IoU result may be zero if there is little translation error during detection. This complicates the comparison of detection algorithms that rely solely on visual data. For a specific category, the AP value is obtained by calculating the precision and recall of the target detection.

Table 2. Confusion matrix

Confusion matrix		Projections	
		Positive	Negative
current situation	Ture	TP	FN
	False	FP	TN

In classification problems, the confusion matrix is often used as an assessment tool. It shows the numerical correlation between the predicted and actual samples within a category. For a problem with N classes, the confusion matrix has dimensions of N^2. If a binary classification issue occurs, Table 2's confusion matrix displays true positive (TP) for predicting a positive scenario, false negative (FN) for a negative scenario, false positive (FP) for a positive scenario, and false negative (FN) for a negative scenario. A positive result indicates a sample incorrectly identified as positive, while TN denotes a sample accurately identified as negative. TP and TN represent accurate predictions, while FN and FP represent inaccurate predictions. Therefore, precision and recall can be computed as follows:

$$Precision = \frac{TP}{TP + FP} \tag{1}$$

$$Recall = \frac{TP}{TP + FN} \tag{2}$$

In model theory evaluation, better model performance is indicated by higher precision and recall. However, in practical prediction situations, these two factors are usually negatively correlated. Precision tends to decrease as recall increases. To better illustrate this relationship, the Precision-Recall (PR) curve is proposed. This graph shows the rate of recall on the horizontal axis and accuracy on the vertical axis. Unlike other detection models that use the area beneath the precision-recall curve as an AP value, a higher mAP indicates superior model performance.

$$mAP = \frac{1}{|C||D|}\sum_{c\epsilon C}\sum_{d\epsilon D} A\,P_{c,d} \tag{3}$$

5.2 Mean True Positive (mTP)

The TP metric series consists of accurate measurements for each predictive frame aligned with a reference frame. The specific metrics in the TP series are as follows: ATE measures the two-dimensional Euclidean distance from the center in meters. ASE is calculated as (1-IOU), using the 3D IOU after aligning the direction and center. AOE is the minimum deviation in yaw angle between the predicted outcome and the actual reality. AVE represents the total velocity discrepancy derived from the L2 norm of the two-dimensional velocity variance in meters per second. AAE is defined as (1-acc), where acc represents the accuracy of attribute classification. The last two metrics are not applicable for obstacle and traffic cone detection. The goal is to minimize the error metrics in the TP series. Class-specific metrics (mATE, mASE, mAOE, mAVE, and mAAE) are determined by averaging the values from each category. To determine the mean value of each TP metric across all categories in the study, follow these procedures:

$$mAP = \frac{1}{C}\sum_{c\epsilon C} TP_c \tag{4}$$

5.3 nuScenes Detection Scores (NDS)

Target detection often uses the mAP combined with the IOU threshold as a standard evaluation metric. However, this metric may not be suitable for all aspects of nuScenes target identification, such as velocity and attribute estimation. Moreover, it combines the contributions of position, size, and orientation to the detection results. To address these limitations, the nuScenes detection score (NDS) is introduced as a more comprehensive evaluation metric. It integrates the performance of target detection algorithms in different aspects, including accuracy, robustness, and efficiency. The NDS aims to assess the effectiveness of an algorithm by determining the rates of recall and false alarms in detecting targets across various error intervals. These rates are then adjusted based on the size and distance of the target to derive an all-encompassing score, computed using the following formula:

$$NDS = \frac{1}{2}\left[5mAP + \sum_{mTP\epsilon TP}\left(1 - min(1, mTP)\right)\right] \tag{5}$$

mTP was obtained using Eq. (3). TP represents the set of five average TP metrics. NDS weights were split in half to detect model performance and quantify detection quality based on bounding box location, size, orientation, attributes, and speed. It is vital to limit each metric to a range of 0 to 1, as mAVE, mAOE, and mATE values may exceed 1.

5.4 Planning KL-Divergence (PKL)

PKL [93] is an evaluation metric that measures perceptual performance by calculating the difference between the planning approach given detector detection and manually labeled detection. The results were consistently non-negative, where higher PKL scores signify inferior detection capabilities. A PKL score of 0 corresponds to the optimal detector. This technique assesses the variance in the ego car's planning actions upon encountering a forecasted object, as opposed to the actual object present in the scene. The sequence of raw sensor observations, S_1,…,S_t, ∈S, corresponds to the ground truth object detection sequences, O_1^*,…, $O_t^* \in O$, and the ego car attitude of the corresponding sequence is represented by x_1,…, x_t , Given an object detector A: S → O that predicts O_t^* □ based on S_t, we define the Planning KL-Divergence (PKL) at time t.

$$PKL(A) = \sum_{0<\Delta\leq T} D_{KL}\left(p_\theta\left(x_{t+\Delta} \mid o^*_{\leq t}\right) \,||\, p_\theta\left(x_{t+\Delta} \mid A(s_{\leq t})\right)\right) \tag{6}$$

The distribution of ground truth trajectories in dataset D was modeled by $p_\theta(x_t \mid o^*_{\leq t})$. The parameter θ is minimized by:

$$\theta = \underset{\theta}{argmin} \sum_{x_t \in D} - log p_\theta(x_t | o^*_{\leq t}) \tag{7}$$

In summary, PKL quantifies the differences between expected and actual objects and their impact on the trajectory planning of the autopilot. The significance of this metric lies in its direct relation to the decision-making process of the autopilot system, which is based on detection results.

Table 3 provides a summary of the performance of the targeting algorithms for MMW radar-camera fusion, evaluated in the nuScenes dataset from 2020 to 2023, using the aforementioned evaluation parameters. The table illustrates that the performance of different algorithms is not optimal across all evaluation parameters. Rather, it is only in certain evaluation parameters that the performance is deemed excellent.

Table 3. Performance of MMW radar and camera fusion in detecting targets on the nuScenes from 2020 to 2023

Method			Metrics							
Author	Data	Name of the algorithm for fusing radar and camera	mAP	mATE (m)	mASE (1-IOU)	mAOE (rad)	mAVE (m/s)	mAAE (1-acc)	NDS	PKL
Kim et al.	2023/11/02	RCM-Fusion [94]	0.506	0.465	0.254	0.384	0.438	0.121	0.587	0.896
Kurniawan et al.	2023/07/22	ClusterFusion [88]	0.341	0.587	0.257	0.424	0.461	0.108	0.487	1.334
Z. Chen et al.	2023/05/29	HVDetFusion [95]	0.609	0.379	0.243	0.382	0.172	0.132	0.674	0.836
Kim et al.	2023/03/07	CRN [96]	0.575	0.416	0.264	0.456	0.365	0.130	0.624	0.929
S. Pang et al.	2022/11/10	TransCAR [89]	0.422	0.63	0.26	0.383	0.495	0.121	0.522	1231
Kim et al.	2022/07/17	CRAFT [97]	0.411	0.467	0.268	0.456	0.519	0.114	0.523	1.145
Nabati et al.	2020/09/28	CenterFusion [43]	0.326	0.631	0.261	0.516	0.614	0.115	0.449	1.446

CenterFusion [43] is the first algorithm to combine MMV radar and camera data from nuScenes dataset. It achieved performance results of 32.6% mAP and 44.9% NDS. This algorithm addresses the problem of correlating critical data by introducing a novel frustum-based approach to align radar detections with the central points of the corresponding objects. This technique uses initial detection to formulate a 3D ROI frustum near the object. Then, it aligns the radar detection with the object's center in the image. Center Fusion, which incorporates radar features, shows a relative increase of 38.1% and 62.1% in NDS and speed error metrics, respectively, compared to CenterNet [98], which relies solely on image inputs. This highlights the efficiency of adopting radar characteristics and the robustness of merging radar and camera within a standard self-driving scenario. As a result, many radar-camera integration algorithms have used CenterFusion as a benchmark.

As an example, the RCBEV technique for merging features at the feature level [99] uses a spatio-temporal encoder to separate radar characteristics and transform image attributes into a representation of BEV. Experimental results show improved characteristic presentation and accuracy of 3D object detection, as evidenced by mAP and NDS rates of 40.6% and 48.6%, respectively.

CRAFT [97] achieves a mAP of 41.1% and an NDS of 52.3% in nuScenes, mainly due to improved localization and velocity measurement. This improvement is supported by a spatial-context fusion converter that leverages the contextual and spatial features of camera and radar information for more precise object detection in 3D space. Currently, CRN [96] stands out as the top detector in the nuScenes dataset among radar-camera fusion techniques, boasting a 57.5% mAP and a 62.4% NDS. This positions CRN as the premier choice for 3D radar-camera fusion. The suggested CRN system achieves enhanced performance through its radar-aided view transform (RVT). By converting features of perspective images into bird's-eye views (BEVs) using infrequent yet precise radar data, this method addresses the spatial data deficiency of the image. After transformation, the image attributes in BEVs are used in the Multimodal Feature Aggregation (MFA) layer to create BEV representations that are both semantically profound and spatially precise.

To address low visibility challenges, REDFormer [100] integrates features of multiple radar points and cameras in the bird's-eye view (BEV) plane of nuScenes dataset. In scenes with low visibility, the model's performance significantly surpasses the baseline model, achieving 50. During wet weather conditions, the algorithm outperforms the standard model with a 50.91% NDS and a 40.36% mAP. At night, it reaches 28.12% NDS and 20.28% mAP.

The HVDetFusion [95] 3D object detection algorithm currently holds the top rank in the nuScenes Camera Radar-based 3D object detection ranking, showing an NDS score of 67.4% and an mAP score of 60.9%. It utilizes a structurally optimized and improved detection method based on Bevdet4D, effectively extracting data from one or more camera sensors in keyframes. Furthermore, it enhances and integrates top view characteristics derived from unprocessed camera data, which are based on positional and radial velocity data captured by radar sensors. This leads to high accuracy in detection.

6. RECOMMENDATIONS FOR FUTURE RESEARCH

The integration of MMW radar and camera systems presents significant challenges, mainly related to the selection of an appropriate fusion methodology and the efficient processing of different data types to improve overall performance. Neglecting these challenges could seriously hinder the decision-making and control mechanisms necessary for auto vehicles to operate safely. This section explores the complex issues and potential research directions related to multimodal data fusion frameworks, which are crucial for advancement.

6.1 Challenges Encountered while Labeling Data

Labeling data is an expensive and time-wasting assignment, especially when dealing with data from multiple sources. This difficulty becomes even more apparent when working with radar-camera fusion, as radar data alone does not provide a direct indication of an object's physical appearance. Researchers have explored the possibility of automating the labeling process of radar data by using actual data from camera photos and a conversion matrix that links radar and camera sensors. However, this labeling technique has proven to be imprecise due to the misalignment of radar targets and the objects in the image. Sengupta et al. [101] raised one camera-assisted approach that used a beforehand YOLOv3[47] network and Hungarian algorithm to automatically label radar point clouds, improving effectiveness and precision. Although there are possible benefits of automating radar data labeling, the challenge of removing irrelevant data surrounding the targeted object remains.

When labeling camera images, it is essential to carefully select appropriate labeled data to minimize labor costs. Supervised learning [102] is widely recognized as an effective method for training deep learning models. It involves using small quantities of labeled image data along with vast quantities of unlabeled image data. The process begins with training the model using pre-labeled data and then using this model to predict and label the unlabeled data. This approach reduces the amount of manual labor required for labeling. Additionally, strategies, including transfer learning [103] and semi-supervised learning [104], can be employed to further reduce the effort involved in labeling.

6.2 Multi-Modal Data

In the fusion of MMW radar and camera data, there is a clear disparity between the nature of information extracted from each sensor. While image data from cameras typically exhibits regularity and structure, radar-generated point cloud data tends to be disordered. Consequently, the integration process poses a significant challenge, with radar data processing being particularly formidable.

The sparse nature of radar clouds presents a risk to neural networks, with the aim of researching their characteristics efficiently. Since the point clouds often lack a comprehensive representation of object shape, color, and other pertinent details, conventional bounding box approaches are considered suboptimal. Researchers often compile data from various radar frames, ranging from 0.25 to 1 second, to tackle sparsity problems. This increases the density of point clouds and improves detection precision [43], [71], [78], [81]. However, overlaying multi-frame point clouds may lead to latency within the system. Consequently, modern studies are increasingly focusing on the potential of 4D MMW radar sensors as a significant research path. These sensors have the advantage of creating more compact point clouds and providing essential altitude data on identified objects, making it easier to align with numerous points on the vehicle. Notably, datasets such as Astyx [105], VoD [71], and TJ4DRadSet [106] demonstrate the extensive spatial coverage provided by 4D radar datasets. Research consistently highlights the benefits of using 4D radar in object detection. For example, Zheng et al. [106] showcased the enhanced capabilities of the 4D radar in 3D perception, attributed to its more compact point arrangement. Furthermore, Palffy et al. [71] emphasized in their VoD dataset that adding height details markedly improves performance in detecting objects.

For the successful integration of multimodal data in deep learning, a substantial amount of training data is essential for strong backing. However, current multimodal datasets that integrate radar and camera data are considerably smaller in scale compared to those that only use images. The ImageNet [107] dataset, currently the largest image-based dataset, encompasses over 14 million images across 20,000 categories. In contrast, the CRUW [69], the most extensive radar-camera fusion dataset, consists of a modest 40,000 frames with 26 objects spread across just three categories, mainly centered on vehicles. Furthermore, developing an extensive multimodal dataset requires meticulous evaluation of authentic driving situations, covering varied settings such as highways, urban roads, and countryside routes. Additionally, the data collection needs to consider intricate meteorological scenarios, such as precipitation, mist, snowfall, and brightness. Creating this type of dataset requires a significant amount of time and workforce to maintain its accuracy and relevance.

6.3 Correlation of Camera and Millimeter Wave Radar Data

Integrating radar and images poses a significant challenge due to their distinct characteristics. One common method is releasing the radar point cloud to the image plane and adjusting data through one calibration matrix. Nonetheless, this projection frequently leads to bad alignment with the core of the object, which makes it hard to map the radar and image information accurately. Nabati and Qi [11] introduced the Radar Proposal Refinement (RPR) network, which aims at matching real situations from radar and camera sources. In fact, they deeply improved the alignment process in CenterFusion[43] by incorporating detection frames and riser extensions, enabling the accurate mapping of radar data to the object's center and resolving overlap issues. Bansal et al. [80] brought forward a presentation named Semantic Point Grid (SPG), which combines semantic data from camera graphs with radar point clouds. This allows for the identification of radar points that correspond to individual camera pixels. From our perspective, a possible approach to connect radar and image data is through attention-driven adaptive threshold correlation. For example, in [81], the RC-PDA technique was introduced to eliminate hidden radar echoes and improve the depth map of the radar projections by creating confidence levels for these associations.

Furthermore, the BEV approach is gradually being employed. The BEV approach provides a comprehensive view from above, which can overcome the limitations of 2D perception, including occlusion and scaling issues. This considers a more detailed comprehension of the surrounding environment and offers a more accurate presentation of vehicle position within it. The conversion of data from disparate sensors into a unified BEV view enables the integration of information in a more efficient manner, thereby facilitating the provision of more accurate environmental perception.

6.4 Model Robustness

A major challenge within the fusion framework is ensuring the stability of the target detection model, especially in situations where sensor signals become unreliable or when an autonomous vehicle encounters impaired visibility. While many target detection methods focus on precision in standard datasets, there are only a few scenarios where a single sensor serves as the sole input. In the case of RadSegNet [80], the SPG coding method autonomously extracts information from both the camera and radar sources. By translating the semantic information from the camera image into a radar point cloud, the SPG coding method relies solely on radar data to ensure reliable functioning in situations where camera input may be unreliable. Additionally, incorporating an attention mechanism proves to be an effective strategy for handling diverse data from multiple sensors. This system facilitates the integration of characteristics from various modalities and aids in the analysis of unprocessed features from a single modality. In our opinion, the unresolved issue lies in establishing a cross-channel within the fusion framework to maintain the model's resilience in different scenarios, underscoring the need for further investigation and the development of solutions.

7. CONCLUSION

The article explores the growing interest in camera-MMW radar fusion as a cost-effective and all-weather solution for smart transportation and autonomous driving technologies. It provides an overview and analysis of research on radar-camera fusion in target detection tasks. The paper examines the importance of combining radar and camera technology in the perception of self-driving vehicles, beginning with fundamental concepts of radar and camera detection. Various fusion methods are thoroughly studied, and their advantages and disadvantages are discussed. Drawing from existing datasets and methods for MMW radar-camera fusion, this text examines the key issues of the widely-used BEV perception fusion and multimodal fusion and proposes potential research directions. Radar-camera integration now leans towards data formats that offer detailed information. Representations, including ADC signals and radar tensors, provide additional raw data, facilitating multimodal fusion. However, the latest 4D radar sensors produce more compact point clouds, improve resolution, and include altitude data, offering broader insights for self-driving vehicles. Notably, the purpose of this survey is to provide comprehensive instruction for researchers and practitioners to foster valuable insights into radar-camera integration.

ACKNOWLEDGEMENTS

The authors are grateful to the Ministry of Higher Education Malaysia for the financial support provided under the FRGS/1/2023/TK10/UMP/02/2 grant (RDU230102). The authors also acknowledge the support of the Universiti Malaysia Pahang Al-Sultan Abdullah through the Postgraduate Research Grants Scheme (PGRS220328).

CONFLICT OF INTEREST

The authors declare that they do not have any conflict of interest.

AUTHORS CONTRIBUTION

Tao Zhenhua (Conceptualisation, Methodology, Data curation, Formal analysis, Writing - original draft)
Ngui Wai Keng (Writing—review & editing, Funding acquisition, Supervision)

REFERENCES

[1] Y. Huang, J. Du, Z. Yang, Z. Zhou, L. Zhang, and H. Chen, "A survey on trajectory-prediction methods for autonomous driving," *IEEE Transactions on Intelligent Vehicles*, vol. 7, no. 3, pp. 652–674, 2022.

[2] S. Teng, X. Hu, P. Deng, B. Li, Y. Li, Y. Ai et al., "Motion planning for autonomous driving: The state of the art and future perspectives," *IEEE Transactions on Intelligent Vehicles*, vol. 8, no. 6, pp. 3692–3711, 2023.

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

[4] D. Feng, C. Haase-Schütz, L. Rosenbaum, et al., "Deep multi-modal object detection and semantic segmentation for autonomous driving: Datasets, methods, and challenges," *IEEE Transactions on Intelligent Transportation Systems*, vol. 22, no. 3, pp. 1341–1360, 2021.

[5] 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*, Long Beach, CA, USA: IEEE, pp. 7345-7353. 2019.

[6] H. Caesar, V. Bankiti, A. H. Lang, S. Vora, V. E. Liong, Q. Xu et al., "nuScenes: A multimodal dataset for autonomous driving," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Seattle, WA, USA: IEEE, pp. 11621-11631, 2020.

[7] H. Liu, C. Wu, and H. Wang, "Real time object detection using Lidar and camera fusion for autonomous driving," *Scientific Reports*, vol. 13, no. 1, p. 8056, 2023.

[8] 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*, Honolulu, HI: IEEE pp. 1907-1915, 2017.

[9] 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*, Madrid: IEEE, pp. 1-8, 2018.

[10] B. Li, T. Zhang, and T. Xia, "Vehicle detection from 3D lidar using fully convolutional network." arXiv, Aug. 29, 2016. Accessed: Jul. 23, 2023. [Online]. Available: http://arxiv.org/abs/1608.07916

[11] R. Nabati and H. Qi, "RRPN: Radar Region Proposal Network for object detection in autonomous vehicles," in *2019 IEEE International Conference on Image Processing (ICIP)*, Taipei, Taiwan: IEEE, pp. 3093–3097, 2019.

[12] P. Li, P. Wang, K. Berntorp, and H. Liu, "Exploiting temporal relations on radar perception for autonomous driving," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, New Orleans, LA, USA: IEEE, pp. 17071-17080, 2022.

[13] B. Major, D. Fontijne, A. Ansari, R.T. Sukhavasi, R. Gowaikar, M. Hamilton et al., "Vehicle detection with automotive radar using deep learning on range-azimuth-doppler tensors," in *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, Seoul, Korea (South): IEEE, pp. 924-932, 2019.

[14] J. Vargas, S. Alsweiss, O. Toker, R. Razdan, and J. Santos, "An overview of autonomous vehicles sensors and their vulnerability to weather conditions," *Sensors*, vol. 21, no. 16, p. 5397, 2021.

[15] Z. Wei, F. Zhang, S. Chang, Y. Liu, H. Wu, and Z. Feng, "MmWave radar and vision fusion for object detection in autonomous driving: A review," *Sensors*, vol. 22, no. 7, p. 2542, 2022.

[16] E. Arnold, O. Y. Al-Jarrah, M. Dianati, S. Fallah, D. Oxtoby, and A. Mouzakitis, "A survey on 3D object detection methods for autonomous driving applications," *IEEE Transactions on Intelligent Transportation Systems*, vol. 20, no. 10, pp. 3782–3795, 2019.

[17] B. Shahian Jahromi, T. Tulabandhula, and S. Cetin, "Real-time hybrid multi-sensor fusion framework for perception in autonomous vehicles," *Sensors*, vol. 19, no. 20, p. 4357, 2019.

[18] "Image sensor," Wikipedia. Mar. 24, 2023. Accessed: Aug. 14, 2023. [Online]. Available: https://en.wikipedia.org/w/index.php?title=Image_sensor&oldid=1146373856

[19] S. Saponara, M.S. Greco, and F. Gini, "Radar-on-Chip/in-package in autonomous driving vehicles and intelligent transport systems: Opportunities and challenges," *IEEE Signal Processing Magazine*, vol. 36, no. 5, pp. 71–84, 2019.

[20] A. Soumya, C. Krishna Mohan, and L.R. Cenkeramaddi, "Recent advances in mmWave-radar-based sensing, its applications, and machine learning techniques: A review," *Sensors*, vol. 23, no. 21, p. 8901, 2023.

[21] Z. Wang, Y. Wu, and Q. Niu, "Multi-sensor fusion in automated driving: A survey," *IEEE Access*, vol. 8, pp. 2847–2868, 2020.

[22] "Detecting static objects in view using—electrical engineering stack exchange," [Online]. Available: https://electronics.stackexchange.com/questions/236484/detecting-static-objects-in-view-using-radar

[23] M. Cho, "A study on the obstacle recognition for autonomous driving RC car using lidar and thermal infrared camera," in *2019 Eleventh International Conference on Ubiquitous and Future Networks (ICUFN)*, Zagreb, Croatia: IEEE, pp. 544–546, 2019.

[24] S. Sun, A. P. Petropulu, and H. V. Poor, "MIMO radar for advanced driver-assistance systems and autonomous driving: Advantages and challenges*," IEEE Signal Processing Magazine*, vol. 37, no. 4, pp. 98–117, 2020.

[25] O. Schumann, M. Hahn, J. Dickmann, and C. Wohler, "Semantic segmentation on radar point clouds," in *2018 21st International Conference on Information Fusion (FUSION)*, Cambridge: IEEE, pp. 2179–2186, 2018.

[26] F. Fent, P. Bauerschmidt, and M. Lienkamp, "RadarGNN: Transformation invariant graph neural network for radar-based perception," in *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, Vancouver, BC, Canada: IEEE, pp. 182–191, 2023.

[27] S. Rahman and D. A. Robertson, "Radar micro-Doppler signatures of drones and birds at K-band and W-band," Scientific Reports, vol. 8, no. 1, p. 17396, 2018.

[28] A. Ouaknine, A. Newson, J. Rebut, F. Tupin, and P. Pérez, "CARRADA Dataset: Camera and automotive radar with range-angle-doppler annotations," arXiv, 2021. Accessed: Jan. 29, 2024. [Online]. Available: http://arxiv.org/abs/2005.01456

[29] D. Gusland, J. M. Christiansen, B. Torvik, F. Fioranelli, S. Z. Gurbuz, and M. Ritchie, "Open radar initiative: Large scale dataset for benchmarking of micro-Doppler recognition algorithms," in *2021 IEEE Radar Conference (RadarConf21),* Atlanta, GA, USA: IEEE, pp. 1–6, 2021.

[30] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *2014 IEEE Conference on Computer Vision and Pattern Recognition*, Columbus, OH, USA: IEEE, pp. 580–587, 2014.

[31] R. Girshick, "Fast R-CNN," in *2015 IEEE International Conference on Computer Vision (ICCV),* Santiago, Chile: IEEE, pp. 1440–1448, 2015.

[32] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 6, pp. 1137–1149, 2017.

[33] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C. Y. Fu et al., "SSD: Single shot multibox detector," in Computer Vision – ECCV 2016, B. Leibe, J. Matas, N. Sebe, and M. Welling, Eds., in Lecture Notes in Computer Science, Cham: Springer International Publishing, vol. 9905, pp. 21–37, 2016.

[34] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR),* Las Vegas, NV, USA: IEEE, 2016.

[35] N. Abdul Razak, N. A. A. Sabri, J. Johari, F. Ahmat Ruslan, M. Md. Kamal, and M. A. Abdul Aziz, "Investigation of object detection and identification at different lighting conditions for autonomous vehicle application," *International Journal of Automotive and Mechanical Engineering*, vol. 20, no. 3, pp. 10649–10658, 2023.

[36] W. Lv, Y. Zhao, S. Xu, J. Wei, G. Wang, Q. Dang et al., "DETRs Beat YOLOs on real-time object detection," *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16965-16974, 2024.

[37] X. Zhu, W. Su, L. Lu, B. Li, X. Wang, and J. Dai, "Deformable DETR: Deformable transformers for end-to-end object detection," arXiv, 2021. Accessed: Jan. 28, 2024. [Online]. Available: http://arxiv.org/abs/2010.04159

[38] F. Liu, H. Wei, W. Zhao, G. Li, J. Peng, and Z. Li, "WB-DETR: Transformer-based detector without backbone," in *2021 IEEE/CVF International Conference on Computer Vision (ICCV),* Montreal, QC, Canada: IEEE, pp. 2959–2967, 2021.

[39] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang et al., "Swin Transformer: Hierarchical vision transformer using shifted windows," in *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, Montreal, QC, Canada: IEEE, pp. 9992–10002, 2021.

[40] Z. Peng, W. Huang, S. Gu, L. Xie, Y. Wang, J. Jiao, et al., "Conformer: Local features coupling global representations for visual recognition," in *2021 IEEE/CVF International Conference on Computer Vision (ICCV),* Montreal, QC, Canada: IEEE, pp. 357–366, 2021.

[41] S. Mehta and M. Rastegari, "MobileViT: Light-weight, general-purpose, and mobile-friendly vision transformer." arXiv, Mar. 04, 2022. Accessed: Jan. 28, 2024. [Online]. Available: http://arxiv.org/abs/2110.02178

[42] Z. Chen, L. Xie, J. Niu, X. Liu, L. Wei, and Q. Tian, "Visformer: The vision-friendly transformer," in *2021 IEEE/CVF International Conference on Computer Vision (ICCV),* Montreal, QC, Canada: IEEE, pp. 569–578, 2021.

[43] R. Nabati and H. Qi, "CenterFusion: Center-based radar and camera fusion for 3D object detection," in *2021 IEEE Winter Conference on Applications of Computer Vision (WACV),* Waikoloa, HI, USA: IEEE, pp. 1526–1535, 2021.

[44] Y. Wang, Z. Jiang, X. Gao, J.-N. Hwang, G. Xing, and H. Liu, "RODNet: Radar object detection using cross-modal supervision," in *2021 IEEE Winter Conference on Applications of Computer Vision (WACV)*, Waikoloa, HI, USA: IEEE, pp. 504–513, 2021.

[45] Y. Kim and T. Moon, "Human detection and activity classification based on micro-doppler signatures using deep convolutional neural networks," *IEEE Geoscience and Remote Sensing Letters*, vol. 13, no. 1, pp. 8–12, 2016.

[46] S. Lee, Y. Yoon, J. Lee, and S. Kim, "Human‐vehicle classification using feature‐based SVM in 77‐GHz automotive FMCW radar," *IET Radar, Sonar & Navigation*, vol. 11, no. 10, pp. 1589‐1596, 2017.

[47] J. Redmon and A. Farhadi, "YOLOv3: An incremental improvement," arXiv, Apr. 08, 2018. Accessed: Jan. 28, 2024. [Online]. Available: http://arxiv.org/abs/1804.02767

[48] X. Dong, P. Wang, P. Zhang, and L. Liu, "Probabilistic oriented object detection in automotive radar," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW),* Seattle, WA, USA: IEEE, pp. 458–467, 2020.

[49] W. Ng, G. Wang, Siddhartha, Z. Lin, and B. J. Dutta, "Range-doppler detection in automotive radar with deep learning," in *2020 International Joint Conference on Neural Networks (IJCNN),* Glasgow, United Kingdom: IEEE, pp. 1–8, 2020.

[50] C. Decourt, R. VanRullen, D. Salle, and T. Oberlin, "DAROD: A deep automotive radar object detector on range-doppler maps," in *2022 IEEE Intelligent Vehicles Symposium (IV),* Aachen, Germany: IEEE, pp. 112–118, 2022.

[51] X. Gao, G. Xing, S. Roy, and H. Liu, "RAMP-CNN: A novel neural network for enhanced automotive radar object recognition," *IEEE Sensors Journal*, vol. 21, no. 4, pp. 5119–5132, 2021.

[52] J. Rebut, A. Ouaknine, W. Malik, and P. Perez, "Raw high-definition radar for multi-task learning," in *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, New Orleans, LA, USA: IEEE, pp. 17000–17009, 2022.

[53] A. Danzer, T. Griebel, M. Bach, and K. Dietmayer, "*2D Car Detection in Radar Data with PointNets," in 2019 IEEE Intelligent Transportation Systems Conference (ITSC)*, Auckland, New Zealand: IEEE, pp. 61–66, 2019.

[54] J. F. Tilly, S. Haag, O. Schumann, F. Weishaupt, B. Duraisamy, J. Dickmann et al., "Detection and tracking on automotive radar data with deep learning," in *2020 IEEE 23rd International Conference on Information Fusion (FUSION)*, Rustenburg, South Africa: IEEE, pp. 1–7, 2020.

[55] A. Dubey, A. Santra, J. Fuchs, M. Lübke, R. Weigel, and F. Lurz, "HARadNet: Anchor-free target detection for radar point clouds using hierarchical attention and multi-task learning," *Machine Learning with Applications*, vol. 8, p. 100275, 2022.

[56] R. Q. Charles, H. Su, M. Kaichun, and L. J. Guibas, "PointNet: Deep learning on point sets for 3D classification and segmentation," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Honolulu, HI: IEEE, pp. 77–85, 2017.

[57] C. R. Qi, L. Yi, H. Su, and L. J. Guibas, "PointNet++: Deep hierarchical feature learning on point sets in a metric space," Jun. 07, 2017, arXiv: arXiv:1706.02413.

[58] C. R. Qi, W. Liu, C. Wu, H. Su, and L. J. Guibas, "Frustum PointNets for 3D object detection from RGB-D data," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Salt Lake City, UT, USA: IEEE, pp. 918–927, 2018.

[59] M. Dreher, E. Ercelik, T. Banziger, and A. Knoll, "Radar-based 2D car detection using deep neural networks," in *2020 IEEE 23rd International Conference on Intelligent Transportation Systems (ITSC)*, Rhodes, Greece: IEEE, pp. 1–8, 2020.

[60] D. Köhler, M. Quach, M. Ulrich, F. Meinl, B. Bischoff, and H. Blume, "Improved multi-scale grid rendering of point clouds for radar object detection networks," in *2023 26th International Conference on Information Fusion (FUSION)*, Charleston, SC, USA: IEEE, pp. 1–8, 2023.

[61] J. Liu, Q. Zhao, W. Xiong, T. Huang, Q.-L. Han, and B. Zhu, "SMURF: Spatial multi-representation fusion for 3D object detection with 4D imaging radar," *IEEE Transactions on Intelligent Vehicles*, pp. 1–14, 2023.

[62] Y. Zhou and O. Tuzel, "VoxelNet: End-to-end learning for point cloud based 3D object detection," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Salt Lake City, UT, USA: IEEE, pp. 4490–4499, 2018.

[63] P. Svenningsson, F. Fioranelli, and A. Yarovoy, "Radar-PointGNN: Graph based object recognition for unstructured radar point-cloud data," in *2021 IEEE Radar Conference (RadarConf21)*, Atlanta, GA, USA: IEEE, pp. 1–6, 2021.

[64] S. Saponara and B. Neri, "Radar sensor signal acquisition and multidimensional FFT processing for surveillance applications in transport systems," *IEEE Transactions on Instrumentation and Measurement*, vol. 66, no. 4, pp. 604–615, 2017.

[65] 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*, Providence, RI: IEEE, pp. 3354–3361, 2012.

[66] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson et al., "The cityscapes dataset for semantic urban scene understanding," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, NV, USA: IEEE, pp. 3213–3223, 2016.

[67] X. Huang, X. Cheng, Q. Geng, B. Cao, D. Zhou, P. Wang et al., "The ApolloScape dataset for autonomous driving," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, Salt Lake City, UT: IEEE, pp. 1067–10676, 2018.

[68] P. Sun, H. Kretzschmar, X. Dotiwalla, A. Chouard, V. Patnaik, P. Tsui et al., "Scalability in perception for autonomous driving: Waymo open dataset," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Seattle, WA, USA: IEEE, pp. 2443–2451, 2020.

[69] Y. Wang, G. Wang, H.-M. Hsu, H. Liu, and J.-N. Hwang, "Rethinking of Radar's Role: A camera-radar dataset and systematic annotator via coordinate alignment," in *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, Nashville, TN, USA: IEEE, pp. 2809–2818, 2021.

[70] A. Ouaknine, A. Newson, J. Rebut, F. Tupin, and P. Perez, "CARRADA Dataset: Camera and automotive radar with range-angle- doppler annotations," in *2020 25th International Conference on Pattern Recognition (ICPR)*, Milan, Italy: IEEE, pp. 5068–5075, 2021.

[71] Palffy, E. Pool, S. Baratam, J. F. P. Kooij, and D. M. Gavrila, "Multi-class road user detection with 3+1D radar in the view-of-delft dataset," *IEEE Robotics and Automation Letters*, vol. 7, no. 2, pp. 4961–4968, 2022.

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

[73] T. Roddick, A. Kendall, and R. Cipolla, "Orthographic feature transform for monocular 3D object detection." arXiv, Nov. 20, 2018. Accessed: Feb. 05, 2024. [Online]. Available: http://arxiv.org/abs/1811.08188

[74] A. Simonelli, S. R. Bulo, L. Porzi, M. L. Antequera, and P. Kontschieder, "Disentangling monocular 3D object detection: From single to multi-class recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 3, pp. 1219–1231, 2022.

[75] X. Zhou, D. Wang, and P. Krähenbühl, "Objects as points." arXiv, Apr. 25, 2019. Accessed: Feb. 05, 2024. [Online]. Available: http://arxiv.org/abs/1904.07850

[76] R. Yadav, A. Vierling, and K. Berns, "Radar + RGB fusion for robust object detection in autonomous vehicle," in *2020 IEEE International Conference on Image Processing (ICIP)*, Abu Dhabi, United Arab Emirates: IEEE, pp. 1986–1990, 2020.

[77] G. Alessandretti, A. Broggi, and P. Cerri, "Vehicle and guard rail detection using radar and vision data fusion," *IEEE Transactions on Intelligent Transportation Systems*, vol. 8, no. 1, pp. 95–105, 2007.

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

[79] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” arXiv, Apr. 10, 2015. Accessed: May 28, 2023. [Online]. Available: http://arxiv.org/abs/1409.1556

[80] K. Bansal, K. Rungta, and D. Bharadia, “RadSegNet: A reliable approach to radar camera fusion,” arXiv, Aug. 07, 2022.

[81] Y. Long, D. Morris, X. Liu, M. Castro, P. Chakravarty, and P. Narayanan, “Radar-camera pixel depth association for depth completion,” in *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12502–12511, 2021.

[82] V. John and S. Mita, “RVNet: Deep sensor fusion of monocular camera and radar for image-based obstacle detection in challenging environments,” in *Image and Video Technology*, C. Lee, Z. Su, and A. Sugimoto, Eds., in Lecture Notes in Computer Science, Cham: Springer International Publishing, vol. 11854, pp. 351–364, 2019.

[83] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770-778, 2016.

[84] X. Guo, J. Du, J. Y. Gao, and W. Wang, “Pedestrian detection based on fusion of millimeter wave radar and vision,” *Proceedings of the 2018 International Conference on Artificial Intelligence and Pattern Recognition*, pp. 38-42, 2018.

[85] S. Chang, Y. Zhang, F. Zhang, X. Zhao, S. Huang, Z. Feng et al., “Spatial attention fusion for obstacle detection using MmWave radar and vision sensor,” *Sensors*, vol. 20, no. 4, p. 956, 2020.

[86] DU Xiaoyu, *Research on vehicle front target detection algorithm based on MMW radar and visual information fusion* [D]. Chongqing University of Science and Technology, 2021.

[87] S. Li, J. Yan, H. Chen, and K. Zheng, “Radar-camera fusion network for depth estimation in structured driving scenes,” *Sensors*, vol. 23, no. 17, p. 7560, 2023.

[88] T. Kurniawan, “ClusterFusion: Leveraging radar spatial features for radar-camera 3D object detection in autonomous vehicles,” *IEEE Access*, vol. 11, pp. 121511–121528, 2023.

[89] S. Pang, D. Morris, and H. Radha, “TransCAR: Transformer-based camera-and-radar fusion for 3D object detection,” arXiv, Apr. 30, 2023. Accessed: Feb. 12, 2024. [Online]. Available: http://arxiv.org/abs/2305.00397

[90] Kim, Y. Kim, and D. Kum, “Low-level sensor fusion network for 3D vehicle detection using radar range-azimuth heatmap and monocular image,” In *Proceedings of the Asian Conference on Computer Vision*, 2020.

[91] H. Jha, V. Lodhi, and D. Chakravarty, “Object detection and identification using vision and radar data fusion system for ground-based navigation,” in *2019 6th International Conference on Signal Processing and Integrated Networks (SPIN)*, pp. 590–593, 2019.

[92] X. Dong, B. Zhuang, Y. Mao, and L. Liu, “Radar camera fusion via representation learning in autonomous driving,” in *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 1672–1681, 2021.

[93] Philion, A. Kar, and S. Fidler, “Learning to evaluate perception models using planner-centric metrics,” arXiv, Apr. 18, 2020. Accessed: Feb. 12, 2024. [Online]. Available: http://arxiv.org/abs/2004.08745

[94] J. Kim, M. Seong, G. Bang, D. Kum, and J. W. Choi, “RCM-Fusion: Radar-camera multi-level fusion for 3D object detection.” arXiv, Feb. 05, 2024. Accessed: Feb. 12, 2024. [Online]. Available: http://arxiv.org/abs/2307.10249

[95] Lei, Z. Chen, S. Jia, and X. Zhang, “HVDetFusion: A simple and robust camera-radar fusion framework,” arXiv, Jul. 20, 2023. Accessed: Feb. 08, 2024. [Online]. Available: http://arxiv.org/abs/2307.11323

[96] Y. Kim, J. Shin, S. Kim, I.-J. Lee, J. W. Choi, and D. Kum, “CRN: Camera radar net for accurate, robust, efficient 3D perception,” in *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, Paris, France: IEEE, pp. 17569–17580, 2023.

[97] Y. Kim, S. Kim, J. W. Choi, and D. Kum, “CRAFT: Camera-radar 3D object detection with spatio-contextual fusion transformer,” *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 37, no. 1, pp. 1160–1168, 2023.

[98] K. Duan, S. Bai, L. Xie, H. Qi, Q. Huang, and Q. Tian, “CenterNet: Keypoint triplets for object detection,” in *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, Seoul, Korea (South): IEEE, pp. 6568–6577, 2019.

[99] T. Zhou, J. Chen, Y. Shi, K. Jiang, M. Yang, and D. Yang, “Bridging the View Disparity Between Radar and Camera Features for Multi-Modal Fusion 3D Object Detection,” IEEE Trans. Intell. Veh., vol. 8, no. 2, pp. 1523–1535, Feb. 2023.

[100] Cui, Y. Ma, J. Lu, and Z. Wang, “Radar enlighten the dark: Enhancing low-visibility perception for automated vehicles with camera-radar fusion.” arXiv, May 26, 2023. Accessed: Feb. 05, 2024. [Online]. Available: http://arxiv.org/abs/2305.17318

[101] A. Sengupta, A. Yoshizawa, and S. Cao, “Automatic radar-camera dataset generation for sensor-fusion applications,” *IEEE Robotics and Automation Letters*, vol. 7, no. 2, pp. 2875–2882, 2022.

[102] E. Haussmann, M. Fenzi, K. Chitta, J. Ivanecky, H. Xu, D. Roy et al., “Scalable active learning for object detection,” in *2020 IEEE Intelligent Vehicles Symposium (IV)*, Las Vegas, NV, USA: IEEE, pp. 1430–1435, 2020.

[103] Z. Zhu, K. Lin, A. K. Jain, and J. Zhou, “Transfer learning in deep reinforcement learning: A survey,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 11, pp. 13344–13362, 2023.

[104] Y. Zhao, Y. Yang, and S. Chen, “An active semi-supervised learning for object detection,” in *2023 International Conference on Culture-Oriented Science and Technology (CoST)*, Xi’an, China: IEEE, pp. 257–261, 2023.

[105] Meyer, Michael, and Georg Kuschk. “Automotive radar dataset for deep learning based 3d object detection,” In *2019 16th european radar conference (EuRAD)*, pp. 129-132, 2019.

[106] L. Zheng, Z. Ma, X. Zhu, B. Tan, S. Li, K. Long et al., “TJ4DRadSet: A 4D radar dataset for autonomous driving,” in *2022 IEEE 25th International Conference on Intelligent Transportation Systems (ITSC)*, Macau, China: IEEE, pp. 493–498, 2022.

[107] J. Deng, W. Dong, R. Socher, L.-J. Li, Kai Li, and Li Fei-Fei, “ImageNet: A large-scale hierarchical image database,” in *2009 IEEE Conference on Computer Vision and Pattern Recognition*, Miami, FL: IEEE, pp. 248–255, 2009.