MDPI

Review

Towards Deep Radar Perception for Autonomous Driving: Datasets, Methods, and Challenges

Yi Zhou [1,2,4,6], Lulu Liu [1,3,4,5], Haocheng Zhao[1,2,4,6], Miguel López-Benítez[6,7], Limin Yu[2], Yutao Yue [1,4,5,*]

1 Institute of Deep Perception Technology, JITRI, Wuxi 214000, China; yueyutao@idpt.org (Y.Y.)
2 Department of Electrical and Electronic Engineering, School of Advanced Technology, Xi'an Jiaotong-Liverpool University, Suzhou 215123, China; zhouyi1023@tju.edu.cn (Y.Z.); haocheng.zhao19@student.xjtlu.edu.cn (H.Z); limin.yu@xjtlu.edu.cn (L.Y.)
3 Department of Mathematical Sciences, School of Science, Xi'an Jiaotong-Liverpool University, Suzhou 215123, China; lulu.liu21@student.xjtlu.edu.cn (L.L.)
4 XJTLU-JITRI Academy of Industrial Technology, Xi'an Jiaotong-Liverpool University, Suzhou, 215123, China
5 Department of Mathematical Sciences, University of Liverpool, Liverpool L69 7ZX, U.K.
6 Department of Electrical Engineering and Electronics, University of Liverpool, Liverpool L69 3GJ, U.K.; mlopben@liverpool.ac.uk (M.LB.)
7 ARIES Research Centre, Antonio de Nebrija University, 28040 Madrid, Spain
* Correspondence: yueyutao@idpt.org (Y.Y.)

Abstract: With recent developments, the performance of automotive radar has improved significantly. The next generation of 4D radar can achieve imaging capability in the form of high-resolution point clouds. In this context, we believe that the era of deep learning for radar perception has arrived. However, studies on radar deep learning are spread across different tasks and a holistic overview is lacking. This review paper attempts to provide a big picture of the deep radar perception stack, including signal processing, datasets, labelling, data augmentation and downstream tasks such as depth and velocity estimation, object detection, and sensor fusion. For these tasks, we focus on explaining how the network structure is adapted to radar domain knowledge. In particular, we summarize three overlooked challenges in deep radar perception, including multi-path effects, uncertainty problems and adverse weather effects, and present some attempts to solve them. We also provide an website for browsing each reference and related codes at https://zhouyi1023.github.io/awesome-radar-perception.

Keywords: automotive radars; radar signal processing; object detection; multi-sensor fusion; deep learning; autonomous driving.

Citation: Zhou, Y.; Lu, L.; Zhao, H.; López-Benítez, M.; Yu, L.; Yue, Y. Towards Deep Radar Perception for Autonomous Driving: Datasets, Methods, and Challenges. *Sensors* **2022**, *1*, 0. https://doi.org/

Received:
Accepted:
Published:

1. Introduction

As autonomous driving technology progresses from the demonstration stage to the landing stage, it puts forward higher requirements for perception ability. Mainstream autonomous driving systems rely on fusion of cameras and LiDARs for perception. Although millimeter wave radar has been widely used in mass-produced cars for active safety functions such as Automatic Emergency Braking (AEB) and Forward Collision Warning (FCW), it is overlooked in autonomous driving. Recently, Tesla announced the removal of radar sensors from its semi-autonomous driving system Autopilot. In the CVPR 2021 workshop [1], Tesla's director of AI Andrej Karpathy explained their reason by presenting three typical scenarios for radar malfunctions, including lost tracking due to significant deceleration of the front vehicle, false slow down under bridges and missed detection of a stationary vehicle parked on the side of the main road. In the first case, radar's close field detection ability is related to side lobes. Conventional radars with a limited number of channels are not good at side lobe compression. The second case is caused by the fact that conventional radar cannot measure height information and therefore confuses the bridge overhead with static objects on the road. The reason for the third case is that conventional

radar has too low angular resolution to capture the shape of a static vehicle. All these challenges can be solved with next-generation high resolution radar.

As a ranging sensor, radar is usually compared to LiDAR. A typical 77 Ghz automotive radar has a wavelength of 3.9 mm, while automotive LiDARs have a much smaller wavelength of 905 nm or 1550 nm. For a small aperture radar, most of the reflected signal is not received by the radar sensor because of the specular reflection. Another problem with small aperture is the low angular resolution, so that two close objects cannot be separated effectively. These two features make the radar point cloud much sparser than the LiDAR point cloud. Conventional automotive radars have a low resolution in elevation and therefore return a two-dimensional point cloud. The next generation of high resolution radar achieves higher angular resolution in both azimuth and elevation. Because it can measure 3D position and Doppler velocity, it is always referred to as a 4D radar in the marketplace. In table 1, typical types of radars and LiDARs are compared. We can find that conventional long-range radars have a low angular resolution in horizontal view and no resolution in vertical view. In contrast, 4D radar can achieve an angular resolution of about 1°in both horizontal and vertical views. Therefore, the classification of static objects is no longer a limitation for 4D radar. Although 4D radar has much higher angular resolution, as shown in fig. 1, the radar point cloud is still much sparser than the 16-beam LiDAR point cloud. However, radar can measure Doppler velocity and radar cross section (RCS), which is expected to better help classify road users. In addition, 4D radar has the advantages of long detection range (up to 300 meters), all-weather operation, low power consumption and low cost. Therefore, we believe that radar is a good complement to LiDAR and vision. The fusion of these sensors enables all-weather, long-range environment perception.

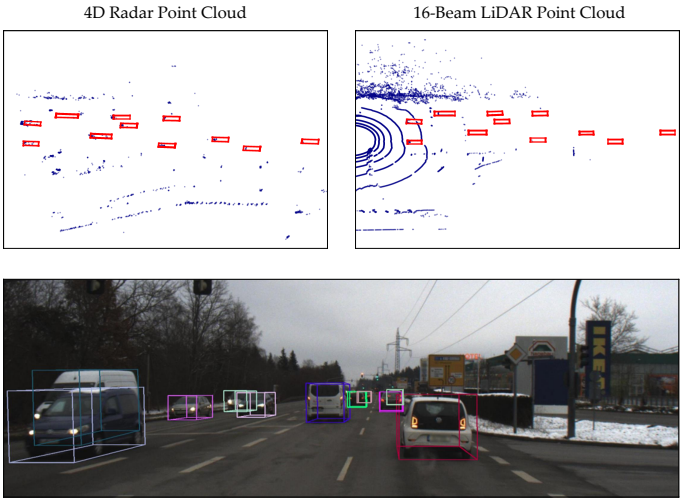


Figure 1. Point clouds of a 4D radar and a 16-beam LiDAR from the Astyx dataset [2]

Table 1. Characteristics of typical radars and LiDARs

	Conventional Radar[1] (Multi-Mode)	**4D Radar**	**16-Beam LiDAR**	**32-Beam LiDAR**	**Solid State LiDAR**
Max Range	f: 250m, n:70m	300m	100m	200m	200m
FoV (H/V)	f: 20°, n: 120°/ ✗	120°/ 30°	360°/ 30°	360°/ 40°	120°25°
Ang Res (H/V)	f: 1.5°, n: 4°/ ✗	1°/ 1°	0.1°/ 2°	0.1°/ 0.3°	0.2°/ 0.2°
Doppler Res	0.1m/s	0.1m/s	✗	✗	✗
Point Density	Low	Medium	High	High	High
All Weather	✓	✓	✗	✗	✗
Power	5W	5W	8W	10W	15W
Expected Cost	Low	Low	Medium	High	Medium

[1] A typical 77GHz 4Tx-6Rx automotive radar, 2Tx-6Rx for far range and 2Tx-6Rx for near range.

In recent years, with the trend of open source, more and more datasets, models and toolboxes have been released. According to our statistics, 10 radar datasets were released in 2021 and 2022. Along with these datasets, some seminal papers are proposed to leverage deep learning in radar perception. However, due to the limited sensing capability of conventional radar, the performance of these methods is far from good enough. Since the introduction of 4D radar, we believe that the era of deep radar perception has arrived. With the power of deep learning, we can design a highly reliable perception system based on the fusion of radar and other modalities.

The application of deep learning in radar perception has drawn extensive attention from autonomous driving researchers. In the past two years, a number of review papers [3–8] have been published in this field. Zhou *et al*. [3] categorize radar perception tasks into dynamic target detection and static environment modelling. They also provide brief introductions to radar-based detection, tracking and localization. Abdu *et al*. [4] summarize the deep learning models for radar perception based on different radar representations, including occupancy grid maps, range-Doppler-azimuth maps, micro-Doppler spectrograms and point clouds. They also introduces approaches for radar and camera fusion based on the classical taxonomy of data-level, feature-level and decision-level. Scheiner *et al*. [5] discuss the information sparsity problem and labelling challenge in learning-based radar perception. Three strategies are recommended to increase radar data density, including the use of pre-CFAR data, the use of high-resolution radar sensors, and the use of polarization information. In this paper, we differ from other review papers in three aspects: Firstly, we provide a detailed summary and description of the public available radar datasets, which is very useful for developing learning-based methods. Secondly, this review does not focus on specific tasks, but aims to present a big picture of the radar perception framework, as illustrated in fig. 2. Thirdly, rather than simply presenting the network structure, we focus on explaining why these modules work from the perspective of radar domain knowledge.

In this article, we systematically review the recent advancements in deep radar perception. In section 2, we introduce radar signal processing pipeline and different radar data representations. In section 3, we summarize the public-available radar datasets (section 3.1) for autonomous driving, as well as the calibration (section 3.2), labelling (section 3.3) and data augmentation techniques (section 3.4). In the following sections, we introduce different radar perception tasks, including radar depth completion (section 4.1), radar full velocity estimation (section 4.2), radar object detection (point-cloud-based in section 5.2 and pre-CFAR-based in section 5.3). In section 6, we classify sensor fusion frameworks into four categories: input fusion (section 6.1), ROI fusion (section 6.2), feature map fusion (section 6.3) and decision fusion (section 6.4). Next, we discuss three challenges toward reliable radar detection, including ghost objects (section 7.1), uncertainty problems (section 7.2), and adverse weather effects (section 7.3). Finally, we propose several interesting research directions in section 8.

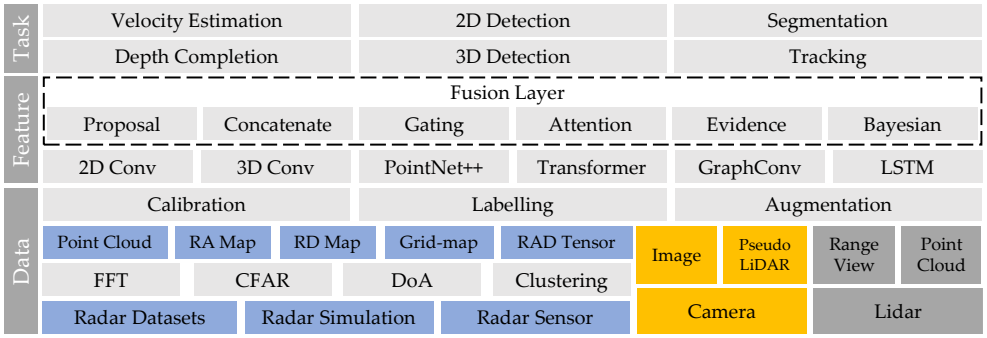


Figure 2. Overview of deep radar perception framework.

2. Radar Signal Processing Fundamentals

Knowledge of radar signal processing is essential for the development of a deep radar perception system. Different radar devices vary in their sensing capabilities. It is important to leverage radar domain knowledge to understand the performance boundary, find key scenarios and solve critical problems. This section outlines the classical signal processing pipeline for automotive radar applications.

2.1. FMCW Radar Signal Processing

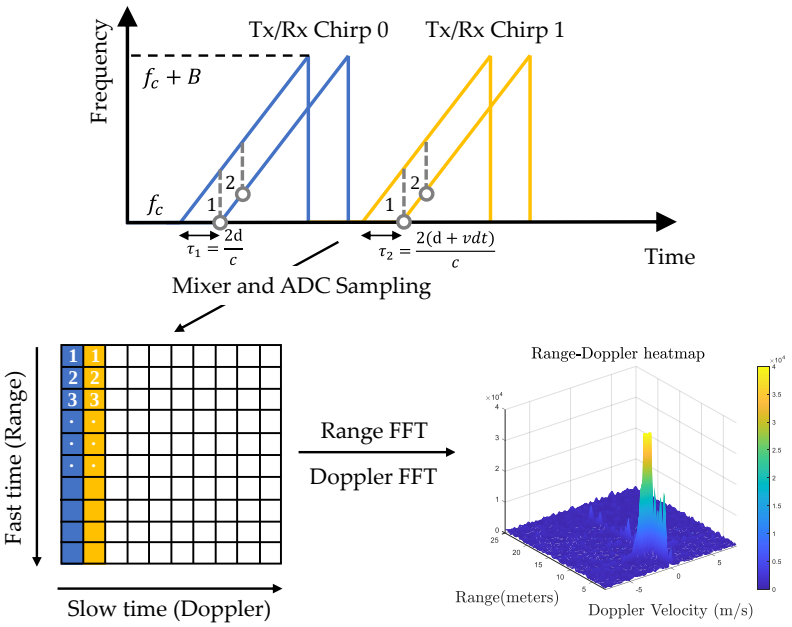


Figure 3. Radar Tx/Rx signals and the resulted range-Doppler map.

Off-the-shelf automotive radars operate with a sequence of linear frequency-modulated continuous-wave (FMCW) signals to simultaneously measure range, angle and velocity. According to regulations, automotive radar are allowed to use two frequency bands in millimeter waves: 24 GHz (24-24.25 GHz) and 77 GHz (77-79 GHz). There is a trend towards 77 GHz due to its larger bandwidth (76-77 GHz for long-range and 77-81 GHz for short-range), higher Doppler resolution and smaller antennas [9]. As shown in fig. 3, the FMCW signal is characterized by the start frequency (also known as the carrier frequency) f_c, the sweep bandwidth B, the chirp duration T_c and the slope $S = B/T_c$. During one chirp duration, the frequency increases linearly from f_c to $f_c + B$ with a slope of S. One FMCW waveform is referred to as a chirp, and radar transmits a frame of N_c chirps equally spaced by chirp cycle time T_c. The total time $T_f = N_c T_c$ is called the frame time, also known as the time on target (TOT). In order to avoid the need of high-speed sampling, a frequency mixer combines the received signal with the transmitted signal to produce two signals with sum frequency $f_T(t) + f_R(t)$ and difference frequency $f_T(t) - f_R(t)$. Then a low-pass filter is used to filter out the sum frequency component and obtain the intermediate frequency (IF) signal. In this way, FMCW radar can achieve GHz performance with only MHz sampling. In practice, a quadrature mixer is used to improve the noise figure [10], resulting in a complex exponential IF signal as

$$x_{\text{IF}}(t) = Ae^{j(2\pi f_{\text{IF}} t + \phi_{\text{IF}})} \tag{1}$$

where A is the amplitude, $f_{\text{IF}} = f_T(t) - f_R(t)$ is referred to as the beat frequency and ϕ_{IF} is the phase of the IF signal. Next, the IF signal is sampled N_s times by an ADC converter, resulting in a discrete-time complex signal. Multiple frames of chirp signals are

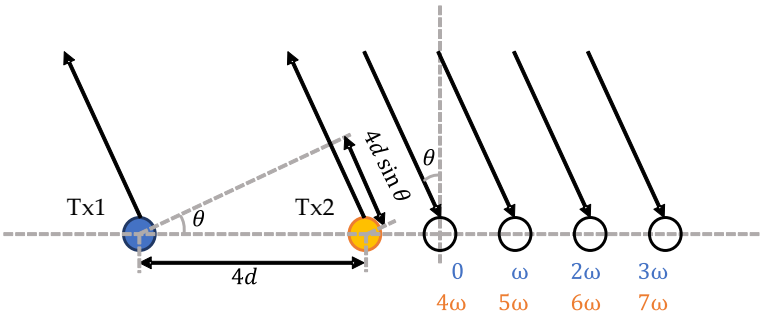
assembled into a two-dimensional matrix. As shown in fig. 3, the dimension of sampling points within a chirp is referred to as fast time, and the dimension of chirp index within one frame is referred to as slow time. Assume one object moving with speed v at distance r, the frequency and phase of the IF signal are given by

$$f_{\mathrm{IF}} = \frac{2S(r + vT_c)}{c}, \phi_{\mathrm{IF}} = \frac{4\pi(r + vT_c)}{\lambda} \tag{2}$$

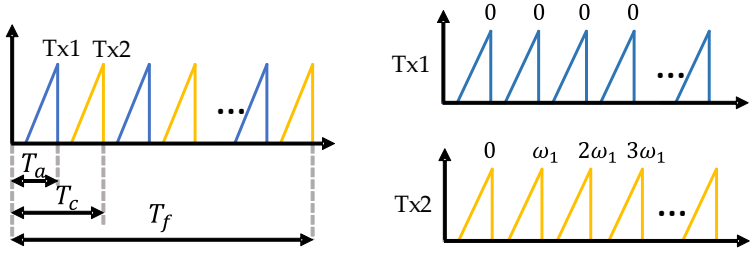
where $\lambda = c/f_c$ is the wavelength of the chirp signal. From eq. (2), we can find that range and Doppler velocity are coupled. Under the following assumptions: 1. the range variations in slow time caused by target motion can be neglected due to the short frame time. 2. the Doppler frequency in fast time can be neglected compared to the beat frequency by utilizing a wideband waveform. Then, range and Doppler can be decoupled. Range can be estimated from the beat frequency as $r = cf_{\mathrm{IF}}/2S$ and Doppler velocity can be estimated from the phase shift between two chirps as $v = \Delta\phi\lambda/4\pi T_c$. Next, a range DFT is applied in the fast-time dimension to resolve the frequency change, followed by a Doppler DFT in the slow-time dimension to resolve the phase change. As a result, we obtain a 2D complex-valued data matrix called Range-Doppler (RD) map. In practice, a window function is applied before DFT to reduce sidelobes. The range and the Doppler velocity of a cell in RD map are given by

$$r_k = k\frac{c}{2B_{\mathrm{IF}}}, v_l = l\frac{\lambda}{2T_f} \tag{3}$$

where k and l denote the indexes of DFT, B_{IF} is the IF bandwidth, and T_f is the frame time. In practice, FFT is applied due to its computational efficiency. Accordingly, the sequence will be zero-padded to the nearest power of 2 if necessary.



(a) MIMO Radar Antenna



(b) TDM (c) DDM

Figure 4. MIMO radar principles.(a) Virtual array configuration of a 2Tx4Rx MIMO radar (b) In TDM mode, Tx1 and Tx2 transmit signals by turns. (c) In DDM mode, a Doppler shift is added to Tx2.

Angle information can be obtained using more than one receive or transmit channel. Single-input multiple-output (SIMO) radars utilize a single transmit (Tx) and multiple receive (Rx) antennas for angle estimation. Suppose one object is located at direction θ. Similar to Doppler processing, the induced frequency change between two adjacent receiver antennas can be neglected, while the induced phase change can be used for calculating

the direction of angle. This phase change is given by $\Delta\phi = 2\pi d \sin\theta/\lambda$, where d is the inter-antenna spacing. To achieve maximum unambiguous angle, the spacing can be set to $\lambda/2$. Then, a third FFT can be applied to the received antenna dimension. For conventional radar with a small number of Rx antennas, the sequence is often padded with $N_{\text{FFT}} - N_{Rx}$ zeros to achieve a smooth display of the spectrum. The angle at index η is given by

$$\theta_\eta = \arcsin \frac{\eta\lambda}{N_{\text{FFT}}} \tag{4}$$

The angular resolution of a SIMO radar depends on the number of Rx antennas. The maximum number of Rx antennas is limited by the additional cost of signal processing chains on the device [11]. Multiple-input multiple-output (MIMO) radar operates with multiple channels in both Tx and Rx. As illustrated in fig. 4 (a), a MIMO radar with N_{Tx} Tx antennas and N_{Rx} Rx antennas can synthesize a virtual array with $N_{\text{Tx}}N_{\text{Rx}}$ channels. In order to separate the transmit signals at the receiver side, the signals from different Tx antennas should be orthogonal. There are multiple ways to realize waveform orthogonal, such as time-division multiplexing (TDM), frequency-division multiplexing (FDM), and Doppler-division multiplexing (DDM) [12,13]. TDM is widely used for its simplicity. In this mode, different Tx antennas transmit chirp signals by turns, as shown in fig. 4 (b). Therefore, at the receiver side, different Tx waveforms can be easily separated in the time domain. An additional phase shift compensation [14] is required to compensate for the motion of detections during the Tx switching time. Another shortcoming of TDM is the reduced detection range due to loss of transmitting power. DDM is also supported by many radar devices. As shown in fig. 4 (c), DDM transmits all Tx waveform simultaneously and separate them in Doppler domain. In order to realize waveform orthogonality, for the the k-th transmitter, a Doppler shift is added to adjacent chirps as

$$\omega_k = \frac{2\pi(k-1)}{N} \tag{5}$$

where N is usually selected as the number of Tx antennas N_{Tx}. One drawback of DDM is its unambiguous Doppler velocity is reduced to $\frac{1}{N}$ of the original one. Empty-band DDM [15] can achieve more robust velocity disambiguation by introducing several empty Doppler subbands. Some example codes are provided in RADIal dataset [16]. After decoupling the received signals, we can obtain a 3D tensor by stacking RD maps with respect to Tx-Rx pairs. Then, DOA can be estimated through angle FFT along the virtual receiver dimension. Some super-resolution methods, such as MUSIC [17], can be applied to improve angular resolution. The resulting 3D tensor is referred to as Range-Azimuth-Doppler (RAD) tensor or radar tensor.

In radar detection pipeline, RD maps are integrated coherently along the virtual receiver dimension to increase SNR. Then, a Constant False Alarm Rate (CFAR) detector [17] is applied to detect peaks in the RD map. Finally, the DOA estimation method is applied for angle estimation. The output is a point cloud with measurements of range, Doppler and angle. For conventional radars, only azimuth angle is resolved, while 4D radars output both azimuth and elevation angles. Since radar is usually used in safety-critical applications, a lower CFAR threshold ($\leq$ 10 dB) is set to achieve high recall. The accuracy of detection is affected by road clutter, interferences and multi-path effects in complex environments. Therefore, additional spatial-temporal filtering is required to improve accuracy. DBSCAN [18] is used to cluster radar detections into object-level targets. Clusters with few detections are considered as outliers and thus be removed. Further, temporal filtering, such as Kalman filtering, is used to filter out outliers and interpolate missed detections.

2.2. Radar Performances

Performance of automotive radar can be evaluated in terms of maximum range, maximum Doppler velocity and field of view (FoV). Equations for these attributes are

summarized in table 2. According to radar equation, the theoretical maximum detection range is given by

$$R_{\max} = \sqrt[4]{\frac{P_t G^2 \lambda^2 \sigma}{(4\pi)^3 P_{\min}}} \tag{6}$$

where P_t is the transmit power, P_{min} is the minimum detectable signal or receiver sensitivity, λ is the transmit wavelength, σ is the target RCS and G is the antenna gain. The wavelength is 3.9 mm for automotive 77Ghz radar. Target RCS is a measure of ability to reflect radar signals back to the radar receiver. It is a statistical quantity that varies with the viewing angle and the target material. According to the test results [19], smaller objects, such as pedestrian and bike, have a average RCS value of around 2-3 dBsm, whereas normal vehicles have an average RCS value of around 10 dBsm and large vehicles of around 20 dBsm. The other parameters, such as transmit power, minimum detectable signal and antenna gain are design parameters aimed at meeting product requirements as well as regulations. Some typical values for these parameters are summarized in table 3. In practice, the maximum range is limited by the supported IF bandwidth B_{IF} and ADC sampling frequency. The maximum unambiguous velocity is inversely proportional to the chirp duration T_c. For MIMO radar, the maximum unambiguous angle is dependent on the spacing of antennas d. The theoretical maximum FoV is 180° if $d = \lambda/2$. In practice, FoV is determined by the antenna gain pattern. Another important characteristic is resolution,*i.e.*, the ability to separate two close targets with respect to range, velocity and angle. As shown in table 2, high range resolution requires a large sweep bandwidth B. High Doppler resolution requires a long integration time, *i.e.*, the frame time $N_c T_c$. The angular resolution depends on the number of virtual receivers N_R, the object angle θ and the inter-antenna spacing d. For the case of $d = \lambda/2$ and $\theta = 0°$, angular resolution is in a simple form of $2/N_R$. From the perspective of antenna theory, angular resolution can also be featured by the half-power beamwidth, *i.e.*, the 3-dB beamwidth [13], which is a function of the array aperture D.

Table 2. Equations for radar performance

Definition	Equation
Max Unambiguous Range	$R_m = \frac{cB_{\text{IF}}}{2S}$
Max Unambiguous Velocity	$v_m = \frac{\lambda}{4T_c}$
Max Unambiguous Angle	$\theta_{\text{FoV}} = \pm \arcsin(\frac{\lambda}{2d})$
Range Resolution	$\Delta R = \frac{c}{2B}$
Velocity Resolution	$\Delta v = \frac{\lambda}{2N_c T_c}$
Angular Resolution	$\Delta\theta_{\text{res}} = \frac{\lambda}{N_R d \cos(\theta)}$
3-dB beamwidth	$\Delta\theta_{3\text{dB}} = 2\arcsin\frac{1.4\lambda}{\pi D}$

The meaning of parameters are consistent in this section. Refer to appendix A for a quick check the meaning.

Table 3. Typical automotive radar parameters[20]

Parameter	Range
Transit power (dBm)	10–13
TX/RX antenna gain (dBi)	10–25
Receiver noise figure (dB)	10–20
Target RCS (dBsm)	(-10)-20
Receiver sensitivity (dBm)	(-120) - (-115)
Minimum SNR (dB)	10-20

In practice, different types of automotive radar are designed for different scenarios. Long Range Radar (LRR) achieves long detection range and high angular resolution at the

cost of a smaller FoV. Short Range Radar (SRR) uses MIMO techniques to achieve high angular resolution and large FoV. In addition, different chirp configurations [21] are used for different applications. For example, Long Range Radar needs to detect fast-moving vehicles at distances, and therefore utilizes small ramp slope for long distance detection, long chirp integration time to increase SNR, small chirp duration to increase maximum velocity and short chirp duration for high velocity resolution [22]. Short Range Radar needs to detect vulnerable road users (VRUs) close to the vehicle, and therefore utilizes higher sweep bandwidth for high range resolution at the cost of short range. Multi-mode radar [21] can work in different modes simultaneously by sending chirps that are switched sequentially with different configurations.

2.3. Open-Source Radar Toolbox

Commercial off-the-shelf radar products can only output point clouds. It can be configured to output either raw point clouds, sometimes referred to as radar detections, or clustered objects wit tracked ids. The signal processing algorithm inside it is a black box and cannot be modified. Alternatively, TI mm-wave radars have been widely used in academic research because of their public nature and flexibility. They support configurable chirps [21] and different MIMO modes [11] to adapt to different tasks. TI also provides a mmWave studio which provides GUIs for radar setup, data capturing, signal processing and visualization. In addition, there are some open-source radar signal processing toolboxes for TI devices, for example, RaDICaL SDK [23,24], PyRapid [25], OpenRadar [26] and Pymmw [27]. These toolboxes enable researchers to build their own datasets using TI devices. While there are a growing number of public radar datasets, most of them provide limited information about the radar configurations they use. This makes it difficult to make a fair comparison between algorithms trained on different datasets. Open radar initiative [28] provides a guideline for radar configuration and encourages researchers to expand this dataset by using the radar device with the same configuration.

3. Datasets, Labelling and Augmentation

Data plays a key role in the learning-based approaches. In the past, radar algorithms were always evaluated on private datasets. Recently, with the trend towards open source, many radar datasets have become publicly available. In this section, we summarize these radar datasets with respect to their data representations, tasks, scenarios, and annotation types, as shown in table 4. To motivate readers to build their own datasets, we also introduce extrinsic calibration and cross-modality labelling techniques. We further investigate data augmentation methods and the potential use of synthetic radar data to improve data diversity.

3.1. Radar Datasets

Different radar datasets use different types of radar. We can classify radar sensors into Low Resolution (LR) and High Resolution (HR). There are different technical routes to achieve high resolution, such as polarimetric radar [29], cooperative radars [30], multi-chip cascaded MIMO radar [13], synthetic aperture radar (SAR) [31] and spinning radar [32]. Most off-the-shelf radars can output a point cloud with range, azimuth angle, Doppler velocity and RCS. Next-generation 4D radar can also measure elevation angle. Some radar prototypes can be configured to output radar raw data, including ADC data, RA/RD maps and RAD tensors.

The role of radar in autonomous driving can be divided into localization and detection. Although this paper focus on radar detection, we also introduce the localization datasets in this section. Since these datasets usually provide synchronized LiDAR and image along with radar data, it is possible to annotate them for detection purpose, as done in [33]. There are various levels of label granularity for radar data. For radar point cloud, it is possible to provide 2D bounding boxes, 3D bounding boxes or point-wise annotations. 2D bounding boxes are labelled in bird's-eye view (BEV) and with orientation information, hence they

are sometimes referred to as pseudo-3D boxes. 3D bounding boxes further capture height information and pitch angle. If properly annotated, point-wise annotation can provide semantic information at a finer granularity than bounding boxes. In fact, radar detections within the bounding box could also be ghost detection or clutter. Therefore, point-wise annotation is a better way to capture the noisy nature of radar point cloud. Similarly, radar pre-CFAR data, including RA/RD maps and RAD tensors, are also annotated point-wise. Some works dilate the annotated points into a dense mask or a bounding box. However, these dilated patches do not necessarily reflect the shape information. Some techniques for precise dilation will be introduced later in section 4.

There are some large scale datasets for autonomous driving which include off-the-shelf 2D radars in their sensor suites. NuScenes [34] is the most popular dataset for its large-scale and diverse scenarios. The capturing vehicle is equipped with a 32 beam LiDAR, 6 cameras, 5 long-range multi-mode radars and a GPS/IMU system. It provides 3D annotations of 23 classes of road users in 1,000 scenes, with a total of 1.3 million frames. However, this dataset is not a good choice for studying the role of radar in perception, because its radar point clouds are too sparse. PixSet dataset [35] also aims at 3D object detection. The vehicle is equipped with a colocated sensor platform consisting of a solid-state LiDAR, a 64-beam LiDAR, a TI AWR1843 radar and a GPS/IMU system. The FoVs of different modalities are largely overlapped and hence are well suited for evaluating sensor fusion algorithms. RadarScenes dataset [36] is a diverse large-scale dataset for instance segmentation of radar point clouds. It uses four 77GHz radars with overlapping FoV in the front of the vehicle. Each radar is in middle-range mode with maximum range of 100m and 60°FoV. Compared to NuScenes dataset, its radar point clouds are much denser. The datasets contains 100km of driving in 158 different scenarios. It provides both point-wise annotations and track IDs for 11 classes of moving road users. All points with zero velocity are labelled as static. Pointillism [37] leverages a multi-radar setup to improve the resolution. Two TI IWR1443 radars were placed at the front of the car, facing forward, at a distance of 1.5 meters. The aim is to study the effect of coherently integrate point clouds from two radar sensors. In order provide ground truth of radar point clouds, the sensor suite also include a 16-beam LiDAR and a camera with overlapping FoV. The dataset contains 54K synchronized frames for five typical driving scenarios under different weather conditions. It also provides 3D box annotations of vehicles. Zendar dataset [38] is a high-resolution radar dataset that uses SAR for moving vehicle detection. It provides time-synchronized image, radar ADC data, 2D SAR point cloud and projected LiDAR point cloud in BEV. Point-wise annotations of moving vehicles are applied to the SAR point cloud. It also provides an SDK for converting raw ADC data to RD maps and visualization.

Several datasets utilize radar sensors in short-range (SR) or ultra-short-range (USR) mode for high resolution close-field imaging. In this mode, close objects will occupy several cells in both range and Doppler dimension (because of the micro-Doppler motion). To fully utilize these spatially spread range and Doppler signatures, annotations are made directly on RA maps or RAD tensors. CARRADA dataset [39] uses TI AWR1843BOOST radar in short-range mode, with a max distance of 50 meters. It provides real-valued RA maps, RD maps and unannotated RAD tensors, as well as synchronized images, for training neural networks. In both RA and RD maps, objects are annotated in point level with a category from pedestrian, car, or cyclist. In addition, the dilated segmentation mask and the bounding box around the cluster are also provided. The data are collected on an empty test track with at most two moving objects in the FoV. RADDet dataset [40] also uses TI AWR1843BOOST radar with a max distance of 50 meters, as well as a stereo camera. It provides 3D bounding boxes for complex-valued RAD tensor and 2D bounding boxes for RA map projected in Cartesian view. The data are captured using a tripod located on the sidewalks and facing the main roads. Therefore, its scenario is much more complex than the CARRADA dataset. CRUW dataset [41] uses a TI AWR1843 radar and a stereo camera for object detection. It adopts a different signal processing pipeline, which directly outputs RA map using range FFT and angle FFT. Then, the object-level point-wise annotation

is applied to complex-valued RA maps. A probabilistic camera-radar fusion approach is used to improve annotation quality. The dataset contains 3.5 hours, 400K frames of camera-radar data in different driving scenarios, including parking lot, campus road, city street, and highway. RaDICaL dataset [42] uses TI IWR1443BOOST radar in multiple configurations for different scenarios, including indoor, parking lot, highway and single human walking. It records radar ADC data together with the RGB-D images and IMU data using ROS. It also provides a signal processing SDK [23] to process and annotate radar data. Ghent VRU dataset [43] collects radar data specifically for VRU detection. The sensor suite includes a TI AWR1243 radar, a camera and a 16-beam LiDAR. The data are recorded by a vehicle driving on public roads in a crowded European city center. It provides radar RAD tensors with segmentation mask annotations for VRUs. To compensate for range-dependent power, many datasets apply logarithmic scaling or normalization to the pre-CFAR data as default. CARRADA dataset [39] and RADDet dataset [40] apply logarithmic scaling to their radar data. The normalization can be applied in different ways, including local power normalizing in Ghent VRU dataset [43], min-max scaling in CRUW dataset [41], and Z-score standardization in RADDet dataset [40]. Here we only summarize some operations that are explicitly mentioned. Further checks are needed when benchmarking the algorithm using different datasets.

As 4D radar is just entering the market, only a few public datasets are available. Astyx dataset [2] is the first public-available 4D radar dataset. The sensors include a 16-beam LiDAR, a camera and a Astyx 6455 HiRes 4D radar. It provides 3D bounding box annotations of seven classes of road users. Each object is also featured with four levels of occlusion and three levels of uncertainty. The dataset is very small, with only 500 annotated frames of short clips, each clip containing less than 10 frames. The class distribution is very imbalanced, with over 90% annotated objects are car. View-of-Delft (VoD) dataset [44] is a recently published 4D radar dataset especially focus on detection of VRUs. The sensor suite includes a ZF FRGen 21 4D radar, a 64-beam LiDAR and a stereo camera. It provides 8693 annotated frames with 3D bounding boxes and tracking ids. Each object is also annotated with two levels of occlusion and four types of activity attribute (stopped, moving, parked, pushed, sitting). The data is collected in campus, suburb and old-town locations, with a preference on scenarios containing VRUs. It provides fine-grained annotations of vehicle, trucks and 10 classes of VRUs. Different classes are equally distributed (21.6% pedestrians, 8.8% cyclists and 21.9% cars). RADIal dataset [45] is a 4D radar dataset for vehicle detection and open space segmentation. The sensors include camera, LiDAR, 4D radar, GPS and vehicle's CAN traces. The 4D radar is a 12Tx 16Rx cascaded radar. A key feature is that they also provide radar raw ADC data, which makes it possible to explore the potential of neural networks in the signal processing stage. This dataset is comparable in size to the VoD dataset, with annotated 8252 frames captured in city street, highway and countryside road. Two kinds of annotation are provided, including vehicle annotations and open space segmentation mask in BEV. The vehicle annotations are in the format of 2D bounding boxes for image and object-level points for LiDAR and radar. Although they do not provide bounding box annotations for the radar point clouds, it is possible for the researchers to annotate them on their own, given the LiDAR point clouds and images. TJ4DRadSet [46] is 4D dataset for 3D detection and tracking. The sensor suite includes a 32 beam LiDAR, a camera, a high performance 4D radar (Oculii Eagle) and GNSS. By utilizing Oculii's virtual aperture imaging technique, this 4D radar can output a much denser point cloud than others. It has a maximum detection range of 400 m and an angular resolution of less than 1°in both azimuth and elevation. The data are captured in a wide range of road conditions in urban driving. The dataset contains a total of 40K frames of synchronized data, where 7757 frames of them are annotated with 3D bounding boxes and track id.

Radar can also be utilized for localization. Compared to camera and LiDAR, radar has advantages of long detection range and robustness to occlusions. Millimeter wave can penetrate certain non-metallic objects, such as glass, Polywood and clay bricks [32], and is less affected by dust, smoke, fog, rain, snow, and ambient lighting conditions [32].

Therefore, radar has great potential for mapping and localization in adverse weathers. The Oxford radar robotcar dataset [47] is the most popular dataset for radar SLAM. The test car is equipped with a rich set of sensors, including an FMCW spinning radar, two 32 beam LiDARs, a stereo camera, three monocular cameras, two 2D LiDARs and a GPS/IMU system. The spinning radar can provide a 360° high-resolution intensity map of surrounding environments. However, it has no Doppler information and is rarely used in production cars due to the high price. Mulran dataset [48] focuses on range-sensor-based place recognition. It uses a spinning radar and a 64 beam LiDAR to capture the surrounding environment. The recorded data are temporally (monthly revisits) and structurally (multi-city) diverse. Borea dataset [49] aims at studying the effect of seasonal variation on long-term localization. The sensor suite includes a spinning radar, a camera, a GPS/IMU system and a 128-beam LiDAR. The data was collected by driving a repeated route over one year, thus capturing seasonal variations and adverse weather conditions. It provides pose ground truth for localization task, as well as 3D bounding box annotations for object detection in sunny weather. Similar to Borea dataset, EU long-term dataset [50] aims at localization in highly dynamic environments and long-term autonomy. Its sensor suite includes two stereo cameras, two 32 beam LiDARs, two fisheye cameras, a four beam LiDAR and a 77Ghz long-range radar and a 2D LiDAR facing the road. Endeavour dataset [51] adopts 5 multi-mode radar to cover the 360°surrounding environment. It is also equipped with LiDARs and RTK-GPS to provide ground truth for radar odometry. ColoRadar dataset [52] utilizes a compact moving sensor rig, which consists of a 64 beam LiDAR, a TI AWR2243 cascaded 4D radar, a TI AWR1843 radar and an IMU. Three levels of radar data representation are provided, including raw ADC samples, Range-Azimuth-Elevation-Doppler (RAED) tensors from the 4D radar and point clouds from the single-chip radar. The data are gathered in a variety of scenarios, including highly diverse indoor environments, outdoor environments and an underground mine.

There are also some radar datasets designed for specific tasks. PREVENTION [53] focuses on predicting inter-vehicle interactions. The data collection car is equipped with one frontal long-range radar, two corner radars, one 32 beam LiDAR and two cameras. It provides annotations of 2D bounding boxes, lane change behaviours, and trajectories. SCORP [54] is a radar dataset for open space segmentation in parking scenarios. It provides three kinds of radar data, including RAD tensor, RA map and BEV map. Radar Ghost dataset [55] aims at studying the effect of multi-path propagation in autonomous driving. It provides point-wise annotations of real targets and four types of ghost targets.

Robust perception under adverse weather is a popular research topic for safe autonomous driving. Although there are some recently published datasets for adverse weather [56–59], only a few include radar in their sensor suite. Dense dataset [60] focus on evaluating multi-modal fusion algorithms under adverse weather. In addition to LiDAR and stereo camera, it is also equipped with several all-weather sensors, including one frontal long-range radar, one gated camera working on NIR band, one FIR camera and one weather station sensor. The data are captured in various natural weather conditions, including rain, snow, light fog and heavy fog, as well as in a controlled lab environment in a fog chamber. However, the dataset only provides sparse radar targets with limited FoV and poor resolution. RADIATE dataset [61] particularly focuses on leveraging radar in adverse weather. The data collection car is equipped with a camera, a LiDAR and a spinning radar. The datasets are captured under different weathers, such as sun, night, rain, fog and snow. It provides annotations for 2D object detection, object tracking and SLAM.

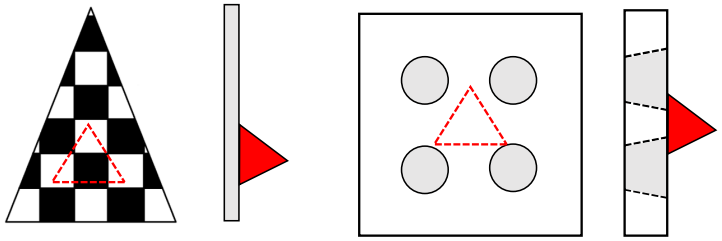
Table 4. Radar Datasets

Name	Year	Task	Radar Type	Data	Doppler	Range	Modalities	Scenarios	Weather	Annotations	Size
Off-the-shelf Radar Datasets for Detection											
nuScenes [34]	2020	DT	LR	PC	✓	SV	CLO	USH	✓	3D	L
Dense [60]	2020	D	LR	PC	✓	LR	CLO	USHT	✓	3D	L
PixSet [35]	2021	DT	LR	PC	✓	MR	CLO	USP	✓	3D	M
RadarScenes [36]	2021	DTS	HR	PC	✓	SV	CO	USHT	✓	P_w	L
Pointillism [37]	2020	D	2LR	PC	✓	MR	CL	U	✓	3D	M
Zendar [38]	2020	D	SAR	ADC,RD,PC	✓	MR	CLO	U	✗	P_w	S
Radar Pre-CFAR Datasets for Detection											
CARRADA [39]	2020	DTS	LR	RAD	✓	SR	C	R	✗	2D,P_w,M	M
CRUW [41]	2021	D	LR	RAD	✓	USR	C	USHP	✗	P_o	L
RADDet [40]	2021	D	LR	RAD	✓	SR	C	US	✗	2D	M
RaDICaL [42]	2021	L	LR	ADC	✓	USR,SR	C_dO	USHIP	✗	2D	L
Ghent VRU [43]	2020	DS	LR	RAD	✓	SR	CL	U	✗	M	M
4D Radar Datasets for Detection											
Astyx [2]	2019	D	HR	PC	✓	MR	CL	SH	✗	3D	S
View-of-Delft [44]	2022	DT	HR	PC	✓	SR	CLO	U	✗	3D,T	S
RADIal [45]	2021	DS	HR	ADC,RAD,PC	✓	MR	CLO	USH	✗	P_o,M	M
TJ4DRadSet [46]	2022	DT	HR	PC	✓	LR	CLO	U	✗	3D, T	M
Radar Datasets for Localization											
ColoRadar [52]	2021	L	HR,LR	ADC,PC	✓	2USR	LO	SIT	✓	P_s	M
Oxford [47]	2020	L	SP	RA	✗	SV	CLO	U	✓	P_s	L
RADIATE [61]	2020	LDT	SP	RA	✗	SV	CLO	USHP	✓	2D,T, P_s	M
Mulran [48]	2020	L	SP	RA	✗	SV	LO	US	✗	P_s	M
Boreas [49]	2022	LD	SP	RA	✗	SV	CLO	S	✓	P_s,3D	L
Endeavour [51]	2021	L	LR	PC	✓	5LR	LO	S	✗	P_s	M
EU Long-term [50]	2020	L	LR	PC	✓	LR	CLO	U	✓	P_s	M
Radar Datasets for Other Tasks											
Ghost [55]	2021	DS	LR	PC	✗	LR	CLO	S	✗	P_w	M
SCORP [54]	2020	S	LR	ADC,RAD	✓	USR	C	P	✗	M	S
PREVENTION [53]	2019	DT	LR	PC	✓	1LR,2SR	CLO	UH	✓	2D	L

Task: 'D', 'T', 'L', 'S' stand for detection, tracking, localization, and segmentation; Type: 'LR', 'HR', 'SP', 'SAR' stand for low-resolution, high-resolution, spinning, and SAR; Range: 'SV', 'LR', 'MR', 'SR', 'USR' stand for surrounding view, long-range (<250m), middle-range (<180m), short-range (<50m), and ultra-short-range (<25m); Modalities: 'C', 'C_d' , 'L', 'O' stand for camera, RGBD camera, LiDAR, and odometry; Scenarios: 'U', 'S', 'H', 'P', 'T', 'R', 'I' stand for urban (city), suburban, highway, parking lot, tunnel, race track and indoor; Size: 'L', 'M', 'S' stand for large, medium, and small; Weather stands for adverse weather; Label: '2D','3D','T', 'P_w','P_o','P_s', 'M' stand for 2D bounding box, 3D bounding box, track ID, point-wise detection, object-level point, pose and segmentation mask.

3.2. Extrinsic Calibration

Multi-sensor extrinsic calibration requires calibration targets to be observed simultaneously by different modalities. The trihedral corner reflector is widely used for radar calibration because of its high RCS. Multiple reflectors are usually placed outdoor to avoid multi-path propagation. The difficulty lies in making the calibration target visible to both radar and other sensors. El Natour *et al*. [62] builds a calibration facility by placing one Luneburg lens and seven trihedral corner reflectors with known inter-distances. To make the reflectors visually detectable, they paint each surface with different colors. Peršić *et al*. [63] design a compact calibration target which can be simultaneously detected by camera, radar and LiDAR. As shown in fig. 5 (a), they place a triangle-shaped checkerboard pattern in front of a trihedral corner reflector. The checkerboard is made of styrofoam and is transparent over a large radio frequency range. So that the millimeter wave can penetrate it and detect the corner reflector behind it. In fig. 5 (b), Domhof *et al*. [64] design a styrofoam board with four circular holes and place a corner reflector at the back. These circular holes are more easily detected by the sparse LiDAR beam since they have no horizontal lines.



(a) Triangle target (b) Target with holes

Figure 5. Two types of radar calibration targets [63,64]. The front board is made of styrofoam. The red triangle is a radar corner reflector.

The extrinsic calibration of a 4D radar and other sensors can be easily done by modifying the classical LiDAR to camera calibration methods[65,66]. However, the calibration of conventional radar is a very difficult task, since it returns a 2D point cloud with no elevation resolution. This leads to the problem of vertical misalignment [63], which is defined as the angular deviation between radar plane and ground plane. Sugimoto *et al*. [67] move a corner reflector up and down to cross radar plane in multiple times. Then, the plane is determined by connecting the peaks with highest intensity in the sequence. Peršić *et al*. [63] propose a two-step optimization method to mitigate the uncertainty caused by the missing elevation angle. They model radar detections as arcs by extending their elevation angle. Similarly, they also convert 3D detections from other sensors to arcs by neglecting the elevation angle. In the first step, they optimize the reprojection error, which is the Euclidean distance of these projected arcs on the ground plane. In the second step, the parameters related to the elevation measurement are refined according to the RCS error. A second-order RCS model is built by fitting RCS measurements with elevation angles. Then, the L2 distance between the expected and measured RCS is minimized. Experiments show their method enables smaller vertical misalignment than Sugimoto's method. In order to improve efficiency, some targetless online calibration approaches [68,69] are proposed to leverage target trajectories for extrinsic estimation.

3.3. Data Labelling

Before introducing the labelling process, we first discuss the time synchronization problem. Different sensors can be synchronised using pulse per second (PPS) triggering signals from the GNSS receiver [70]. However, in most of radar datasets, sensors differ in their triggering time and sampling frequency. Some of them select one sensor as the lead, and choose the closest frames from other modalities for synchronization. Assuming a tolerable time offset of 50 ms, a vehicle with a relative speed of 20 m/s will lead to

an offset of 1 m. Therefore, it is necessary to compensate for synchronization errors in high-speed scenarios. Kaul *et al*. [71] design a pose chain method to interpolate inter-frame measurements. The translational and rotational transformations are determined by a constant velocity model and Spherical Linear Interpolation (SLERP) [72] respectively.

Labelling radar data is a difficult task. Both radar point clouds and pre-CFAR data are hard to interpret by human labellers. To reduce the labelling efforts, most of datasets adopt a semi-automatic labelling framework, which includes two steps: cross-modality pre-labelling and fine-tuning.

In the first step, a well-trained detector on other modalities is leveraged for radar labelling. For 3D tasks, radar point clouds can be annotated by 3D boxes predicted by the detector trained with images and LiDAR point clouds [2]. If we want to get point-wise annotations for radar point cloud, we can first predict a dense semantic map for the corresponding image using a visual segmentation network, such as mask R-CNN [73] or DeepLab V3 [74]. To avoid scale ambiguity, it is better to project the masked image to radar frames using depth measured by LiDAR [71] or stereo camera [40,41]. Then, each radar point can be associated with the corresponding semantic labels. Point-wise annotation of RAD tensors or RA maps is in a similar process. We can firstly use CFAR to detect peaks as detections, then annotate these point detections with the aligned visual semantic map. CRUW dataset [41] propose a post-processing method to obtain point-wise object annotations for RA maps. The authors define an object location similarity (OLS) metric which jointly considers the similarities in distance, scale and class. Then, they propose a location-based Non-Maximum Suppression (NMS) method that selects one object point out of the adjacent points based on the OLS metric. Compared to RA map, RD map alone is much more difficult to label. It needs both depth and radial velocity to associate an RD cell to a pixel or a LiDAR point. Radial velocity can be estimated by visual scene flow [75] or by tracking [39].

In the second step, manual inspection is required to correct pre-labelling errors. Identifying radar errors involves domain knowledge and therefore requires hiring of radar experts. As a result, building a high-quality, large-scale radar dataset is both time-consuming and financially expensive. To improve labelling efficiency, one way is to reduce the amount of data to be labelled. Dimitrievski *et al*. [76] leverage a tracking algorithm to interpolate annotations between key frames. The intermediate position is estimated by a Kalman filter with optical flow as measurements. Meyer *et al*. [2] adopt an active learning [77] framework to reduce labelling efforts in building the Astyx dataset. The core idea is to only label the most informative data. Specifically, they first label a small number of frames and train a detector with this data subset. The trained detector is then used to make predictions on the remaining unlabelled data. Next, the top N uncertain data are again manually labelled and added to the training subset. This process is repeated until convergence of the validation performance.

3.4. Data Augmentation

Data augmentation plays an essential role in improving generalization of deep learning models. It is well studied for image [78], LiDAR point cloud [79] and audio spectrogram [80], but overlooked in radar perception. According to the summary report of Radar Object Detection 2021 (ROD2021) Challenge [81], data augmentation techniques significantly improve the performance of RA-map-based radar detection. Considering the radar representation, we can divide the augmentation techniques into spectral and point cloud-based. Augmentation methods can also be featured by local or global depending on whether the entity being augmented is a single object or the entire scene.

Spectral augmentation techniques are used for radar pre-CFAR data. DANet [82] adopts several global augmentation techniques borrowed from computer vision to radar RA maps. The methods include mirroring, resizing, random combination, adding Gaussian noise, and temporal reversing. Although physical fidelity is not explicitly considered, the performance gain proves the effectiveness of these augmentation methods. RADIO [83]

implements four types of spectral augmentations, including attenuation, resolution change, adding speckle noise, and background shift. The first two methods are applied to a local patch around each detected object. The attenuation effect is approximated by dampening the cells according to an empirical relationship between the received power and range. The resolution change is modelled by nearest-neighbor interpolation according to the object size. The speckle noise can be approximated as multiplicative truncated exponential distribution [84] or multiplicative Gaussian noise [83]. Background shift is done by adding or subtracting a constant value to background cells. RAMP-CNN [85] applies global geometric augmentations to RA maps. They translate and rotate RA maps in Cartesian coordinate, then project back to the original polar coordinate. The out-of-boundaries areas are cropped off, and the blank areas are filled with background noises. Energy loss and antenna gain loss due to the transformation are compensated according to the radar equation.

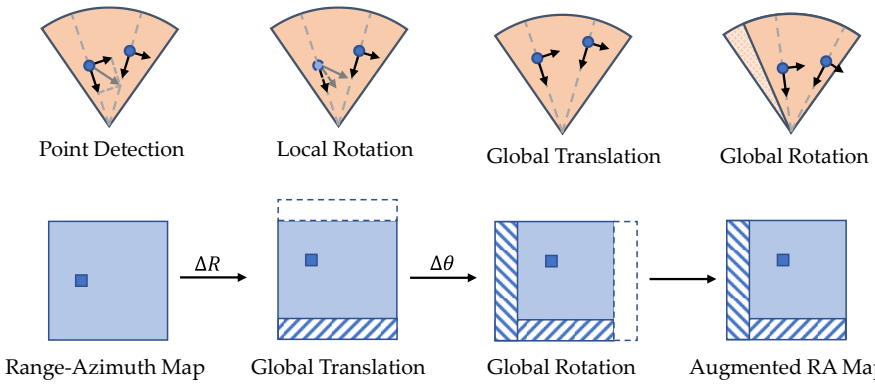


Figure 6. Radar data augmentation techniques. The Doppler velocity measured by radar is a scalar, so local rotation of the radar detection will cause a misalignment between the Doppler velocity and the true velocity. Global translation and rotation are free from such misalignment. When augmenting radar RA map, it is necessary to interpolate the background area and compensate the intensity of detection.

Point cloud augmentation aims to introduce invariance to geometric transformations and improve the signal-to-clutter ratio. Compared with spectral augmentation, point cloud augmentation methods can be easily extended to multi-modality by properly handling occlusion issues [86,87]. Geometric augmentation can be applied locally or globally, depending on whether the transformation is applied to a single target or the entire scene. For radar point cloud, the Doppler velocity and RCS need further consideration. As illustrated in fig. 6, rotating objects locally will affect the radial velocity, and rotating the radar point cloud globally will affect the ego-motion compensated radial velocity if the ego-motion is not rotated accordingly. Therefore, Palffy *et al*. [44] advise only use mirroring and scaling along the longitudinal axis as augmentation. Another applicable technique is the copy-paste augmentation, which copies the detected object from other frames and pastes it into the same location in the current frame as done in [88]. A limitation of these two methods is that they do not change the distribution of detections, while radar points are actually randomly distributed over the object in different frames. According to experiments [89], most of the radar detections are located in the proximity of the vehicle contour and wheel rims. The number of detections per object is inversely proportional to the distance, and the probability of detection on the contour depends heavily on the orientation. Simulation-based methods, which will be introduced in the next section, is more suitable to capture such randomness.

To handle with the sparsity issue, many works utilize augmentation to increase point cloud density. One simple method is accumulating radar points from multiple frames into the current frame. However, accumulation without motion compensation will lead to point cloud aliasing. Long *et al*. [90] compensates the accumulated radar point cloud with the

estimated full velocity, achieving better performance in bounding box regression. Plaffy *et al*. [44] augment the accumulated radar point cloud by appending a temporal index to each point as an additional channel. Along with the increased density, this index augmentation is expected to effectively retain the temporal information. Alternatively, Bansal *et al*. [37] leverages space coherent of two radar sensors to increase the point cloud density. They fuse point clouds from two radars with overlapping FoVs in a probabilistic manner. They firstly cluster the raw point clouds and then associate clusters from two radars by defining a distance-dependent potential function. Points with low confidence are filtered out as outliers, and the remaining points within the same cluster are coherently accumulated.

3.5. Synthetic Data

Synthetic datasets are widely used in computer vision [91,92] and LiDAR perception [93,94] for autonomous driving. Experiments [95] show the network trained with synthetic data can generalize well in the real-world. By using synthetic radar data, the labelling cost can be completely avoided. Moreover, simulation can be used to generate the safety-critical long-tail scenarios [96]. Physics-based simulation methods, such as ray tracing [97,98], are widely applied to generate synthetic radar point clouds. Experiments [98] show that ray tracing can successfully model the multi-path propagation and separability issue of close objects. However, it is difficult to capture the RCS variation in azimuth with current methods. Another type of simulation is to build a probabilistic model of radar detections, also known as model-based augmentation. The spatial distribution of radar detections over the vehicle can be approximated by the surface-volume model, including volcanormal measurement model [99], variational Gaussian mixture model (GMM) [99], and hierarchical truncated Gaussian (HTG) [100]. Model parameters can be learned from data. Using this model, we can augment new synthetic radar detections to real point clouds. It is arguable that what level of fidelity is necessary for downstream tasks. In [101], model-based and ray tracing methods are compared with respect to multiple target tracking. Experiments indicate that the ray-tracing-based model achieves the lowest simulation-to-reality gap.

There are some seminal works utilizing learning-based generative models for radar simulation. For example, deep stochastic radar model [102] adopts a conditional-VAE architecture. The encoder consists of two heads, one for RAD tensor and one for object list. The extracted features are concatenated and further processed with an MLP. The decoder generates a radar intensity map in polar grid conditioned on the encoded feature and random noise. Generative models can also be used in cross-modality data generation, for example, GAN-based LiDAR-to-radar generation [103], GAN-based radar-to-image generation [104] and VAE-based radar-to-image generation [105].

4. Radar Depth and Velocity Estimation

Radar can measure range and Doppler velocity, but both of them cannot be directly used for downstream tasks. The range measurements are sparse and therefore difficult to associate with their visual correspondences. The Doppler velocity is measured in radial axis and therefore cannot be directly used for tracking. In this section, we summarize depth completion and velocity estimation methods using radar point cloud.

4.1. Depth Estimation

Recently, pseudo-LiDAR-based visual object detection [106–108] has became a popular research topic. The core idea is to project pixels into a pseudo point cloud to avoid distortions induced by inverse projective mapping (IPM). The pseudo LiDAR detection is built on depth estimation. Visual depth estimation is a ill-posed problem because of the scale ambiguity. However, learning-based methods, either in supervised [109] or self-supervised [110], can successfully predict dense depth maps with camera only. Roughly speaking, these methods learn a priori knowledge of object size from the data and are therefore vulnerable to some data-related problems, such as sensitivity to input image quality [110] and learning

non-causal correlations, such as object and shadow correlations [111]. These limitations can be mitigated with the help of range sensors, such as LiDAR and radar. Depth completion is a sub-problem of depth estimation. It aims to recover a dense depth map for image using the sparse depth measured by range sensors. Compared to LiDAR, radar has advantages of low price, long range and robustness to adverse weather. Meanwhile, it faces the problems of noisy detections, no height measurements and sparsity. As shown in fig. 7, due to multi-path propagation, radar can see the non-line-of-sight highly reflective objects, such as wheel rims and occluded vehicles. In [112], the authors refer this phenomenon as the see through effect. It is beneficial in 3D coordinate, but brings difficulty in associating radar detections with visual objects in image view.

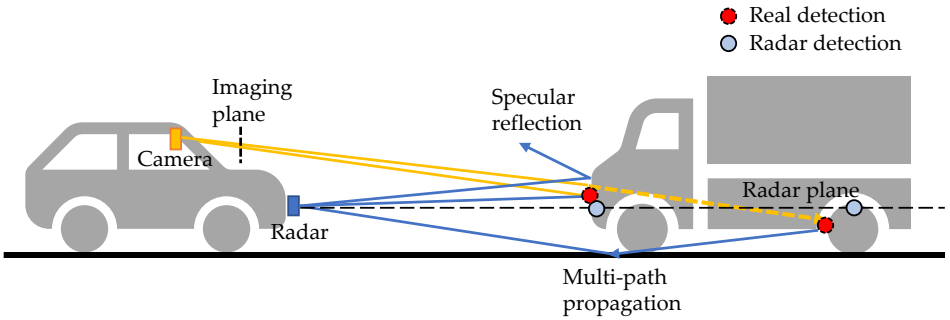
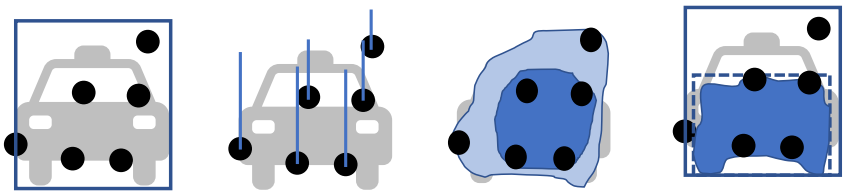


Figure 7. Radar range measurements. Off-the-shelf radars return detections on a 2D radar plane. The detections are sparsely spread on objects due to specular reflection. Due to multi-path propagation, radar can see through occlusions, and meanwhile this can cause some noisy detections.

The two-stage architecture is widely applied for image guided radar depth completion task. Lin *et al*. [113] adopt a two-stage coarse-to-fine architecture with LiDAR supervision. In the first stage, a coarse radar depth is estimated by an encoder-decoder network. Radar and image are processed independently by two encoders and fuse together in feature-level. Then, the decoder output a coarse dense depth map in image view. The predicted depth with large errors are filtered out according to a range-dependent threshold. Next, the original sensor inputs and the filtered depth map are sent to a second encoder-decoder to output a fine-grained dense map. In the first stage, the quality of association can be improved by expanding radar detections to better match visual objects. As shown in fig. 8 (b), Lo *et al*. [114] apply height extension to radar detections to compensate for the missed height information. A fixed height is assumed for each detection and is projected onto the image view according to the range. Then, the extended detections are sent to a two-stage achitecture to output a denoised radar depth map. Long *et al*. [115] propose a probabilistic association method to model the uncertainties of radar detections. As shown in fig. 8 (c), radar points are transformed into a multi-channel enhanced radar (MER) image, with each channel representing the expanded radar depth at a specific confidence level of association. In this way, the occluded detections and imprecise detections at the boundary are preserved but with a low confidence. Gasperini *et al*. [112] use radar as supervision to train a monocular depth estimation model. Therefore, they apply a strict filtering to only retain detections with high confidence. In the pre-processing, they remove clutters inside the bounding box that exceeded the range threshold, and discard points in the upper 50% and outer 20% of the box as well as the overlapping regions to avoid the see-through effect. All the background detections are also discarded. For association, they first apply a bilateral filtering, *i.e*., an edge preserving filtering, to constrain the expansion to be within the object boundary. They further clip the association map close to the edge to get rid of imprecise boundary estimations. To compensate for height information, they directly use the height of the bounding box as reference. Considering the complexity of the vehicle shape, they extend the detections to the lower third of its bounding box to capture the flat front surface of the vehicle.



(a) Radar detections (b) Height extension (c) Multi-channel (d) Box-based filtering

Figure 8. Radar detection expansion techniques. (b) Extend radar detections in height. (c) Build a probabilistic map, where the dark/light blue indicates channel with high/low confidence threshold. (d) Apply a strict filtering according to the bounding box, where only detections corresponding to the frontal surface are retained.

As ground truth, LiDAR has some inherent defects, such as sparsity, limited range and holes with no reflections. Long *et al*. [115] suggest to pre-process LiDAR points for better supervision. They accumulate multiple frames of LiDAR point clouds to improve density. Pixels with no LiDAR reaches are assigned zero values. Since LiDAR and camera do not share the same FoV, the LiDAR points projected to image view also have occlusion problem. Therefore, the occluded points are filtered out by two criterions, one is the difference between visual optical flow and LiDAR scene flow, and the other is the difference between segmentation mask and bounding boxes. Lee *et al*. [116] suggest to use both visual semantic mask and LiDAR as supervision signals. Visual semantic segmentation can detect smaller objects at a distance, thus compensating for the limited range of LiDAR. To extract better representations, they leverage a shared decoder to learn depth estimation and semantic segmentation concurrently. Both the LiDAR measurement and the visual semantic mask annotations are used as supervision. Accordingly, the loss function consists of three parts: a depth loss with LiDAR points as ground truth, a visual semantic segmentation loss and a semantic guided regularization term for smoothness.

Projecting radar to image view will lose the advantages of the see through effect. Alternatively, Niesen *et al*. [117] leverage radar RA maps for depth prediction. They use a short range radar with maximum range of 40 meters. Because of the low angular resolution, the azimuth smearing effect is obvious,*i.e*., the detections are smeared as a blurry horizontal line in RA maps. It is expected that fusion of image and RA map can mitigate this effect. Therefore, they use a two branch encoder-decoder network with radar RA map and image as inputs. A dense LiDAR depth map is used as ground truth. Different from the above methods that align LiDAR to image, they crop, downsample and quantize LiDAR detections to match radar's FoV and resolution. The proposed method is tested with their self-collected data. Although the effectiveness of RA map and point cloud is not compared, it provides a new direction to explore radar in depth estimation task.

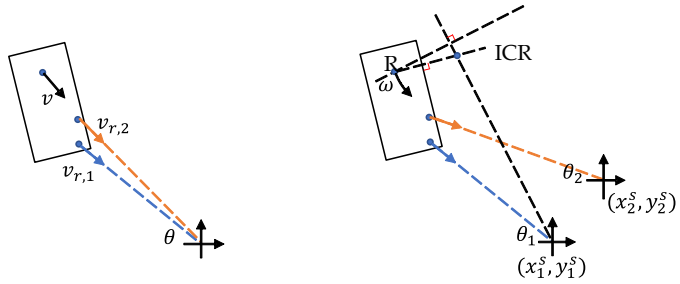
4.2. Velocity Estimation

For autonomous driving, velocity estimation is helpful for trajectory prediction and path planning. Radar can accurately measure the Doppler velocity, *i.e*., radial velocity in polar coordinate. If a vehicle moves parallel to ego-vehicle at a distance, its actual velocity can be approximated by the measured Doppler velocity. But this only applies in highway scenarios. On urban road, it is possible for an object to move tangentially while crossing the road, then its Doppler velocity will be close to zero. Therefore, Doppler velocity cannot replace full velocity. Recovering full velocity from the Doppler velocity needs two steps: first compensate the ego-motion, then estimate the tangential velocity. In the first step, the ego-motion can be estimated by visual-inertial odometry (VIO) and GPS. Radar-inertial odometry [118,119] can also be used in visual-degraded or GPS-denied environments. Then, the Doppler velocity is compensated by subtracting the ego-velocity. In the second step, the full velocity is estimated according to the geometric constraints. Suppose that radar observes several detections of an object and that the object is in linear motion. As

shown in fig. 9 (a), the relationship between the predicted linear velocity (v_x, v_y) and the measured Doppler velocity $v_{r,i}$ is given by

$$v_{r,i} = v_x \cos(\theta_i) + v_y \sin(\theta_i) \tag{7}$$

where the subscript i denotes the i-th detection, θ_i is the measured azimuth angle. By observing N detections per object, we can solve the linear velocity using the least square method. However, the L2 loss is not robust to outliers, such as clutter and mirco-Doppler motion of wheels. Kellner *et al*. [120] apply RANSAC to remove outliers then use Orthogonal Distance Regression (ODR) to find the optimal velocity.



(a) Linear motion model (b) Curvilinear motion model

Figure 9. Radar motion model. (a) Linear motion model needs multiple detections for the object. (b) Curvilinear motion model requires either two radars observe the same objects, or the determination of vehicle boundary and rear axle.

Although the linear motion model is widely used for its simplicity, it will generate large position errors for motion with high curvature [121]. Alternatively, as shown in fig. 9 (b), the curvilinear motion model is given by

$$v_{r,i} = \omega(y_c - y^S)\cos(\theta_i) - \omega(x_c - x^S)\sin(\theta_i) \tag{8}$$

where ω is the angular velocity, θ is the angle of the detected point,(x_c, y_c) represents the position of the instantaneous center of rotation (ICR) and (x^S, y^S) represents the known radar position. In order to decouple angular velocity and position of the ICR, we need at least two radar sensors that observe the same object. Then, we can transform eq. (8) into a linear form as

$$y_j^S \cos(\theta_{ji}) - x_j^S \sin(\theta_{ji}) = y_c \cos(\theta_{ji}) - x_c \sin(\theta_{ji}) - v_{ji}^D \omega^{-1} \tag{9}$$

where the subscript j denotes the j-th radar. Similarly, RANSAC and ODR can be used to find the unbiased solution of both angular velocity and position of the ICR [122]. For single radar setting, it is also possible to derive a unique solution of eq. (8) if we can correctly estimate the vehicle shape. According to the Ackermann steering geometry, the position of the ICR should be located on a line extending from the rear axle. By adding this constraints to eq. (8), the full velocity can be determined in closed form [123].

Above methods predict velocity in object level under the assumption of rigid motion. However, the micro-motion of object parts, such as the swinging arms of pedestrians, are also useful for classification. Capturing these non-rigid motions requires velocity estimation at the point level. It can be achieved by fusing with other modalities or by using temporal consistency between adjacent radar frames. Long *et al*. [90] estimate point-wise velocity by fusion of radar and camera. They first estimate the dense global optical flow and the association between radar points and image pixels through neural network models. Next, they derive the closed-form full velocity based on the geometric relationship between optical flow and Doppler velocity. Ding *et al*. [124] estimate the scene flow for 4D radar point cloud in a self-supervised learning framework. Scene flow is a 3D motion field and can be roughly considered as the linear velocity field. Their model consists of two steps: flow estimation and static flow refinement. In the flow estimation step, they adopt a

similar structure with PointPWCNet [125]. To compensate for the positional randomness of detections between frames, a cost-volume layer is utilized for patch-to-patch correlation. The features and correlation maps are then sent to a decoder network for flow regression. In the static flow refinement step, they assume that most radar detections are static and therefore use the Kabsch algorithm [126] to robustly estimate the ego-motion. They then filter out moving objects based on the coarse ego-motion, and apply the Kabsch algorithm again to all static points for fine-grained ego-motion estimation. The self-supervised loss consists of three parts: a radial displacement loss, which penalize errors between the estimated velocity projected along radial axis and the measured Doppler velocity, a soft Chamfer distance loss, which encourage temporal consistence between two consecutive point clouds, and a soft spatial smoothness loss, which encourage the spatial consistence for the estimated velocities with their neighbours. The soft version of loss is used to model spatial sparsity and temporal randomness of radar point cloud.

5. Radar Object Detection

Due to low resolution, classical radar detection algorithm has limited classification capability. In recent years, the performance of automotive radar has been greatly improved. At hardware level, next generation imaging radars can output high-resolution point clouds. At algorithmic level, neural networks show their potentials to learn better features from the dataset. In this section, we consider a broader definition of radar detection, including point-wise detection, 2D/3D bounding box detection and instance segmentation. We first introduce the classical detection pipeline and recent improvements on clustering and feature selection. As shown in fig. 10, neural networks can be applied to different stages in the classical pipeline. According to input data structure, we classify the deep radar detection into point-cloud-based and pre-CFAR-based. Radar point cloud and pre-CFAR data are similar to LiDAR point cloud and visual image respectively. Accordingly, architectures for LiDAR and vision tasks can be adapted for radar detection. We focus on how knowledge from the radar domain can be incorporated into these networks to address the low SNR problem.

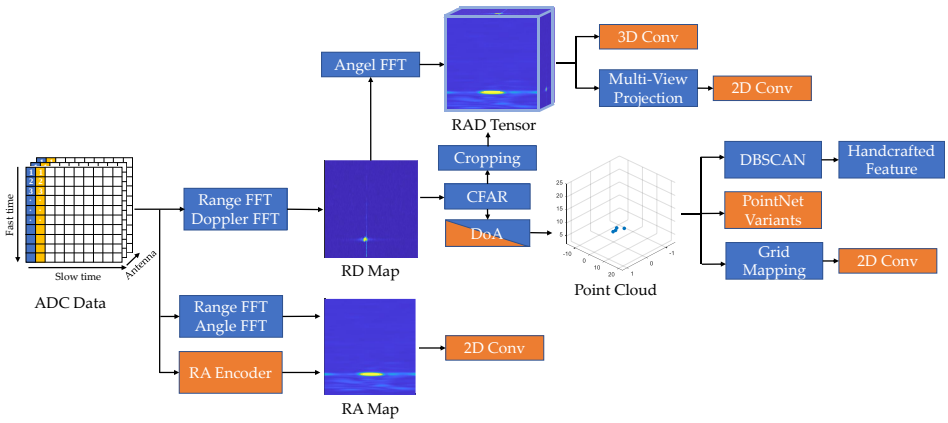


Figure 10. Overview of radar detection frameworks: Blue boxes indicate classical radar detection modules. Orange boxes represent AI-based substitutions.

5.1. Classical Detection Pipeline

As shown in fig. 10, conventional radar detection pipeline consists of four steps: CFAR detection, clustering, feature extraction, and classification. Firstly, a CFAR detector is applied to detect peaks in RD heatmap as a list of targets. Then, the moving targets are projected to Cartesian coordinate and clustered by DBSCAN [18]. Static targets are usually filtered out before clustering because they are indistinguishable from environmental clutter. Within each cluster, hand-crafted features, such as statistics of measurements and shape descriptors, are extracted and sent to a machine learning classifier. Improvements can be

made upon each of these four steps. CFAR is usually executed in an on-chip DSP, so the choice of method is restricted by hardware support. Cell-Averaging (CA) CFAR [17] is widely used due to its efficiency. It estimates the noise as the average power of neighbouring cells around the cell under test (CUT) within a CFAR window. A threshold is set to achieve a constant false alarm rate for Rayleigh distributed noise. The next generation high-resolution radar chips also support Order-Statistics (OS) CFAR [17]. It sorts neighbouring cells around the CUT according to the received power, and selects the k-th cell to represent the noise value. OS-CFAR has advantages in distinguishing close targets, but introduces a slightly increased false alarm rate and additional computational costs. More sophisticated CFAR variants are summarized in [127], but are rarely used in automotive applications. Deep learning methods can be used to improve noise estimation [128] and peak classification [127] in CFAR. Clustering is the most important stage in radar detection pipeline, especially for the next generation high resolution radar [129]. DBSCAN is favored for several reasons: It does not require a pre-specified number of clusters, it fits arbitrary shapes, and it runs fast [130]. Some works improve DBSCAN by explicitly considering characteristics of radar point cloud. Grid-based DBSCAN [131] suggests clustering radar points in a RA grid map to avoid the range-dependent resolution variations in Cartesian coordinate. Multi-stage clustering [132] proposes a coarse-to-fine two-stage framework to alleviate the negative impact of clutter. It applies a second cluster merging based on the velocity and spatial trajectory of clusters estimated from the first stage.

With the improvement of automotive radar resolution, radar target classification has become a hot research topic. For moving objects, the micro-Doppler velocity of moving components such as wheels and arms, can be useful for classification. To better observe these micro-motion, short-time Fourier transform (STFT) is applied to extract Doppler spectrograms. Different types of VRUs can be classified according to their micro-Doppler signatures[133,134]. For static objects, Cai *et al*. [135] suggest the use of statistical RCS and time-domain RCS as useful features for classification of vehicles and pedestrians. Some researchers work on exploiting a large number of features for better classification. Scheiner *et al*. [136] consider a large set of 98 features and use the heuristic-guided backward elimination for feature selection. They find range and Doppler features are most important for classification, while angle and shape features are usually discarded, probably because of the low angular resolution. Schumann *et al*. [137] compares the performance of random forest and LSTM for radar classification. Experiments show that LSTM with an input of 8-frame sequences performs slightly better than random forests, especially in the classification of classes with similar shape, such as pedestrians and pedestrian groups, and for false alarms. But LSTM is more sensitive to the amount of training examples. To cope with class imbalance in radar datasets, Scheiner *et al*. [138] suggest using classifier binarization techniques, which can be divided into two variants: one-vs-all (OVA) and one-vs-one (OVO). OVA trains N classifiers to separate one class from the other $N-1$ classes, and OVO trains $\binom{N}{2}$ classifiers for every class pair. During inference, the results are decided by max-voting.

5.2. Point Cloud Detector

End-to-end object detectors are expected to replace the conventional pipelines based on hand-crafted features. However, convolutional neural network is not well designed for sparse data structure[139]. It is necessary to increase the input density of radar point cloud for better performance. Dreher *et al*. [140] accumulate radar points into an occupancy grid mapping (OGM), then apply YOLOv3 [141] for object detection. Some works [142–144] utilize point cloud segmentation networks, such as PointNet [145] and PointNet++ [146], followed by a bounding box regression module for 2D radar detection. The original 3D point cloud input is replaced by a 4D radar point cloud with two spatial coordinates in x-y plane, Doppler velocity and RCS. Scheiner *et al*. [144] compare performances of two-stage clustering method, OGM-based method and PointNet-based method with respect to 2D detection. Experiments show that OGM-based method performs best, while PointNet-

based method performs far worse than others probably due to sparsity. Liu *et al*. [147] suggest that incorporating global information can help with the sparsity issue of radar point cloud. Therefore, they add a gMLP [148] block to each set abstraction layer in PointNet++. The gMLP block is expected to extract global features at an affordable computational cost.

Most radar detection methods only apply to moving targets, since static objects are difficult to classify due to low angular resolution. Schumann *et al*. [149] propose a scene understanding framework to detect both static and dynamic objects simultaneously. For static objects, they first build a RCS histogram grid map through temporal integration of multiple frames, and send it to an fully convolutional network (FCN) [150] for semantics segmentation. For dynamic objects, they adopt a two-branch recurrent architecture: one is the point feature generation module, which uses PointNet++ to extract features from the input point cloud. The other is the memory abstraction module, which learns temporal features from the temporal neighbors in the memorized point cloud. The resulted features are concatenated together and sent to a instance segmentation head. In addition, a memory update module is proposed to integrate targets into the memorized point cloud. Finally, static and dynamic points are combined into a single semantic point cloud. The proposed framework can successfully detect moving targets such as cars and pedestrians, as well as static targets like parked cars,infrastructures, poles and vegetation.

As 4D radars have gradually come to the market, radar point cloud density has increased considerably. A major advantage of 4D radar is that static objects can be classified based on elevation measurements without the need to build an occupancy grid map. Therefore, it is possible to train a single detector for both static and dynamic objects. Plaffy *et al*. [44] applies PointPillars [151] to 4D radar point clouds for 3D detection of multi-class road users. They find the performance can be improved by temporal integration and by introducing of additional features, such as elevation, Doppler velocity and RCS. Among them, the Doppler velocity is essential for detecting pedestrians and bicyclists. However, the performance of the proposed 4D radar detector (mAP 47.0) is still far inferior than their LiDAR detector on 64-beam LiDAR (mAP 62.1). They argue this performance gap comes from radar's poor ability in determining the exact 3D position of objects. RPFA-Net [152] improves PointPillars by introducing a Radar Pillar Features Attention (PFA) module. It leverages self-attention to extract the global context feature from pillars. The global features are then residually connected to the original feature map and sent to a CNN-based detection network. The idea behind is to explore the global relationship between objects for a better heading angle estimation. In fact, self-attention is basically an set operator, so it is well suited for sparse point clouds. Radar transformer *et al*. [153] is a classification network constructed entirely by self-attention modules. The 4D radar point cloud is first sent to an MLP network for input embedding. The following feature extraction network consists of two branches. In the local feature branch, it uses three stacked set abstraction modules [146] and vector attention modules [154] to extract hierarchical local features. In the global feature branch, the extracted local features at each hierarchy are concatenated with global feature map at the previous hierarchy and fed into a vector attention module for feature extraction. In the last hierarchy, a scalar-attention, *i.e*., the conventional self-attention, is used for feature integration. Finally, the feature map is sent to a classification head. Experiments show the proposed radar transformer outperforms other point cloud networks in terms of classification. The above two attention-based approaches show their potential in modeling the global context and extracting semantic information. Further works should focus on combine these two advantages into a fully attention-based detection network.

5.3. Pre-CFAR Detector

There are some attempts to explore the potential of pre-CFAR data for detection. Radar pre-CFAR data encode rich information of both targets and backgrounds, but is hard to interpret by human. Neural network is expected to better utilize these information. One option is to use neural network to replace CFAR [155] or DOA estimation[75,156]. Readers

can refer to [157] for a detailed survey of learning-based DOA estimation. Alternatively, there are also some efforts to perform end-to-end detection through neural networks. Deep radar detector [158] jointly trains two cascaded networks for CFAR and DOA estimation respectively. Zhang *et al*. [159] use stacked complex RD maps as input to a FCN for 3D detection. In order to remove the DC component in phase, they perform a phase normalization by using RD cells in the first receiver as normalizers. They argue that phase normalization is crucial for successful training. Rebut *et al*. [45] design a DDM-MIMO encoder with complex RD map as input. In DDM configuration, as illustrated in fig. 3, all Tx antennas transmit signals at the same time. Instead of performing waveform separation, they directly apply range FFT and Doppler FFT to ADC signals received by Rx antennas. In this way, targets detected from different Tx antennas should be located separately with a fixed Doppler shifts in RA map. To extract these features, they design a two-layer MIMO encoder, consisting of a dilated convolutional layer to separate Tx channels, followed by a convolutional layer to mix the information. This MIMO encoder are jointly trained with the following RA encoder, detection head and segmentation head.

In close-field applications that require large bandwidth and high resolution, RD maps are not suitable because the extended Doppler profile can lead to false alarms. RA map, on the other hand, does not suffer from the same problem. For each detection point on RA map, the micro-Doppler information in the slow time can be utilized for better classification. RODNet [41] uses complex RA maps as input for object detection. It performs range FFT followed by angle FFT to get a complex RA map for each sampled chirp. It is difficult to separate static clutter and moving objects using RA map alone without Doppler dimension. To utilize the motion information, it samples a few chirps within a frame. Then, the sequences of RA maps corresponding to these chirps are sent to a temporal convolution layer. Specifically, it first uses 1x1 convolutions along the chirp dimension to aggregate temporal information. Then, a 3D convolution layer is used to extract temporal features. Finally, the features are merged along the chirp dimension by max-pooling. Experiments indicate sampling 8 chirps out of 255 can achieve a comparable performance with using the full chirp sequences.

Training neural network to utilize phase information in complex RA or RD map is a difficult task. Alternatively, some works attempt to use the real-valued RAD tensor as input. A key issue in using the 3D RAD tensor as input is the curse of dimensionality. Therefore, many techniques are proposed to reduce the computational cost of 3D tensor processing. RADDet [40] normalizes and reshapes the RAD tensor to an image-like data structure. The Doppler dimension is treated as channel of 2D RA maps. Then, YOLO is applied to the RA map for object detection. One disadvantage is that this method fails to utilize the spatial distribution of Doppler velocities. Alternatively, 3D convolution can be used to extract features from all three dimensions in a 3D tensor, but requires huge computation and memory overheads [160]. RODNet [41] samples chirp sequences, as described above, to reduce input dimensionality. RTCNet [161] reduces tensor size by cropping a small cube around each point detected by CFAR, and then uses 3D CNN to classify these small cubes. However, its detection performance is limited by the CFAR detector. To fully exploit the information encoded in RAD tensors, some works [85,162,163] adopt the multi-view encoder-decoder architecture. Major *et al*. [162] and Ouaknine *et al*. [163] both utilize a similar multi-view structure. The RAD tensor is projected into three 2D views. Then, three decoders extract features from these views respectively. To fuse these features, Ouaknine *et al*. direclty concatenate three feature maps. Major *et al*. recover the tensor shape by duplicating these 2D feature maps along the missing dimension, then use a 3D convolution layer to fuse them. Next, the Doppler dimension is suppressed by pooling to recover the shape of RA feature map. Finally the fused feature maps are sent to a decoder for downstream segmentation tasks. Another difference is Major *et al*. use skip-connection while Ouaknine *et al*. adopt a ASPP [74] pathway to encode information from different resolutions. RAMP-CNN *et al*. [85] is also built in a multi-view architecture, but it uses three

encoder-decoders for feature map extraction. Their fusion method is similar to Major's but in 2D.

Radar pre-CFAR data are captured in a polar coordinate. For object detection, polar-to-Cartesian transformation is necessary to obtain the correct bounding box. Major *et al*. [162] compare three configurations for coordinate transformation: pre-processed input transformation, learning from neural networks, and transformation on middle-layer feature map. Experiments show applying explicit polar-to-Cartesian transformation to the last-layer feature map achieves the best performance, the implicit learning-based transformation is slightly worse and the pre-processed transformation is far inferior than other two. They attribute this poor performance to distorted azimuth side-lobes in the input. In fact, conventional 2D convolution is not the best choice for radar pre-CFAR data, since range, Doppler and azimuth dimension vary in their dynamic ranges and resolutions. Instead of 2D convolution, PolarNet [164] uses cascade of two 1D convolutions, including a column-wise convolution to extract range-dependent features, followed by a row-wise convolution to mix information from spatial neighbours. A similar idea is used in google's RadarNet [165] for gesture recognition. They first extract range-wise features then summarized them together in the later stage. Meyer *et al*. [166] use a isotropic graph convolution network (GCN) [167] to encode the RAD tensor and achieves more than 10% improvement in AP for 3D detection. They argue that the performance gain comes from the ability of GCN to aggregate information from neighboring nodes.

Incorporating temporal information is an effective way to improve the performance of pre-CFAR detectors. There are multiple ways to add temporal information to the network. Major *et al*. [162] use a convolutional LSTM layer to process a sequence of feature maps from the encoder network. Experiments indicate the temporal layer enables more accurate detection and significantly better velocity estimation. Ouaknine *et al*. [163] compare the performance between the static model with accumulated inputs and the temporal model with stacked inputs. For the static model, RAD tensors within 3 frames are accumulated into one single tensor, and fed to a multi-view encoder-decoder for segmentation. For the temporal model, RAD tensors within 5 frames are stacked to form a 4D tensor and then sent to a multi-view encoder-decoder. In each branch, multiple 3D convolution layers are used to leverage spatial-temporal information. The results show that the introduction of the temporal dimension can significantly improve detection performance. Pervsic *et al*. [68] discuss the effect of the number of stacked radar frames. They find too long frames will introduce many background clutter, which in turn makes the model difficult to learn target correspondences. According to their experiments, stacking of 5 frames is the most suitable choice. RODNet [41] investigate on stacking multiple frames on feature-level. It concatenates the extracted per-frame features and sends them to a 3D CNN layer. For motion compensation, they apply deformable convolution [168] on the chirp dimension in the first few layers. In addition, an inception module with different temporal length are used in the later layers. Despite the introduction of additional computational costs, these two temporal modules significantly improve the average precision. Li *et al*. [169] explicitly model the temporal relationship between features extracted from two consecutive frames using a attention module. Firstly, they stack RA maps in two orders, *i.e*.current frame on top and previous frame on top. Then, they use two encoders to extract features from these two inputs and concatenate the features together. A positional encoding is further added to compensate the positional imprecision. Next, the features are sent to a masked attention module. The mask is used to disable cross-object attention in the same frame. Finally, the temporally enhanced features are sent to a encoder for object detection. This attention-based approach is more semantically interpretable and avoids the locality constraint induced by convolution.

6. Sensor Fusion for Detection

Different sensors observe and represent an object with different features. Sensor fusion can be considered as the mapping of different modalities into a common latent space where different features of the same object can be associated together. In this section, we focus on sensor fusion for detection. We argue that the conventional taxonomy of fusion architectures into early (input), middle (feature) and late (decision) fusion is ambiguous for neural network based detection. For example, in the definition of late fusion, we cannot distinguish between ROI-level (without category information) fusion and object-level (with category information) fusion. Therefore, we explicitly classify fusion methods according to the fusion stage. This is beneficial because different fusion stages correspond to different levels of semantics, *i.e.*, the classification capabilities. As shown in fig. 11, we classify fusion architectures into four categories: input fusion, ROI fusion, feature map fusion and decision fusion.

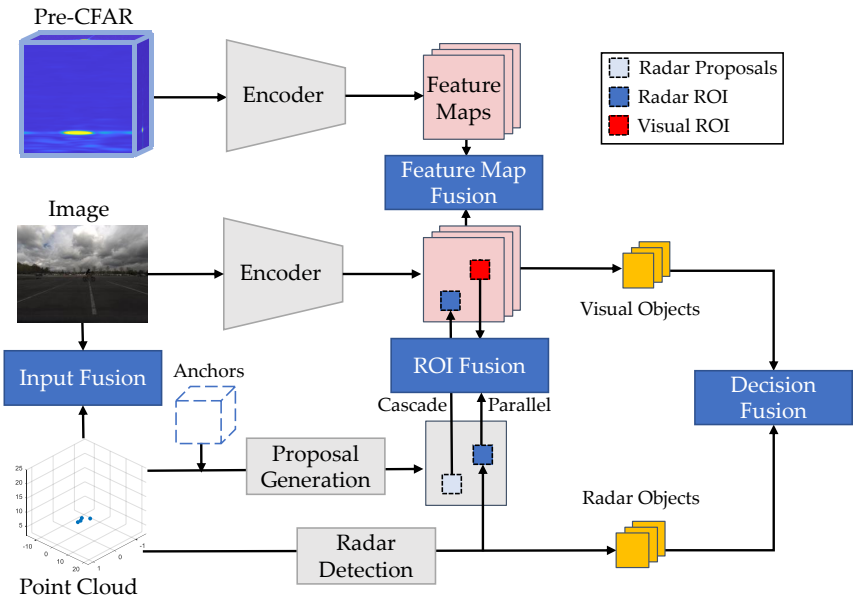


Figure 11. Overview of radar and camera fusion frameworks. We classify the fusion frameworks into input fusion, ROI fusion, feature map fusion and decision fusion. For ROI fusion, we further investigate two architectures: cascade fusion which projects radar proposals to image view, and parallel fusion which fuses radar ROIs and visual ROIs.

6.1. Input Fusion

Input fusion is applied to radar point cloud. It projects radar points into a pseudo-image with range, velocity, and RCS as channels [170,171]. Then, similar to a RGB-Depth image, the radar pseudo-image and the visual image are concatenated as a whole. Finally, a visual detector can be applied to this multi-channel image for detection. Input fusion does not make independent use of the detection capability of radar. In other words, the radar and vision modalities are tightly coupled. Assuming good alignment between modalities, it makes the network easier to learn joint feature embeddings. However, an obvious disadvantage is that the architecture is not robust to sensor failures.

The fusion performance depends on the alignment of radar detections with visual pixels. As mentioned in section 4.1, the difficulties lie in three aspects: Firstly, the radar point cloud is highly sparse. Many reflections from the surface are bounced away due to specular reflections. As a result, the detected points are sparsely distributed over the object. In addition to the sparsity, the lateral imprecision of radar measurements leads to further difficulties. The radar points can be out of the visual bounding box. The imprecision comes from different aspects, *e.g.* , imprecise extrinsic calibration, multi-path effects and

low angular resolution. The third limitation is that low-resolution radar does not provide height information. To address these difficulties, some association techniques are required. Relying on the network to implicitly learn association is a hard task, because the network tends to simply ignore the weak modality such as radar. The expansion methods described in section 4.1 can be applied as a pre-processing stage for input fusion. However, object detection does not require such a strict association as depth completion, so some of the expansion methods are too costly for real-time processing. Nobis *et al*. [170] utilize the light-weight height extension as pre-processing. Both Chadwick *et al*. [171] and Yadav *et al*. [172] add a one-layer convolution to radar input before concatenation. This convolutional layer can be considered as a lightweight version of the association network. Radar detections at different ranges requires different size of receptive field for association. Therefore, Nobis *et al*. [170] concatenate the radar pseudo-image with image feature maps at multi-scales.

6.2. ROI Fusion

ROI fusion is adapted from the classical two-stage detection framework [173]. Region of Interests (ROIs) can be considered as a set of object candidates without category information. The fusion architecture can be further divided into cascade fusion and parallel fusion. In cascade fusion, radar detections are directly used for region proposal. Radar points are projected into image view as the candidate locations for anchors. Then, the ROI is determined with the help of visual semantics. In the second stage, each ROI is classified and its position is refined. Nabati *et al*. [174] adopt two techniques to improve the anchor quality. They add offsets to anchors to model the positional imprecision of radar detections. To mitigate the scale ambiguity in the image view, they rescale anchor size according to the range measurements. In their following work [175], they directly propose 3D bounding boxes and then map these boxes to the image view. In this way, the rescaling step can be avoided. It is also possible to propose region on radar point cloud using visual ROIs. For example, CenterFusion [176] propose a frustum-based association to generate radar ROI frustums using visual bounding boxes.

Cascade fusion is particularly well suited for low-resolution radars, where the radar point cloud has a high detection recall but is very sparse. However, there are two potential problems with the cascade structure. Firstly, the performance is limited by the completeness of proposed ROIs in the first stage. In other words, if an object is missed, we cannot recover it in the second stage. The second problem is that the cascade structure cannot take advantage of modality redundancy. If the radar sensor is nonfunctional, the whole sensing system will fail. Therefore, it is necessary to introduce a parallel structure to ROI fusion. Nabati *et al*. [175] adopt a two-branch structure for ROI fusion. The radar and visual ROIs are generated independently. Then, the fusion module merges radar ROIs and visual ROIs by taking an set union, while the redundant ROIs are removed through NMS. To enable adaptive fusion of modalities, Kim *et al*. [177] propose a Gated Region of Interest Fusion (GRIF) module for ROI fusion. It first predicts a weight for each ROI through a convolutional-sigmoid layer. Then, the ROIs from radar and vision are multiplied by their corresponding weights and element-wise added together.

6.3. Feature Map Fusion

Feature-map fusion leverage the semantics from both radar and image. From section 5.3, we find that high resolution radars can provide sufficient semantic cues for classification. Therefore, feature-map fusion utilizes two encoders to map radar and image into the same latent space with high-level semantics. The detection frameworks are flexible, including one stage methods [178,179] and two-stage methods [33,180,181]. The one-stage method leverage two branches of neural networks to extract feature maps from radar and image respectively, and then concatenate the feature maps together. The two-stage fusion methods are adapted from the classical fusion architecture AVOD [182]. It firstly fuses the ROIs proposed from radar and image in the first stage. In the second stage, the fused ROIs are projected to radar and visual feature maps respectively. The feature maps inside

the ROIs are cropped and resized to an equal-sized feature crop. The feature crop pairs from radar and image are then fused by element-wise mean and sent to a detection head. Generally speaking, the two-stage method has better performance, but it is much slower than the one-stage method. Anchor free methods [183,184] further avoid the complicated computation related to anchor boxes, such as calculating IOU score during training.

Feature-map fusion allows the network to flexibly combine radar and visual semantics. However, the fusion network may face the problem of overlooking weak modalities and modality synergies [185]. Some training techniques are needed to force the network to learn from radar input. Nobis *et al*. [170] adopt a modality-wise dropout approach that randomly deactivates image branch during training. Lim *et al*. [178] use a weight freezing strategy to fix the weights of the pre-trained feature extractors when training the fusion branch. Experiments show that freezing only the image branch works best. However, fusion of multiple modalities is not guaranteed to always be better than using single modality. Sometimes we want the network to lower the weight of radar branch if it gives noisy inputs. To achieve adaptive fusion, Cheng *et al*. [186] adopt self-attention and global channel attention [187] in their network. The self-attention is used to enhance real target points and weaken clutter points. Then, the global attention module is applied to estimate modality-wise weights. Bijelic *et al*. [60] estimate the sensor entropy as the modality weight. For each modality, the entropy is evaluated pixel-wise as a weight mask. Then these weight masks are multiplied with the corresponding feature maps at each fusion layer.

6.4. Decision Fusion

Decision fusion assumes that objects are detected independently by different modalities and fuses them according to their spatial-temporal relationships. This structure realizes sensing redundancy at the system level and is therefore robust to modality-wise error. Due to the low resolution of radar, most existing studies do not explicitly consider the category information estimated by radar. In other words, they only fuse the location information from radar and vision branches, while retaining the category information estimated by vision. Since the next generation 4D radar can provide classification capabilities, it is expected that future fusion frameworks should consider both location and category information.

The location can be optimal fused in a tracking framework. Different objects are first associated and then sent to a Bayesian tracking module for fusion. Due to the low resolution of radar, association is difficult to achieve in some scenarios, *e.g.* , a truck splitting into two vehicles or two close objects merging into one. Such association ambiguity can be mitigated using a track-to-track fusion architecture [188]. By estimating tracks, temporal information can be leveraged to filter out false alarms and interpolate missed detections. Some researchers exploit deep learning to make a better association between radar and other modalities. RadarNet [183] propose an attention-based late fusion to optimize the estimated velocity. Firstly, they train a fiver-layer MLP with softmax to estimate the normalized association scores between each bounding box and its nearby radar detections. Then, they predict the velocity by weighted averaging the radar-measured velocities using the association scores. AssociationNet [189] attempts to map the radar detections to a better representation space in contrastive learning framework. It first projects radar objects and visual bounding boxes to the image plane as pseudo images. To utilize the visual semantics, they concatenate these pseudo images with the original image. Next, the concatenated images are sent to an encoder-decoder network to output a feature map. Representation vectors are extracted from the feature map according to the locations of radar detections. A contrastive loss is designed to pull together the representation vectors of positive samples and push away the representation vectors of negative examples. During inference, they compute the Euclidean distance between the representation vectors of all possible radar-visual pairs. The pairs with distance below the threshold are considered associative.

Category information, especially the conflict in category predictions, is difficult to handle in sensor fusion. BayesOD [190] proposes a probabilistic framework for fusing bounding boxes with category. The locations of bounding boxes are modelled by Gaussian

distributions. The category prior is modelled as a Dirichlet distribution, thereby allowing a Dirichlet posterior to be computed in closed form. Then, the bounding box with the highest categorical score is considered as the cluster center, while the other bounding boxes are treated as measurements. Finally, Bayesian inference is used to optimally fuse the location and category information of these bounding boxes. Probabilistic methods have their inherent shortage in modelling the lack of knowledge [191]. For example, a uniform distribution brings confusion of either the network has no confidence in its prediction or the input is indeed ambiguous for classification. In contrast, set-based methods have no such problem. Chavez *et al.* [192] leverage the evidential theory to fuse LiDAR, camera and radar. They consider the frame of discernment, *i.e.*, the set of mutually exclusive hypotheses, as $\Omega = \{pedestrains(p), bikes(b), cars(c), truck(t)\}$, and assign each possible hypothesis, *i.e.*, a subset of Ω, with a belief. In the case of object detection, possible hypotheses are selected according to sensor characteristics. For example, a car is sometimes confused as part of a truck. Thus, if a car is detected, evidence should be also put into the set $\{c, t\}$ and the set of ignorance Ω. Accordingly, we can assign the belief m to a car detection as

$$m(\{c\}) = \gamma_c \alpha_c, \quad m(\{c,t\}) = \gamma_c(1-\alpha_c), \quad m(\Omega) = 1-\gamma_c \tag{10}$$

where γ_c is a discounting factors to model the uncertainty of misdetection, and α_c is the accurateness, *i.e.*, the rate of correct predictions in car detecting. Suppose there are two sources of evidence S_1 and S_2 from different modalities. Each of these sources provides a list of detections as $A = \{a_1, a_2, ..., a_m\}$ and $B = \{b_1, b_2, ..., b_n\}$. Then, three propositions can be defined regarding the possible association of two detections a_i and b_j as

- $\{1\}$ if a_i and b_j are the same object;
- $\{0\}$ if a_i and b_j are not the same object
- $\{0,1\}$ for the ignorance of association.

The belief of association can be determined according to both location and category similarities. The evidence for location similarity is defined according to the Mahalanobis distance as

$$\begin{aligned} m^p_{a_i,b_j}(\{0\}) &= \alpha(1 - f(d_{a_i,b_j}))m^p_{a_i,b_j} \\ m^p_{a_i,b_j}(\{1\}) &= \alpha f(d_{a_i,b_j}) \quad m^p_{a_i,b_j}(\{1,0\}) = 1-\alpha \end{aligned} \tag{11}$$

where $f(d_{a_i,b_j}) = \exp(-\lambda d_{a_i,b_j}) \in [0,1]$ measure the similarity with respect to the Mahalanobis distance d_{a_i,b_j} and a scaling factor λ, and α is an evidence discounting factor. For the category similarity, two detections belong to the same category is too weak to provide evidence that they are the same object. However, if two detections are of different categories, it is reasonable to assign evidence to the proposition that they are not the same object. Accordingly, the evidence for category similarity is given by

$$\begin{aligned} m^c_{a_i,b_j}(\{0\}) &= \sum_{A \cap B = \emptyset} m^c_{a_i}(A) m^c_{b_j}(B) \quad \forall A, B \subset \Omega \\ m^c_{a_i,b_j}(\{1\}) &= 0, \quad m_{a_i,b_j}(\{0,1\}) = 1 - m^c_{a_i,b_j}(\{0\}) \end{aligned} \tag{12}$$

where the mass evidences are fused if no common category hypothesis is shared, *i.e.*, $A \cap B = \emptyset$. And the rest of the evidence are placed in the ignorance hypothesis. Finally, for each detection pairs, the category similarity and the location similarity are fused according to Yager's combination rule [193]. Evidential fusion provides an reliable framework for information fusion. However, it cannot be directly applied to neural-network-based detectors that make predictions on a single hypothesis. To address this problem, conformal prediction [194] can be used to generate confidence sets from a trained network using a small amount of calibration data.

7. Challenges

Although deep radar perception shows good performance on datasets, there are few studies investigating the generalization of these methods. In fact, some challenging situations are overlooked, but may prohibit the use of these methods in real-world scenarios. For example, the ghost objects caused by multi-path propagation are common in complex scenarios. Over-confidence is a general problem with neural networks. Since radar is always used for safety-critical applications, it is important to calibrate the detection network and output the predictive uncertainty. Even though we always refer to radar as an all-weather sensor, robustness in adverse weather is not well tested in many radar fusion methods. In this section, we present these three challenges and summarize some recent works that attempts to solve them.

7.1. Ghost Object Detection

Multi-path is a phenomenon in the physics of waves where a wave from a target travels to a detector through two or more paths. Because of the multi-path propagation, radar receives both direct reflections and indirect time-shifted reflections of targets. If the target reflections and the multi-path reflections occupy the same RD cell, the performance of DOA estimation is affected. Otherwise, if they occupy different cells, it can produce ghost targets in multi-path directions. In the latter case, since ghost detection has similar dynamics to the real target, it is difficult to eliminate them in the traditional detection pipeline. The multi-path effect can be classified into three types [195]. The first type is the reflection between ego-vehicle and targets. Therefore, the distance and velocity of clutter should be multiple times of the true measurement. The second type is the underbody reflection. It usually happens under the truck, resulting in points with longer distances. This see-through effect is sometimes beneficial, since occluded vehicles can be detected. The third type is mirrored ghost detections caused by the reflective surface. Because of the large wavelength of automotive 77GHz radar, many flat facilities, such as concrete walls, guardrails and noise cancellation walls, can be regarded as reflective surfaces. As shown in fig. 12, this kind of multi-path effect can be further categorized into type 1 and type 2 depending on whether the final reflection occurs on the target or the surface [196]. The number of reflections is referred to as the order of the multi path. Usually, only orders below 3 need to be considered, since higher order reflections return little energy due to signal diffusion.

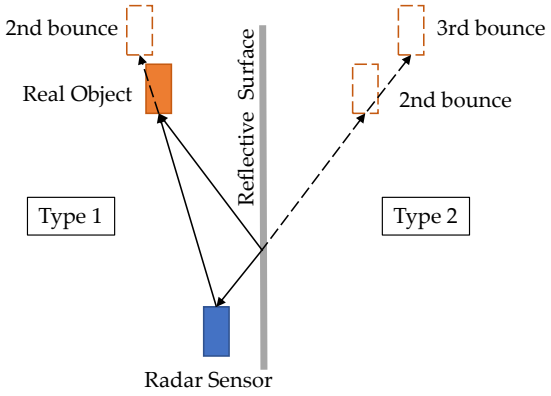


Figure 12. Multi-path effect: The solid orange box is the real object. Dotted boxes are ghost objects caused by multi-path propagation.

A high quality dataset is necessary for the performance evaluation of ghost detection. However, the labelling of ghost objects is a difficult task and requires expert knowledge. Chamseddine *et al.* [88] propose a method to automatically identify radar ghost objects by comparing with the LiDAR point cloud. However, LiDAR measurements are not perfect. It has its own inherent defects, such as sparsity, limited range and holes with no reflections.

Therefore, using LiDAR as ground truth could be sometimes problematic. In the Radar Ghost dataset [55], ghost objects are manually annotated with the help of a helper tool. This tool can automatically calculate the locations of potential ghosts based on real objects and reflective surfaces. As a result, four types of multi-path effects are annotated, including type-1 second-order bounces, type-2 second-order bounces, type-2 third-order bounces, and the other higher-order bounces. In addition, they also provide a synthetic dataset by overlaying objects from different frames within the same scene.

Unlike clutter, ghost objects cannot be filtered by temporal tracking because they have the same kinematic properties as real targets. Instead, they can be detected by geometric methods [195,197]. With a radar ghost dataset, it is also possible to train a neural network for ghost detection, such as PointNet-based methods [88] and PointNet++-based methods [196,198]. Because of the signal diffusion, the higher order reflections can be safely ignored. Thus, ghost objects usually occur in a ring-shaped region with the similar distance as the real target. Accordingly, Griebel *et al*. [198] design a ring grouping to replace the multi-scale grouping in PointNet++. The scene structure and relationship between detections are important cues to identify ghost objects. Garcia *et al*. [199] suggest the occupancy grid map can provide information of scene structure. Therefore, they use the occupancy grid map and the list of moving objects as inputs to FCN, to predict a heatmap of moving ghost detections. Wang *et al*. [200] propose to use multimodal transformers to capture the semantic affinity between ghost objects and real objects with LiDAR as reference. They design a multimodal attention block, which consists of two modules. The first one is a self-attention module for radar point cloud. It is expected to model the similarities of real objects and mirrored ones. The feature maps from the radar and LiDAR branches are then fused by a second multimodal attention module. This fusion module can be seen as calculating the correlation between LiDAR detections and real radar detections.

7.2. Uncertainty in Radar Detection

Learning-based radar detection shows its potential in classifying different road users. However, the performance evaluated on research datasets could be biased due to class imbalance and simple scenarios. Palffy *et al*. [44] summarize some failure cases for radar detection in VoD dataset: Two close pedestrians can be detected as one bicyclist. One large object, for example a truck or a bus, can be split into two smaller ones. Distant objects with few reflections may be missed by the detector. Strong reflections from metal poles and high curbs can mask real objects. Most of these failures come from the imperfection of radar sensors with respect to angular resolution and dynamic range. To make it worse, neural networks tend to be overconfident in their incorrect predictions [201]. For autonomous driving, the misspecified confidence in perception can leak to downstream tasks like sensor fusion and decision making, potentially leading to catastrophic failure. Patel *et al*. [202] investigate the class uncertainty of a learning-based radar classifier under different perturbations, including domain shift, signal corruptions and out-of-distribution data. Experiments indicate their baseline network are severely over-confident under these perturbations.

There are two kinds of uncertainty: Data uncertainty, also known as aleatoric uncertainty, is caused by noisy input. Model uncertainty, also known as epistemic uncertainty, is caused by insufficient or inappropriate training of the network. Sources of model uncertainty include three cases: covariate shift ($p(x)$ changes), label shift ($p(y)$ changes) and open set recognition (unseen y) [203]. The sum of data uncertainty and model uncertainty is referred to as predictive uncertainty. For the task of probabilistic object detection [204], uncertainties of two parameters are of interest: class uncertainty, which encodes the confidence in the classification, and spatial uncertainty, which represents the reliability of the bounding boxes. Class uncertainty can be seen as model uncertainty, while spatial uncertainty is more relevant to data uncertainty introduced by noisy input.

For classification tasks, the simplest way is to learn a function that maps the pseudo probability output by softmax layer into true probability. The true probability is defined

as the class-wise accuracy on the training set. This process is usually called network calibration. Since it is a post-processing method, both the model size and inference time are not affected. The calibration method is mainly concerned with the aleatoric part of the overall uncertainty [191]. To calibrate the radar classifier, Patel *et al*. [205] compares different post-processing techniques, including temperature scaling [201], latent Gaussian process (GP) [206] and mutual information maximization [207]. Mutual information maximization achieves the best balance between performance and inference time. Some recent researches indicate that soft-label augmentation techniques, such as label smoothing [208] and mixup [209,210], can effectively mitigate the over-confidence problem, thus helps network calibration. Patel *et al*. [205] suggests the use of label smoothing regularization in radar classification. The core idea is that the classifier should give lower confidence to distant objects with low received power. Therefore, they propose two label smoothing techniques to generate soft labels according to the range and the received power respectively. Experiments show that both of them can significantly improve the calibration performance.

In addition to calibrating the class uncertainty, we are also interested in estimating the spatial uncertainty in bounding box regression. Monte-Carlo Dropout [211] and Deep Ensembles [212] are popular in estimating predictive uncertainty. However, experiments [213] show that these methods only provide marginal improvements in object detection, but at a high cost. Direct modelling [214] is widely used to estimate the aleatoric uncertainty in bounding box regression. The idea is to let the network estimate both mean and variance of a prediction. The loss is constructed as

$$\mathcal{L}(\theta) = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{2\sigma(x_i)^2}\|y_i - f(x_i)\|^2 + \frac{1}{2}\log\sigma(x_i)^2 \tag{13}$$

where $\sigma(x_i)$ is the estimated variance which reduces the penalty with high variance and penalises high variance at the same time. Dong *et al*. [215] estimate the spatial uncertainty in radar detection using direct modelling. Experiments indicate that adding variance prediction for bounding box parameters can improve detection performance, especially under high IoU threshold.

7.3. Fusion in Adverse Weather

Adverse weather conditions, such as heavy rain, snow and fog, can be a significant threat to safe driving. Different sensors operate in different electromagnetic wavebands, thus having different robustness to environments. A comparison of the weather effects on different sensors can be found in [216]. Visual perception is susceptible to blur, noise, and brightness distortions [217,218]. In adverse weather, LiDAR suffers from reduced detection range and blocked view in powder snow [219], heavy rain [220], and strong fog [220]. In contrast, radar is more robust under adverse weather. The effect of weather on radar can be divided into attenuation and backscattering [221]. The attenuation effect decreases the received power of the signal, and the backscattering effect increases the interference at the receiver. Experiments [60,222–224] reveal attenuation and backscattering under dust, fog and light rain are negligible for radar, while the performance of radar degrades under heavy rainfall. Zang *et al*. [221] summarize the mathematical models for attenuation and backscattering effects of rain. They suggest the detection range of radar can be reduced by up to 45% under severe rainfall conditions (150 mm/h). For close targets with small RCS, the backscattering effect is more severe and can cause additional performance degradation.

Driving in adverse weather can be considered as a corner case [225] for autonomous driving. The concept of operational design domain(ODD) [226] is proposed to define the conditions under which autonomous vehicles are designed to operate safely. If the monitoring system [227] detects a violation of OOD requirements, control will be handed over to the driver. However, changes in the operational environment are usually rapid and unpredictable. Therefore, the hand-over mechanism is controversial in terms of safety. In the future, a fully autonomous vehicle (SAE Level 5) is expected to work under all environmental and weather conditions. However, most fusion methods are not designed

to explicitly consider weather effects. Networks trained in good weather may experience performance degradation in adverse weather.

There are some possible ways to adapt the network to different weather conditions. One way is to add a scene switching module [228], then use different networks for different weathers. This method is straightforward but introduces additional computational and memory costs. The other option is to add some dynamic mechanisms into the network. Qian *et al*. [33] add a two-stage attention block in the fusion module. They first apply self-attention to each modality and then mix them through cross attentions. Experiments show that the fusion mechanism performs robustly in foggy weather. They further investigate the domain generalization problem, *i.e*., training with a good-weather dataset and inference in foggy weather. The result shows a significant accuracy drop compared with training with data in both good and foggy weather, indicating that their model relies on data to generalize. Malawade *et al*. [229] propose a gating strategy to rank each modality and pick the top 3 reliable modalities for fusion. They compares three types of gating methods: knowledge-based, CNN-based and attention-based gating. Knowledge-based gating use a set of pre-defined modality-wise weights for each weather condition, while CNN-based and attention-based learn the weights from data. Experiments on RADIATE dataset [61] indicate gating methods outperform fusion methods under adverse weather, and attention-based gating can achieve the best performance. Alternatively, Bijelic *et al*. [60] proposes an entropy-steered fusion network which uses the sensor entropy as modality-wise weights. Specifically, they use a deep fusion architecture that continuously fuses feature maps from different modalities. The pixel-wise entropy is used as the attention map for each sensor branch.Since the entropy map is conditioned only on sensor inputs, the fusion network can perform robustly in unseen adverse weather. According to uncertainty theory, sensor entropy can be considered as a measure of data uncertainty. To utilize both data and model uncertainties, Ahuja *et al*. [230] propose an uncertainty-aware fusion framework. They leverage a decision-level fusion architecture and expect each branch to output both data uncertainty and model uncertainty. A gating function is used to apply a weighted average to each modality according to the predicted uncertainty. Then, they design two modules to handle the data with high uncertainty. One for failure detection. A sensor with data uncertainty consistently above a threshold is considered to be malfunctioning. The other is used for continuous learning. Data with model uncertainty above a threshold will be added to the training set for continuous learning.

8. Future Research Directions

In this paper, we summarize the recent developments on deep radar perception. As we can see, many research efforts have focused on developing models for detection tasks. However, there are also some unexplored research topics or fundamental questions to be addressed. In this section, we propose some interesting research directions to the automotive radar community.

8.1. High Quality Dataset

Deep learning revolution started with the introduction of ImageNet dataset [231]. However, radar perception has not yet seen its ImageNet moment. Although many datasets exist, they differ in scale, resolution, data representation, scenario, and labelling granularity. The granularity and quality of labelling is also a key issue for radar datasets. Therefore, it is hard to fairly compare different models trained on different datasets. Since the introduction of 4D imaging radar to the market, we anticipate an urgent need for datasets with high quality annotations and diverse scenes.

8.2. Radar Domain Knowledge

In the absence of high-quality datasets, we need to avoid treating AI in radar as a data fitting game. It is essential to exploit domain knowledge to develop a generalizable perception model. Radar domain knowledge need to be considered at many stages, such as

labelling, data augmentation, model structure, training techniques and evaluation metrics. Take ghost detection as an example. From a data perspective, we need to use our expert knowledge to label ghost objects [55]. From a network perspective, we can design an attention module [200] or utilize graph convolution [166] to model the relationship between real and ghost objects. We hope that researchers put more focus on solving these critical problems in radar perception.

8.3. Uncertainty Quantification

As introduced in section 7.2, uncertainty quantification is important for applying AI in safety-critical applications. Due to the low SNR of radar data and the small size of radar datasets, both high data and model uncertainties are expected for CNN-based radar detectors. However, there is still very little work touching on this topic. Although many uncertainty quantification methods have been proposed, they are not necessarily helpful for a specific task. For example, Feng *et al*. [232] find that sampling-based methods are not useful for visual object detection. Similarly, we need empirical experiments and theoretical explanations to demonstrate the necessity and effectiveness of uncertainty quantization methods for radar perception.

8.4. Motion Forecasting

An overlooked feature of radar is Doppler velocity. In addition to being a feature of moving road users, Doppler velocity is valuable for motion forecasting. Motion forecasting is a popular research topic in autonomous driving [233]. By accurately estimating the motion of road users, the down-stream path planning module can better react to future interactions. Lin *et al*. [234] predict trajectories by building a constant velocity model with binarized RA maps as input. However, experiments show that the constant velocity model performs poorly in predicting vehicle trajectories [233]. As mentioned in section 4, a second-order nonlinear motion model can be developed using the measured Doppler velocity. We believe that radar has great potential to play an important role in motion forecasting.

8.5. Interference Mitigation

For FMCW radar, mutual interference is a challenging task to solve. It occurs when multiple radars operate simultaneously in direct line of sight [235]. Depending on if the chirp configuration, *i.e.*slope and chirp duration, is same between interferer and victim radar, interference can be classified as coherent and incoherent [236]. Coherent interference occurs when the same chirp configuration is used and leads to ghost detections. Incoherent interference is caused by different types of chirps, resulting in significantly increased noise floor, masked weak target and thus reduced probability of detection. In reality, partially coherent interference is more widely seen where interferer has a slightly different chirp configuration. Oyedare *et al*. [237] summarize deep learning methods for interference mitigation. Although these methods achieve better performance than classical zeroing methods, they are generally designed for specific types of disturbances and require significant computational costs. Future research should consider interference mitigation and downstream tasks (e.g., detection) as a whole, and build an end-to-end learning framework to optimize them together.

9. Conclusions

The purpose of this review article is to provide a big picture of deep radar perception. We first summarize the principles of radar signal processing. Then, we present a detailed summary of radar datasets for autonomous driving. To encourage researchers to build their own datasets, we also present methods for calibration and labelling. We further investigate data augmentation and synthetic radar data to improve data diversity. Radar can be used for depth completion and velocity estimation. For ease of depth completion, several expansion methods are introduced to better associate the radar detections with the

image pixels. The full velocity can be recovered as a geometric optimization problem or in a self-supervised learning way.

Radar detection is the main focus of this paper. We classify deep radar detectors into point-cloud-based and pre-CFAR-based. PointNet variant networks and multi-view encoedr-decoders are popular choices for radar point clouds and radar tensors respectively. By increasing spatial density and exploiting temporal information, significant performance improvements can be achieved. Some new operators, such as depth-wise convolution, attention and graph convolution, are leveraged for larger receptive field. In practical applications, radar is often fused with cameras and LIDAR. We classify fusion frameworks into four categories. Input fusion requires a lightweight pre-processing to explicitly handle radar position imprecision. Cascaded ROI fusion is not robust to sensor failures, while parallel ROI fusion improves it. Feature map fusion provides the network with greater flexibility to combine radar and visual semantics, but requires specific training techniques for effective learning. Decision fusion takes advantage of modal redundancy and is therefore popular in real-world applications. Location information can be robustly fused in a track-to-track architecture or with the help of network semantics. Category information can be fused with Bayesian inference or evidence theory.

We summarize three challenges for deep radar perception. Firstly, multi-path effects need to be explicitly considered in object detection. Secondly, we need to alleviate the problem of overconfidence in radar classification and estimate the uncertainty in bounding box regression. Thirdly, the fusion architecture should have adaptive mechanisms to take full advantage of radar's all-weather capabilities. Finally, some future research directions are proposed. There is an urgent need for high quality radar datasets. Radar domain knowledge and uncertainty quantification can help us to develop a generalizable AI model. Considering the perceptual system as a whole, we can extend the end-to-end learning framework forward, *i.e.*, joint learning with interference mitigation, or backward, *i.e.*, predicting motion.

Author Contributions: Conceptualization, Y.Z., L.L., H.Z.; investigation, Y.Z.; writing—original draft preparation, Y.Z.; writing—review and editing, Y.Z.; visualization, Z.Y.; supervision, M.LB., L.Y., Y.Y.; project administration, Y.Y. All authors have read and agreed to the published version of the manuscript.

Funding: This research was partially funded by XJTLU-JITRI Academy of Industrial Technology, Institute of Deep Perception Technology (IDPT) and the Research Enhancement Fund of XJTLU (REF-19-01-04).

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

RA	Range-Azimuth
RD	Range-Doppler
RAD	Range-Azimuth-Doppler
RCS	Radar Cross Section
FoV	Field of View
BEV	Bird's Eye View
FMCW	Frequency-Modulated Continuous-Wave
IF	Intermediate Frequency
TDM	Time-Division Multiplexing
DDM	Doppler-Division Multiplexing
SNR	Signal-to-Noise Ratio
VRU	Vulnerable Road User
ICR	Instantaneous Center of Rotation
OGM	Occupancy Grid Mapping
FCN	Fully Convolutional Network
ROI	Region of Interest
NMS	Non-Maximum Suppression
ODD	Operational Design Domain
RFS	Random Finite Set

Appendix A

Table A1. Parameters used in the radar signal processing section.

Parameter	Meaning	Parameter	Meaning
c	Light Speed (m/s^2)	N_{Tx}	Number of Tx Antennas
λ	Wavelength (m)	N_{Rx}	Number of Rx Antennas
f_c	Carrier Frequency (Hz)	d	Inter Antenna Spacing (m)
B	Sweep Bandwidth (dB)	D	Array Aperture (m)
S	Chirp Slope	P_t	Transmit Power (dBW)
N_c	Number of Chirps	G	Antenna Gain (dB)
T_c	Chirp Duratoin (s)	P_{min}	Minimum Detectable Power (dBw)
T_f	Frame Duration (s)	σ	Radar Cross Section (dBm^2)
B_{IF}	IF Bandwidth (dB)	SNR	Signal-to-Noise Ration

References

1. Karpathy, A. Keynotes at CVPR Workshop on Autonomous Driving. https://cvpr2021.wad.vision/, 2021.
2. Meyer, M.; Kuschk, G. Automotive radar dataset for deep learning based 3d object detection. 2019 16th European Radar Conference (EuRAD). IEEE, 2019, pp. 129–132.
3. Zhou, T.; Yang, M.; Jiang, K.; Wong, H.; Yang, D. MMW Radar-Based Technologies in Autonomous Driving: A Review. *Sensors* **2020**, *20*, 7283.
4. Abdu, F.J.; Zhang, Y.; Fu, M.; Li, Y.; Deng, Z. Application of Deep Learning on Millimeter-Wave Radar Signals: A Review. *Sensors* **2021**, *21*, 1951.
5. Scheiner, N.; Weishaupt, F.; Tilly, J.F.; Dickmann, J. New Challenges for Deep Neural Networks in Automotive Radar Perception. In *Automatisiertes Fahren 2020*; Springer, 2021; pp. 165–182.
6. Wei, Z.; Zhang, F.; Chang, S.; Liu, Y.; Wu, H.; Feng, Z. MmWave Radar and Vision Fusion for Object Detection in Autonomous Driving: A Review. *arXiv preprint arXiv:2108.03004* **2021**.
7. Tang, X.; Zhang, Z.; Qin, Y. On-road object detection and tracking based on radar and vision fusion: A review. *IEEE Intelligent Transportation Systems Magazine* **2021**.
8. Ravindran, R.; Santora, M.J.; Jamali, M.M. Multi-Object Detection and Tracking, Based on DNN, for Autonomous Vehicles: A Review. *IEEE Sensors Journal* **2021**, *21*, 5668–5677.

9. Hakobyan, G.; Yang, B. High-performance automotive radar: A review of signal processing algorithms and modulation schemes. *IEEE Signal Processing Magazine* **2019**, *36*, 32–44.
10. Ramasubramanian, K.; Instruments, T. Using a complex-baseband architecture in FMCW radar systems. *Texas Instruments* **2017**, *19*.
11. Rao, S. MIMO Radar. *Application Report SWRA554A, Texas Instruments* **2017**.
12. Sun, H.; Brigui, F.; Lesturgie, M. Analysis and comparison of MIMO radar waveforms. 2014 International Radar Conference. IEEE, 2014, pp. 1–6.
13. Sun, S.; Petropulu, A.P.; Poor, H.V. MIMO radar for advanced driver-assistance systems and autonomous driving: Advantages and challenges. *IEEE Signal Processing Magazine* **2020**, *37*, 98–117.
14. Bechter, J.; Roos, F.; Waldschmidt, C. Compensation of motion-induced phase errors in TDM MIMO radars. *IEEE Microwave and Wireless Components Letters* **2017**, *27*, 1164–1166.
15. Gupta, J. High-End Corner Radar Reference Design. *Design Guide TIDEP-01027, Texas Instruments* **2022**.
16. Rebut, J.; Ouaknine, A.; Malik, W.; Pérez, P. RADIal Dataset. https://github.com/valeoai/RADIal, 2022.
17. Richards, M.A. *Fundamentals of radar signal processing*; Tata McGraw-Hill Education, 2005.
18. Schubert, E.; Sander, J.; Ester, M.; Kriegel, H.P.; Xu, X. DBSCAN revisited, revisited: why and how you should (still) use DBSCAN. *ACM Transactions on Database Systems (TODS)* **2017**, *42*, 1–21.
19. Muckenhuber, S.; Museljic, E.; Stettinger, G. Performance evaluation of a state-of-the-art automotive radar and corresponding modeling approaches based on a large labeled dataset. *Journal of Intelligent Transportation Systems* **2021**, pp. 1–20.
20. Gamba, J. *Radar signal processing for autonomous driving*; Springer, 2020.
21. Dham, V. Programming chirp parameters in TI radar devices. *Application Report SWRA553, Texas Instruments* **2017**.
22. Hasch, J.; Topak, E.; Schnabel, R.; Zwick, T.; Weigel, R.; Waldschmidt, C. Millimeter-wave technology for automotive radar sensors in the 77 GHz frequency band. *IEEE Transactions on Microwave Theory and Techniques* **2012**, *60*, 845–860.
23. Lim, T.Y.; Markowitz, S.; Do, M.N. RaDICaL Dataset SDK. https://github.com/moodoki/radical_sdk, 2021.
24. Lim, T.Y.; Markowitz, S.; Do, M.N. IWR Raw ROS Node. https://github.com/moodoki/iwr_raw_rosnode, 2021.
25. Mostafa, A. pyRAPID. http://radar.alizadeh.ca, 2020.
26. Pan, E.; Tang, J.; Kosaka, D.; Yao, R.; Gupta, A. OpenRadar. https://github.com/presenseradar/openradar, 2019.
27. Constapel, M.; Cimdins, M.; Hellbrück, H. A Practical Toolbox for Getting Started with mmWave FMCW Radar Sensors. Proceedings of the 4th KuVS/GI Expert Talk on Localization, 2019.
28. Gusland, D.; Christiansen, J.M.; Torvik, B.; Fioranelli, F.; Gurbuz, S.Z.; Ritchie, M. Open Radar Initiative: Large Scale Dataset for Benchmarking of micro-Doppler Recognition Algorithms. 2021 IEEE Radar Conference (RadarConf21). IEEE, 2021, pp. 1–6.
29. Visentin, T. *Polarimetric radar for automotive applications*; Vol. 90, KIT Scientific Publishing, 2019.
30. Gottinger, M.; Hoffmann, M.; Christmann, M.; Schütz, M.; Kirsch, F.; Gulden, P.; Vossiek, M. Coherent automotive radar networks: The next generation of radar-based imaging and mapping. *IEEE Journal of Microwaves* **2021**, *1*, 149–163.
31. Laribi, A.; Hahn, M.; Dickmann, J.; Waldschmidt, C. Performance investigation of automotive SAR imaging. 2018 IEEE MTT-S International Conference on Microwaves for Intelligent Mobility (ICMIM). IEEE, 2018, pp. 1–4.
32. Adams, M.; Adams, M.D.; Jose, E. *Robotic navigation and mapping with radar*; Artech House, 2012.
33. Qian, K.; Zhu, S.; Zhang, X.; Li, L.E. Robust Multimodal Vehicle Detection in Foggy Weather Using Complementary Lidar and Radar Signals. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 444–453.
34. Caesar, H.; Bankiti, V.; Lang, A.H.; Vora, S.; Liong, V.E.; Xu, Q.; Krishnan, A.; Pan, Y.; Baldan, G.; Beijbom, O. nuscenes: A multimodal dataset for autonomous driving. Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2020, pp. 11621–11631.
35. Déziel, J.L.; Merriaux, P.; Tremblay, F.; Lessard, D.; Plourde, D.; Stanguennec, J.; Goulet, P.; Olivier, P. PixSet: An Opportunity for 3D Computer Vision to Go Beyond Point Clouds With a Full-Waveform LiDAR Dataset. *arXiv preprint arXiv:2102.12010* **2021**.
36. Schumann, O.; Hahn, M.; Scheiner, N.; Weishaupt, F.; Tilly, J.F.; Dickmann, J.; Wöhler, C. RadarScenes: A real-world radar point cloud data set for automotive applications. *arXiv preprint arXiv:2104.02493* **2021**.
37. Bansal, K.; Rungta, K.; Zhu, S.; Bharadia, D. Pointillism: Accurate 3d bounding box estimation with multi-radars. Proceedings of the 18th Conference on Embedded Networked Sensor Systems, 2020, pp. 340–353.
38. Mostajabi, M.; Wang, C.M.; Ranjan, D.; Hsyu, G. High-Resolution Radar Dataset for Semi-Supervised Learning of Dynamic Objects. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, 2020, pp. 100–101.
39. Ouaknine, A.; Newson, A.; Rebut, J.; Tupin, F.; Pérez, P. CARRADA Dataset: Camera and Automotive Radar with Range-Angle-Doppler Annotations. *arXiv preprint arXiv:2005.01456* **2020**.
40. Zhang, A.; Nowruzi, F.E.; Laganiere, R. RADDet: Range-Azimuth-Doppler based radar object detection for dynamic road users. 2021 18th Conference on Robots and Vision (CRV). IEEE, 2021, pp. 95–102.
41. Wang, Y.; Jiang, Z.; Li, Y.; Hwang, J.N.; Xing, G.; Liu, H. RODNet: A Real-Time Radar Object Detection Network Cross-Supervised by Camera-Radar Fused Object 3D Localization. *IEEE Journal of Selected Topics in Signal Processing* **2021**, *15*, 954–967.
42. Lim, T.Y.; Markowitz, S.; Do, M.N. RaDICaL: A Synchronized FMCW Radar, Depth, IMU and RGB Camera Data Dataset with Low-Level FMCW Radar Signals. *IEEE Journal of Selected Topics in Signal Processing* **2021**.

43. Dimitrievski, M.; Shopovska, I.; Van Hamme, D.; Veelaert, P.; Philips, W. Weakly supervised deep learning method for vulnerable road user detection in FMCW radar. 2020 IEEE 23rd International Conference on Intelligent Transportation Systems (ITSC). IEEE, 2020, pp. 1–8.
44. Palffy, A.; Pool, E.; Baratam, S.; Kooij, J.; Gavrila, D. Multi-class Road User Detection with 3+ 1D Radar in the View-of-Delft Dataset. *IEEE Robotics and Automation Letters* **2022**.
45. Rebut, J.; Ouaknine, A.; Malik, W.; Pérez, P. Raw High-Definition Radar for Multi-Task Learning. *arXiv preprint arXiv:2112.10646* **2021**.
46. Zheng, L.; Ma, Z.; Zhu, X.; Tan, B.; Li, S.; Long, K.; Sun, W.; Chen, S.; Zhang, L.; Wan, M.; et al. TJ4DRadSet: A 4D Radar Dataset for Autonomous Driving. *arXiv preprint arXiv:2204.13483* **2022**.
47. Barnes, D.; Gadd, M.; Murcutt, P.; Newman, P.; Posner, I. The oxford radar robotcar dataset: A radar extension to the oxford robotcar dataset. 2020 IEEE International Conference on Robotics and Automation (ICRA). IEEE, 2020, pp. 6433–6438.
48. Kim, G.; Park, Y.S.; Cho, Y.; Jeong, J.; Kim, A. Mulran: Multimodal range dataset for urban place recognition. 2020 IEEE International Conference on Robotics and Automation (ICRA). IEEE, 2020, pp. 6246–6253.
49. Burnett, K.; Yoon, D.J.; Wu, Y.; Li, A.Z.; Zhang, H.; Lu, S.; Qian, J.; Tseng, W.K.; Lambert, A.; Leung, K.Y.; et al. Boreas: A Multi-Season Autonomous Driving Dataset. *arXiv preprint arXiv:2203.10168* **2022**.
50. Yan, Z.; Sun, L.; Krajník, T.; Ruichek, Y. EU long-term dataset with multiple sensors for autonomous driving. 2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). IEEE, 2020, pp. 10697–10704.
51. Huang, R.; Zhu, K.; Chen, S.; Xiao, T.; Yang, M.; Zheng, N. A High-precision and Robust Odometry Based on Sparse MMW Radar Data and A Large-range and Long-distance Radar Positioning Data Set. 2021 IEEE International Intelligent Transportation Systems Conference (ITSC). IEEE, 2021, pp. 98–105.
52. Kramer, A.; Harlow, K.; Williams, C.; Heckman, C. ColoRadar: The Direct 3D Millimeter Wave Radar Dataset. *arXiv preprint arXiv:2103.04510* **2021**.
53. Izquierdo, R.; Quintanar, A.; Parra, I.; Fernández-Llorca, D.; Sotelo, M. The prevention dataset: a novel benchmark for prediction of vehicles intentions. 2019 IEEE Intelligent Transportation Systems Conference (ITSC). IEEE, 2019, pp. 3114–3121.
54. Nowruzi, F.E.; Kolhatkar, D.; Kapoor, P.; Al Hassanat, F.; Heravi, E.J.; Laganiere, R.; Rebut, J.; Malik, W. Deep open space segmentation using automotive radar. 2020 IEEE MTT-S International Conference on Microwaves for Intelligent Mobility (ICMIM). IEEE, 2020, pp. 1–4.
55. Kraus, F.; Scheiner, N.; Ritter, W.; Dietmayer, K. The Radar Ghost Dataset–An Evaluation of Ghost Objects in Automotive Radar Data. 2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). IEEE, 2021, pp. 8570–8577.
56. Sakaridis, C.; Dai, D.; Van Gool, L. ACDC: The adverse conditions dataset with correspondences for semantic driving scene understanding. Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 10765–10775.
57. Kenk, M.A.; Hassaballah, M. DAWN: vehicle detection in adverse weather nature dataset. *arXiv preprint arXiv:2008.05402* **2020**.
58. Jin, J.; Fatemi, A.; Lira, W.M.P.; Yu, F.; Leng, B.; Ma, R.; Mahdavi-Amiri, A.; Zhang, H. Raidar: A rich annotated image dataset of rainy street scenes. Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 2951–2961.
59. Pitropov, M.; Garcia, D.E.; Rebello, J.; Smart, M.; Wang, C.; Czarnecki, K.; Waslander, S. Canadian adverse driving conditions dataset. *The International Journal of Robotics Research* **2021**, *40*, 681–690.
60. Bijelic, M.; Gruber, T.; Mannan, F.; Kraus, F.; Ritter, W.; Dietmayer, K.; Heide, F. Seeing through fog without seeing fog: Deep multimodal sensor fusion in unseen adverse weather. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 11682–11692.
61. Sheeny, M.; De Pellegrin, E.; Mukherjee, S.; Ahrabian, A.; Wang, S.; Wallace, A. RADIATE: A Radar Dataset for Automotive Perception. *arXiv preprint arXiv:2010.09076* **2020**.
62. El Natour, G.; Aider, O.A.; Rouveure, R.; Berry, F.; Faure, P. Radar and vision sensors calibration for outdoor 3D reconstruction. 2015 IEEE International Conference on Robotics and Automation (ICRA). IEEE, 2015, pp. 2084–2089.
63. Peršić, J.; Marković, I.; Petrović, I. Extrinsic 6dof calibration of a radar–lidar–camera system enhanced by radar cross section estimates evaluation. *Robotics and Autonomous Systems* **2019**, *114*, 217–230.
64. Domhof, J.; Kooij, J.F.; Gavrila, D.M. An extrinsic calibration tool for radar, camera and lidar. 2019 International Conference on Robotics and Automation (ICRA). IEEE, 2019, pp. 8107–8113.
65. Geiger, A.; Moosmann, F.; Car, Ö.; Schuster, B. Automatic camera and range sensor calibration using a single shot. 2012 IEEE international conference on robotics and automation. IEEE, 2012, pp. 3936–3943.
66. Dhall, A.; Chelani, K.; Radhakrishnan, V.; Krishna, K.M. LiDAR-camera calibration using 3D-3D point correspondences. *arXiv preprint arXiv:1705.09785* **2017**.
67. Sugimoto, S.; Tateda, H.; Takahashi, H.; Okutomi, M. Obstacle detection using millimeter-wave radar and its visualization on image sequence. Proceedings of the 17th International Conference on Pattern Recognition, 2004. ICPR 2004. IEEE, 2004, Vol. 3, pp. 342–345.
68. Peršić, J.; Petrović, L.; Marković, I.; Petrović, I. Spatio-temporal multisensor calibration based on gaussian processes moving object tracking. *arXiv preprint arXiv:1904.04187* **2019**.
69. Peršić, J.; Petrović, L.; Marković, I.; Petrović, I. Online multi-sensor calibration based on moving object tracking. *Advanced Robotics* **2021**, *35*, 130–140.

70. Faizullin, M.; Kornilova, A.; Ferrer, G. Open-Source LiDAR Time Synchronization System by Mimicking GPS-clock. *arXiv preprint arXiv:2107.02625* **2021**.
71. Kaul, P.; De Martini, D.; Gadd, M.; Newman, P. Rss-net: Weakly-supervised multi-class semantic segmentation with FMCW radar. 2020 IEEE Intelligent Vehicles Symposium (IV). IEEE, 2020, pp. 431–436.
72. Shoemake, K. Animating rotation with quaternion curves. Proceedings of the 12th annual conference on Computer graphics and interactive techniques, 1985, pp. 245–254.
73. He, K.; Gkioxari, G.; Dollár, P.; Girshick, R. Mask r-cnn. Proceedings of the IEEE international conference on computer vision, 2017, pp. 2961–2969.
74. Chen, L.C.; Papandreou, G.; Schroff, F.; Adam, H. Rethinking atrous convolution for semantic image segmentation. *arXiv preprint arXiv:1706.05587* **2017**.
75. Grimm, C.; Fei, T.; Warsitz, E.; Farhoud, R.; Breddermann, T.; Haeb-Umbach, R. Warping of Radar Data into Camera Image for Cross-Modal Supervision in Automotive Applications. *arXiv preprint arXiv:2012.12809* **2020**.
76. Dimitrievski, M.; Shopovska, I.; Van Hamme, D.; Veelaert, P.; Philips, W. Automatic labeling of vulnerable road users in multi-sensor data. 2021 IEEE International Intelligent Transportation Systems Conference (ITSC). IEEE, 2021, pp. 2623–2630.
77. Settles, B. Active learning literature survey **2009**.
78. Shorten, C.; Khoshgoftaar, T.M. A survey on image data augmentation for deep learning. *Journal of Big Data* **2019**, *6*, 1–48.
79. Hahner, M.; Dai, D.; Liniger, A.; Van Gool, L. Quantifying data augmentation for lidar based 3d object detection. *arXiv preprint arXiv:2004.01643* **2020**.
80. Park, D.S.; Chan, W.; Zhang, Y.; Chiu, C.C.; Zoph, B.; Cubuk, E.D.; Le, Q.V. SpecAugment: A Simple Data Augmentation Method for Automatic Speech Recognition. *Proc. Interspeech 2019* **2019**, pp. 2613–2617.
81. Wang, Y.; Hwang, J.N.; Wang, G.; Liu, H.; Kim, K.J.; Hsu, H.M.; Cai, J.; Zhang, H.; Jiang, Z.; Gu, R. ROD2021 Challenge: A Summary for Radar Object Detection Challenge for Autonomous Driving Applications. Proceedings of the 2021 International Conference on Multimedia Retrieval, 2021, pp. 553–559.
82. Ju, B.; Yang, W.; Jia, J.; Ye, X.; Chen, Q.; Tan, X.; Sun, H.; Shi, Y.; Ding, E. DANet: Dimension Apart Network for Radar Object Detection. Proceedings of the 2021 International Conference on Multimedia Retrieval, 2021, pp. 533–539.
83. Sheeny, M.; Wallace, A.; Wang, S. Radio: parameterized generative radar data augmentation for small datasets. *Applied Sciences* **2020**, *10*, 3861.
84. Ding, J.; Chen, B.; Liu, H.; Huang, M. Convolutional neural network with data augmentation for SAR target recognition. *IEEE Geoscience and remote sensing letters* **2016**, *13*, 364–368.
85. Gao, X.; Xing, G.; Roy, S.; Liu, H. Ramp-cnn: A novel neural network for enhanced automotive radar object recognition. *IEEE Sensors Journal* **2020**, *21*, 5119–5132.
86. Wang, C.; Ma, C.; Zhu, M.; Yang, X. Pointaugmenting: Cross-modal augmentation for 3d object detection. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 11794–11803.
87. Zhang, W.; Wang, Z.; Change Loy, C. Multi-modality cut and paste for 3d object detection. *arXiv e-prints* **2020**, pp. arXiv–2012.
88. Chamseddine, M.; Rambach, J.; Stricker, D.; Wasenmuller, O. Ghost Target Detection in 3D Radar Data using Point Cloud based Deep Neural Network. 2020 25th International Conference on Pattern Recognition (ICPR). IEEE, 2021, pp. 10398–10403.
89. Berthold, P.; Michaelis, M.; Luettel, T.; Meissner, D.; Wuensche, H.J. Radar reflection characteristics of vehicles for contour and feature estimation. 2017 Sensor Data Fusion: Trends, Solutions, Applications (SDF). IEEE, 2017, pp. 1–6.
90. Long, Y.; Morris, D.; Liu, X.; Castro, M.; Chakravarty, P.; Narayanan, P. Full-Velocity Radar Returns by Radar-Camera Fusion. Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 16198–16207.
91. Cabon, Y.; Murray, N.; Humenberger, M. Virtual kitti 2. *arXiv preprint arXiv:2001.10773* **2020**.
92. Tremblay, J.; Prakash, A.; Acuna, D.; Brophy, M.; Jampani, V.; Anil, C.; To, T.; Cameracci, E.; Boochoon, S.; Birchfield, S. Training deep networks with synthetic data: Bridging the reality gap by domain randomization. Proceedings of the IEEE conference on computer vision and pattern recognition workshops, 2018, pp. 969–977.
93. Hurl, B.; Czarnecki, K.; Waslander, S. Precise synthetic image and lidar (presil) dataset for autonomous vehicle perception. 2019 IEEE Intelligent Vehicles Symposium (IV). IEEE, 2019, pp. 2522–2529.
94. Rong, G.; Shin, B.H.; Tabatabaee, H.; Lu, Q.; Lemke, S.; Možeiko, M.; Boise, E.; Uhm, G.; Gerow, M.; Mehta, S.; et al. Lgsvl simulator: A high fidelity simulator for autonomous driving. 2020 IEEE 23rd International conference on intelligent transportation systems (ITSC). IEEE, 2020, pp. 1–6.
95. Johnson-Roberson, M.; Barto, C.; Mehta, R.; Sridhar, S.N.; Rosaen, K.; Vasudevan, R. Driving in the matrix: Can virtual worlds replace human-generated annotations for real world tasks? *arXiv preprint arXiv:1610.01983* **2016**.
96. Wang, J.; Pun, A.; Tu, J.; Manivasagam, S.; Sadat, A.; Casas, S.; Ren, M.; Urtasun, R. Advsim: Generating safety-critical scenarios for self-driving vehicles. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 9909–9918.
97. Schüßler, C.; Hoffmann, M.; Bräunig, J.; Ullmann, I.; Ebelt, R.; Vossiek, M. A Realistic Radar Ray Tracing Simulator for Large MIMO-Arrays in Automotive Environments. *IEEE Journal of Microwaves* **2021**, *1*, 962–974.
98. Holder, M.; Rosenberger, P.; Winner, H.; D'hondt, T.; Makkapati, V.P.; Maier, M.; Schreiber, H.; Magosi, Z.; Slavik, Z.; Bringmann, O.; et al. Measurements revealing challenges in radar sensor modeling for virtual validation of autonomous driving. 2018 21st International Conference on Intelligent Transportation Systems (ITSC). IEEE, 2018, pp. 2616–2622.

99. Scheel, A.; Dietmayer, K. Tracking multiple vehicles using a variational radar model. *IEEE Transactions on Intelligent Transportation Systems* **2018**, *20*, 3721–3736.
100. Xia, Y.; Wang, P.; Berntorp, K.; Svensson, L.; Granström, K.; Mansour, H.; Boufounos, P.; Orlik, P.V. Learning-Based Extended Object Tracking Using Hierarchical Truncation Measurement Model With Automotive Radar. *IEEE Journal of Selected Topics in Signal Processing* **2021**, *15*, 1013–1029.
101. Ngo, A.; Bauer, M.P.; Resch, M. A Multi-Layered Approach for Measuring the Simulation-to-Reality Gap of Radar Perception for Autonomous Driving. *arXiv preprint arXiv:2106.08372* **2021**.
102. Wheeler, T.A.; Holder, M.; Winner, H.; Kochenderfer, M.J. Deep stochastic radar models. 2017 IEEE Intelligent Vehicles Symposium (IV). IEEE, 2017, pp. 47–53.
103. Wang, L.; Goldluecke, B.; Anklam, C. L2R GAN: LiDAR-to-radar translation. Proceedings of the Asian Conference on Computer Vision, 2020.
104. Lekic, V.; Babic, Z. Automotive radar and camera fusion using generative adversarial networks. *Computer Vision and Image Understanding* **2019**, *184*, 1–8.
105. Ditzel, C.; Dietmayer, K. GenRadar: Self-Supervised Probabilistic Camera Synthesis Based on Radar Frequencies. *IEEE Access* **2021**, *9*, 148994–149042.
106. Wang, Y.; Chao, W.L.; Garg, D.; Hariharan, B.; Campbell, M.; Weinberger, K.Q. Pseudo-lidar from visual depth estimation: Bridging the gap in 3d object detection for autonomous driving. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 8445–8453.
107. Weng, X.; Kitani, K. Monocular 3d object detection with pseudo-lidar point cloud. Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops, 2019, pp. 0–0.
108. Qian, R.; Garg, D.; Wang, Y.; You, Y.; Belongie, S.; Hariharan, B.; Campbell, M.; Weinberger, K.Q.; Chao, W.L. End-to-end pseudo-lidar for image-based 3d object detection. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 5881–5890.
109. Fu, H.; Gong, M.; Wang, C.; Batmanghelich, K.; Tao, D. Deep ordinal regression network for monocular depth estimation. Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 2002–2011.
110. Godard, C.; Mac Aodha, O.; Firman, M.; Brostow, G.J. Digging into self-supervised monocular depth estimation. Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 3828–3838.
111. Dijk, T.v.; Croon, G.d. How do neural networks see depth in single images? Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 2183–2191.
112. Gasperini, S.; Koch, P.; Dallabetta, V.; Navab, N.; Busam, B.; Tombari, F. R4Dyn: Exploring radar for self-supervised monocular depth estimation of dynamic scenes. 2021 International Conference on 3D Vision (3DV). IEEE, 2021, pp. 751–760.
113. Lin, J.T.; Dai, D.; Van Gool, L. Depth estimation from monocular images and sparse radar data. 2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). IEEE, 2020, pp. 10233–10240.
114. Lo, C.C.; Vandewalle, P. Depth Estimation From Monocular Images And Sparse Radar Using Deep Ordinal Regression Network. 2021 IEEE International Conference on Image Processing (ICIP). IEEE, 2021, pp. 3343–3347.
115. Long, Y.; Morris, D.; Liu, X.; Castro, M.; Chakravarty, P.; Narayanan, P. Radar-camera pixel depth association for depth completion. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 12507–12516.
116. Lee, W.Y.; Jovanov, L.; Philips, W. Semantic-guided radar-vision fusion for depth estimation and object detection. BMVC, the 32nd British Machine Vision Conference, Proceedings. BMVA Press, 2021, p. 13.
117. Niesen, U.; Unnikrishnan, J. Camera-Radar Fusion for 3-D Depth Reconstruction. 2020 IEEE Intelligent Vehicles Symposium (IV). IEEE, 2020, pp. 265–271.
118. Kramer, A.; Stahoviak, C.; Santamaria-Navarro, A.; Agha-Mohammadi, A.A.; Heckman, C. Radar-inertial ego-velocity estimation for visually degraded environments. 2020 IEEE International Conference on Robotics and Automation (ICRA). IEEE, 2020, pp. 5739–5746.
119. Cen, S.H.; Newman, P. Precise ego-motion estimation with millimeter-wave radar under diverse and challenging conditions. 2018 IEEE International Conference on Robotics and Automation (ICRA). IEEE, 2018, pp. 6045–6052.
120. Kellner, D.; Barjenbruch, M.; Dietmayer, K.; Klappstein, J.; Dickmann, J. Instantaneous lateral velocity estimation of a vehicle using Doppler radar. Proceedings of the 16th International Conference on Information Fusion. IEEE, 2013, pp. 877–884.
121. Schubert, R.; Richter, E.; Wanielik, G. Comparison and evaluation of advanced motion models for vehicle tracking. 2008 11th international conference on information fusion. IEEE, 2008, pp. 1–6.
122. Kellner, D.; Barjenbruch, M.; Klappstein, J.; Dickmann, J.; Dietmayer, K. Instantaneous full-motion estimation of arbitrary objects using dual Doppler radar. 2014 IEEE Intelligent Vehicles Symposium Proceedings. IEEE, 2014, pp. 324–329.
123. Schlichenmaier, J.; Yan, L.; Stolz, M.; Waldschmidt, C. Instantaneous actual motion estimation with a single high-resolution radar sensor. 2018 IEEE MTT-S International Conference on Microwaves for Intelligent Mobility (ICMIM). IEEE, 2018, pp. 1–4.
124. Ding, F.; Pan, Z.; Deng, Y.; Deng, J.; Lu, C.X. Self-Supervised Scene Flow Estimation with 4D Automotive Radar. *arXiv preprint arXiv:2203.01137* **2022**.
125. Sun, D.; Yang, X.; Liu, M.Y.; Kautz, J. Pwc-net: Cnns for optical flow using pyramid, warping, and cost volume. Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 8934–8943.

126. Kabsch, W. A solution for the best rotation to relate two sets of vectors. *Acta Crystallographica Section A: Crystal Physics, Diffraction, Theoretical and General Crystallography* **1976**, *32*, 922–923.
127. Cao, Z.; Fang, W.; Song, Y.; He, L.; Song, C.; Xu, Z. DNN-Based Peak Sequence Classification CFAR Detection Algorithm for High-Resolution FMCW Radar. *IEEE Transactions on Geoscience and Remote Sensing* **2021**.
128. Lin, C.H.; Lin, Y.C.; Bai, Y.; Chung, W.H.; Lee, T.S.; Huttunen, H. DL-CFAR: A Novel CFAR target detection method based on deep learning. 2019 IEEE 90th Vehicular Technology Conference (VTC2019-Fall). IEEE, 2019, pp. 1–6.
129. Scheiner, N.; Schumann, O.; Kraus, F.; Appenrodt, N.; Dickmann, J.; Sick, B. Off-the-shelf sensor vs. experimental radar-How much resolution is necessary in automotive radar classification? 2020 IEEE 23rd International Conference on Information Fusion (FUSION). IEEE, 2020, pp. 1–8.
130. Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; et al. Comparing different clustering algorithms on toy datasets. https://scikit-learn.org/0.15/auto_examples/cluster/plot_cluster_comparison.html#example-cluster-plot-cluster-comparison-py.
131. Kellner, D.; Klappstein, J.; Dietmayer, K. Grid-based DBSCAN for clustering extended objects in radar data. 2012 IEEE Intelligent Vehicles Symposium. IEEE, 2012, pp. 365–370.
132. Scheiner, N.; Appenrodt, N.; Dickmann, J.; Sick, B. A multi-stage clustering framework for automotive radar data. 2019 IEEE Intelligent Transportation Systems Conference (ITSC). IEEE, 2019, pp. 2060–2067.
133. Angelov, A.; Robertson, A.; Murray-Smith, R.; Fioranelli, F. Practical classification of different moving targets using automotive radar and deep neural networks. *IET Radar, Sonar & Navigation* **2018**, *12*, 1082–1089.
134. Gao, X.; Xing, G.; Roy, S.; Liu, H. Experiments with mmwave automotive radar test-bed. 2019 53rd Asilomar Conference on Signals, Systems, and Computers. IEEE, 2019, pp. 1–6.
135. Cai, X.; Giallorenzo, M.; Sarabandi, K. Machine Learning-Based Target Classification for MMW Radar in Autonomous Driving. *IEEE Transactions on Intelligent Vehicles* **2021**, *6*, 678–689.
136. Scheiner, N.; Appenrodt, N.; Dickmann, J.; Sick, B. Radar-based road user classification and novelty detection with recurrent neural network ensembles. 2019 IEEE Intelligent Vehicles Symposium (IV). IEEE, 2019, pp. 722–729.
137. Schumann, O.; Wöhler, C.; Hahn, M.; Dickmann, J. Comparison of random forest and long short-term memory network performances in classification tasks using radar. 2017 Sensor Data Fusion: Trends, Solutions, Applications (SDF). IEEE, 2017, pp. 1–6.
138. Scheiner, N.; Appenrodt, N.; Dickmann, J.; Sick, B. Radar-based feature design and multiclass classification for road user recognition. 2018 IEEE Intelligent Vehicles Symposium (IV). IEEE, 2018, pp. 779–786.
139. Graham, B.; van der Maaten, L. Submanifold sparse convolutional networks. *arXiv preprint arXiv:1706.01307* **2017**.
140. Dreher, M.; Erçelik, E.; Bänziger, T.; Knol, A. Radar-based 2D Car Detection Using Deep Neural Networks. 2020 IEEE 23rd International Conference on Intelligent Transportation Systems (ITSC). IEEE, 2020, pp. 1–8.
141. Redmon, J.; Farhadi, A. Yolov3: An incremental improvement. *arXiv preprint arXiv:1804.02767* **2018**.
142. Schumann, O.; Hahn, M.; Dickmann, J.; Wöhler, C. Semantic segmentation on radar point clouds. 2018 21st International Conference on Information Fusion (FUSION). IEEE, 2018, pp. 2179–2186.
143. Danzer, A.; Griebel, T.; Bach, M.; Dietmayer, K. 2d car detection in radar data with pointnets. 2019 IEEE Intelligent Transportation Systems Conference (ITSC). IEEE, 2019, pp. 61–66.
144. Scheiner, N.; Kraus, F.; Appenrodt, N.; Dickmann, J.; Sick, B. Object detection for automotive radar point clouds–a comparison. *AI Perspectives* **2021**, *3*, 1–23.
145. Qi, C.R.; Su, H.; Mo, K.; Guibas, L.J. Pointnet: Deep learning on point sets for 3d classification and segmentation. Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 652–660.
146. Qi, C.R.; Yi, L.; Su, H.; Guibas, L.J. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. *Advances in neural information processing systems* **2017**, *30*.
147. Liu, J.; Xiong, W.; Bai, L.; Xia, Y.; Huang, T.; Ouyang, W.; et al. Deep Instance Segmentation with Automotive Radar Detection Points **2022**.
148. Liu, H.; Dai, Z.; So, D.; Le, Q. Pay attention to MLPs. *Advances in Neural Information Processing Systems* **2021**, *34*.
149. Schumann, O.; Lombacher, J.; Hahn, M.; Wöhler, C.; Dickmann, J. Scene understanding with automotive radar. *IEEE Transactions on Intelligent Vehicles* **2019**, *5*, 188–203.
150. Long, J.; Shelhamer, E.; Darrell, T. Fully convolutional networks for semantic segmentation. Proceedings of the IEEE conference on computer vision and pattern recognition, 2015, pp. 3431–3440.
151. Lang, A.H.; Vora, S.; Caesar, H.; Zhou, L.; Yang, J.; Beijbom, O. Pointpillars: Fast encoders for object detection from point clouds. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 12697–12705.
152. Xu, B.; Zhang, X.; Wang, L.; Hu, X.; Li, Z.; Pan, S.; Li, J.; Deng, Y. RPFA-Net: a 4D RaDAR Pillar Feature Attention Network for 3D Object Detection. 2021 IEEE International Intelligent Transportation Systems Conference (ITSC). IEEE, 2021, pp. 3061–3066.
153. Bai, J.; Zheng, L.; Li, S.; Tan, B.; Chen, S.; Huang, L. Radar transformer: An object classification network based on 4d mmw imaging radar. *Sensors* **2021**, *21*, 3854.
154. Zhao, H.; Jia, J.; Koltun, V. Exploring self-attention for image recognition. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 10076–10085.

155. Cheng, Y.; Su, J.; Chen, H.; Liu, Y. A New Automotive Radar 4D Point Clouds Detector by Using Deep Learning. ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2021, pp. 8398–8402.
156. Gall, M.; Gardill, M.; Horn, T.; Fuchs, J. Spectrum-based single-snapshot super-resolution direction-of-arrival estimation using deep learning. 2020 German Microwave Conference (GeMiC). IEEE, 2020, pp. 184–187.
157. Fuchs, J.; Gardill, M.; Lübke, M.; Dubey, A.; Lurz, F. A Machine Learning Perspective on Automotive Radar Direction of Arrival Estimation. *IEEE Access* **2022**.
158. Brodeski, D.; Bilik, I.; Giryes, R. Deep radar detector. 2019 IEEE Radar Conference (RadarConf). IEEE, 2019, pp. 1–6.
159. Zhang, G.; Li, H.; Wenger, F. Object detection and 3d estimation via an FMCW radar using a fully convolutional network. ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2020, pp. 4487–4491.
160. Mittal, S.; et al. A survey of accelerator architectures for 3D convolution neural networks. *Journal of Systems Architecture* **2021**, *115*, 102041.
161. Palffy, A.; Dong, J.; Kooij, J.F.; Gavrila, D.M. CNN based road user detection using the 3D radar cube. *IEEE Robotics and Automation Letters* **2020**, *5*, 1263–1270.
162. Major, B.; Fontijne, D.; Ansari, A.; Teja Sukhavasi, R.; Gowaikar, R.; Hamilton, M.; Lee, S.; Grzechnik, S.; Subramanian, S. Vehicle detection with automotive radar using deep learning on range-azimuth-doppler tensors. Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops, 2019, pp. 0–0.
163. Ouaknine, A.; Newson, A.; Pérez, P.; Tupin, F.; Rebut, J. Multi-View Radar Semantic Segmentation. Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 15671–15680.
164. Nowruzi, F.E.; Kolhatkar, D.; Kapoor, P.; Heravi, E.J.; Hassanat, F.A.; Laganiere, R.; Rebut, J.; Malik, W. PolarNet: Accelerated Deep Open Space Segmentation Using Automotive Radar in Polar Domain. *arXiv preprint arXiv:2103.03387* **2021**.
165. Hayashi, E.; Lien, J.; Gillian, N.; Giusti, L.; Weber, D.; Yamanaka, J.; Bedal, L.; Poupyrev, I. Radarnet: Efficient gesture recognition technique utilizing a miniature radar sensor. Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems, 2021, pp. 1–14.
166. Meyer, M.; Kuschk, G.; Tomforde, S. Graph convolutional networks for 3d object detection on radar data. Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 3060–3069.
167. Kipf, T.N.; Welling, M. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907* **2016**.
168. Dai, J.; Qi, H.; Xiong, Y.; Li, Y.; Zhang, G.; Hu, H.; Wei, Y. Deformable convolutional networks. Proceedings of the IEEE international conference on computer vision, 2017, pp. 764–773.
169. Li, P.; Wang, P.; Berntorp, K.; Liu, H. Exploiting Temporal Relations on Radar Perception for Autonomous Driving. *arXiv preprint arXiv:2204.01184* **2022**.
170. Nobis, F.; Geisslinger, M.; Weber, M.; Betz, J.; Lienkamp, M. A deep learning-based radar and camera sensor fusion architecture for object detection. 2019 Sensor Data Fusion: Trends, Solutions, Applications (SDF). IEEE, 2019, pp. 1–7.
171. Chadwick, S.; Maddern, W.; Newman, P. Distant vehicle detection using radar and vision. 2019 International Conference on Robotics and Automation (ICRA). IEEE, 2019, pp. 8311–8317.
172. Yadav, R.; Vierling, A.; Berns, K. Radar+ RGB Fusion For Robust Object Detection In Autonomous Vehicle. 2020 IEEE International Conference on Image Processing (ICIP). IEEE, 2020, pp. 1986–1990.
173. Girshick, R. Fast r-cnn. Proceedings of the IEEE international conference on computer vision, 2015, pp. 1440–1448.
174. Nabati, R.; Qi, H. Rrpn: Radar region proposal network for object detection in autonomous vehicles. 2019 IEEE International Conference on Image Processing (ICIP). IEEE, 2019, pp. 3093–3097.
175. Nabati, R.; Qi, H. Radar-camera sensor fusion for joint object detection and distance estimation in autonomous vehicles. *arXiv preprint arXiv:2009.08428* **2020**.
176. Nabati, R.; Qi, H. Centerfusion: Center-based radar and camera fusion for 3d object detection. Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2021, pp. 1527–1536.
177. Kim, Y.; Choi, J.W.; Kum, D. GRIF Net: Gated Region of Interest Fusion Network for Robust 3D Object Detection from Radar Point Cloud and Monocular Image. 2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). IEEE, 2020, pp. 10857–10864.
178. Lim, T.Y.; Ansari, A.; Major, B.; Fontijne, D.; Hamilton, M.; Gowaikar, R.; Subramanian, S. Radar and camera early fusion for vehicle detection in advanced driver assistance systems. Machine Learning for Autonomous Driving Workshop at the 33rd Conference on Neural Information Processing Systems, 2019, Vol. 2.
179. Zhang, J.; Zhang, M.; Fang, Z.; Wang, Y.; Zhao, X.; Pu, S. RVDet: Feature-level Fusion of Radar and Camera for Object Detection. 2021 IEEE International Intelligent Transportation Systems Conference (ITSC). IEEE, 2021, pp. 2822–2828.
180. Kim, J.; Kim, Y.; Kum, D. Low-level Sensor Fusion Network for 3D Vehicle Detection using Radar Range-Azimuth Heatmap and Monocular Image. Proceedings of the Asian Conference on Computer Vision, 2020.
181. Meyer, M.; Kuschk, G. Deep learning based 3d object detection for automotive radar and camera. 2019 16th European Radar Conference (EuRAD). IEEE, 2019, pp. 133–136.
182. Ku, J.; Mozifian, M.; Lee, J.; Harakeh, A.; Waslander, S.L. Joint 3d proposal generation and object detection from view aggregation. 2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). IEEE, 2018, pp. 1–8.
183. Yang, B.; Guo, R.; Liang, M.; Casas, S.; Urtasun, R. Radarnet: Exploiting radar for robust perception of dynamic objects. European Conference on Computer Vision. Springer, 2020, pp. 496–512.

184. Shah, M.; Huang, Z.; Laddha, A.; Langford, M.; Barber, B.; Zhang, S.; Vallespi-Gonzalez, C.; Urtasun, R. Liranet: End-to-end trajectory prediction using spatio-temporal radar fusion. *arXiv preprint arXiv:2010.00731* **2020**.
185. Liu, Y.; Fan, Q.; Zhang, S.; Dong, H.; Funkhouser, T.; Yi, L. Contrastive multimodal fusion with tupleinfonce. Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 754–763.
186. Cheng, Y.; Xu, H.; Liu, Y. Robust Small Object Detection on the Water Surface Through Fusion of Camera and Millimeter Wave Radar. Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 15263–15272.
187. Hu, J.; Shen, L.; Sun, G. Squeeze-and-excitation networks. Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 7132–7141.
188. Matzka, S.; Altendorfer, R. A comparison of track-to-track fusion algorithms for automotive sensor fusion. In *Multisensor Fusion and Integration for Intelligent Systems*; Springer, 2009; pp. 69–81.
189. Dong, X.; Zhuang, B.; Mao, Y.; Liu, L. Radar Camera Fusion via Representation Learning in Autonomous Driving. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 1672–1681.
190. Harakeh, A.; Smart, M.; Waslander, S.L. Bayesod: A bayesian approach for uncertainty estimation in deep object detectors. 2020 IEEE International Conference on Robotics and Automation (ICRA). IEEE, 2020, pp. 87–93.
191. Hüllermeier, E.; Waegeman, W. Aleatoric and epistemic uncertainty in machine learning: An introduction to concepts and methods. *Machine Learning* **2021**, *110*, 457–506.
192. Chavez-Garcia, R.O.; Aycard, O. Multiple sensor fusion and classification for moving object detection and tracking. *IEEE Transactions on Intelligent Transportation Systems* **2015**, *17*, 525–534.
193. Florea, M.C.; Jousselme, A.L.; Bossé, É.; Grenier, D. Robust combination rules for evidence theory. *Information Fusion* **2009**, *10*, 183–197.
194. Angelopoulos, A.N.; Bates, S. A gentle introduction to conformal prediction and distribution-free uncertainty quantification. *arXiv preprint arXiv:2107.07511* **2021**.
195. Kopp, J.; Kellner, D.; Piroli, A.; Dietmayer, K. Fast Rule-Based Clutter Detection in Automotive Radar Data. 2021 IEEE International Intelligent Transportation Systems Conference (ITSC). IEEE, 2021, pp. 3010–3017.
196. Kraus, F.; Scheiner, N.; Ritter, W.; Dietmayer, K. Using machine learning to detect ghost images in automotive radar. 2020 IEEE 23rd International Conference on Intelligent Transportation Systems (ITSC). IEEE, 2020, pp. 1–7.
197. Kamann, A.; Held, P.; Perras, F.; Zaumseil, P.; Brandmeier, T.; Schwarz, U.T. Automotive radar multipath propagation in uncertain environments. 2018 21st International Conference on Intelligent Transportation Systems (ITSC). IEEE, 2018, pp. 859–864.
198. Griebel, T.; Authaler, D.; Horn, M.; Henning, M.; Buchholz, M.; Dietmayer, K. Anomaly Detection in Radar Data Using PointNets. 2021 IEEE International Intelligent Transportation Systems Conference (ITSC). IEEE, 2021, pp. 2667–2673.
199. Garcia, J.M.; Prophet, R.; Michel, J.C.F.; Ebelt, R.; Vossiek, M.; Weber, I. Identification of ghost moving detections in automotive scenarios with deep learning. 2019 IEEE MTT-S International Conference on Microwaves for Intelligent Mobility (ICMIM). IEEE, 2019, pp. 1–4.
200. Wang, L.; Giebenhain, S.; Anklam, C.; Goldluecke, B. Radar ghost target detection via multimodal transformers. *IEEE Robotics and Automation Letters* **2021**, *6*, 7758–7765.
201. Guo, C.; Pleiss, G.; Sun, Y.; Weinberger, K.Q. On calibration of modern neural networks. International Conference on Machine Learning. PMLR, 2017, pp. 1321–1330.
202. Patel, K.; Beluch, W.; Rambach, K.; Cozma, A.E.; Pfeiffer, M.; Yang, B. Investigation of Uncertainty of Deep Learning-based Object Classification on Radar Spectra. 2021 IEEE Radar Conference (RadarConf21). IEEE, 2021, pp. 1–6.
203. Geng, C.; Huang, S.j.; Chen, S. Recent advances in open set recognition: A survey. *IEEE transactions on pattern analysis and machine intelligence* **2020**, *43*, 3614–3631.
204. Hall, D.; Dayoub, F.; Skinner, J.; Zhang, H.; Miller, D.; Corke, P.; Carneiro, G.; Angelova, A.; Sünderhauf, N. Probabilistic object detection: Definition and evaluation. Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2020, pp. 1031–1040.
205. Patel, K.; Beluch, W.; Rambach, K.; Pfeiffer, M.; Yang, B. Improving Uncertainty of Deep Learning-based Object Classification on Radar Spectra using Label Smoothing. *arXiv preprint arXiv:2109.12851* **2021**.
206. Wenger, J.; Kjellström, H.; Triebel, R. Non-parametric calibration for classification. International Conference on Artificial Intelligence and Statistics. PMLR, 2020, pp. 178–190.
207. Patel, K.; Beluch, W.H.; Yang, B.; Pfeiffer, M.; Zhang, D. Multi-Class Uncertainty Calibration via Mutual Information Maximization-based Binning. International Conference on Learning Representations, 2020.
208. Müller, R.; Kornblith, S.; Hinton, G.E. When does label smoothing help? *Advances in neural information processing systems* **2019**, *32*.
209. Thulasidasan, S.; Chennupati, G.; Bilmes, J.A.; Bhattacharya, T.; Michalak, S. On mixup training: Improved calibration and predictive uncertainty for deep neural networks. *Advances in Neural Information Processing Systems* **2019**, *32*.
210. Hendrycks, D.; Mu, N.; Cubuk, E.D.; Zoph, B.; Gilmer, J.; Lakshminarayanan, B. AugMix: A Simple Data Processing Method to Improve Robustness and Uncertainty. International Conference on Learning Representations, 2019.
211. Gal, Y.; Ghahramani, Z. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. international conference on machine learning. PMLR, 2016, pp. 1050–1059.
212. Fort, S.; Hu, H.; Lakshminarayanan, B. Deep ensembles: A loss landscape perspective. *arXiv preprint arXiv:1912.02757* **2019**.

213. Feng, D.; Wang, Z.; Zhou, Y.; Rosenbaum, L.; Timm, F.; Dietmayer, K.; Tomizuka, M.; Zhan, W. Labels are not perfect: Inferring spatial uncertainty in object detection. *IEEE Transactions on Intelligent Transportation Systems* **2021**.
214. Kendall, A.; Gal, Y. What uncertainties do we need in bayesian deep learning for computer vision? *Advances in neural information processing systems* **2017**, *30*.
215. Dong, X.; Wang, P.; Zhang, P.; Liu, L. Probabilistic oriented object detection in automotive radar. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, 2020, pp. 102–103.
216. Mohammed, A.S.; Amamou, A.; Ayevide, F.K.; Kelouwani, S.; Agbossou, K.; Zioui, N. The perception system of intelligent ground vehicles in all weather conditions: a systematic literature review. *Sensors* **2020**, *20*, 6532.
217. Hendrycks, D.; Dietterich, T. Benchmarking Neural Network Robustness to Common Corruptions and Perturbations. International Conference on Learning Representations, 2019.
218. Secci, F.; Ceccarelli, A. On failures of RGB cameras and their effects in autonomous driving applications. 2020 IEEE 31st International Symposium on Software Reliability Engineering (ISSRE). IEEE, 2020, pp. 13–24.
219. Jokela, M.; Kutila, M.; Pyykönen, P. Testing and validation of automotive point-cloud sensors in adverse weather conditions. *Applied Sciences* **2019**, *9*, 2341.
220. Carballo, A.; Lambert, J.; Monrroy, A.; Wong, D.; Narksri, P.; Kitsukawa, Y.; Takeuchi, E.; Kato, S.; Takeda, K. LIBRE: The multiple 3D LiDAR dataset. 2020 IEEE Intelligent Vehicles Symposium (IV). IEEE, 2020, pp. 1094–1101.
221. Zang, S.; Ding, M.; Smith, D.; Tyler, P.; Rakotoarivelo, T.; Kaafar, M.A. The impact of adverse weather conditions on autonomous vehicles: how rain, snow, fog, and hail affect the performance of a self-driving car. *IEEE vehicular technology magazine* **2019**, *14*, 103–111.
222. Brooker, G.; Hennessey, R.; Lobsey, C.; Bishop, M.; Widzyk-Capehart, E. Seeing through dust and water vapor: Millimeter wave radar sensors for mining applications. *Journal of Field Robotics* **2007**, *24*, 527–557.
223. Guan, J.; Madani, S.; Jog, S.; Gupta, S.; Hassanieh, H. Through fog high-resolution imaging using millimeter wave radar. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 11464–11473.
224. Gourova, R.; Krasnov, O.; Yarovoy, A. Analysis of rain clutter detections in commercial 77 GHz automotive radar. 2017 European Radar Conference (EURAD). IEEE, 2017, pp. 25–28.
225. Breitenstein, J.; Termöhlen, J.A.; Lipinski, D.; Fingscheidt, T. Corner Cases for Visual Perception in Automated Driving: Some Guidance on Detection Approaches. *arXiv preprint arXiv:2102.05897* **2021**.
226. Koopman, P.; Fratrik, F. How many operational design domains, objects, and events? Safeai@ aaai, 2019.
227. Antonante, P.; Spivak, D.I.; Carlone, L. Monitoring and diagnosability of perception systems. 2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). IEEE, 2020, pp. 168–175.
228. Zheng, Z.; Yue, X.; Keutzer, K.; Sangiovanni Vincentelli, A. Scene-aware Learning Network for Radar Object Detection. Proceedings of the 2021 International Conference on Multimedia Retrieval, 2021, pp. 573–579.
229. Malawade, A.V.; Mortlock, T.; Faruque, M.A.A. HydraFusion: Context-Aware Selective Sensor Fusion for Robust and Efficient Autonomous Vehicle Perception. *arXiv preprint arXiv:2201.06644* **2022**.
230. Ahuja, N.; Alvarez, I.J.; Krishnan, R.; Ndiour, I.J.; Subedar, M.; Tickoo, O. Robust multimodal sensor fusion for autonomous driving vehicles, 2020. US Patent App. 16/911,100.
231. Deng, J.; Dong, W.; Socher, R.; Li, L.J.; Li, K.; Fei-Fei, L. Imagenet: A large-scale hierarchical image database. 2009 IEEE conference on computer vision and pattern recognition. Ieee, 2009, pp. 248–255.
232. Feng, D.; Harakeh, A.; Waslander, S.L.; Dietmayer, K. A review and comparative study on probabilistic object detection in autonomous driving. *IEEE Transactions on Intelligent Transportation Systems* **2021**.
233. Ettinger, S.; Cheng, S.; Caine, B.; Liu, C.; Zhao, H.; Pradhan, S.; Chai, Y.; Sapp, B.; Qi, C.R.; Zhou, Y.; et al. Large scale interactive motion forecasting for autonomous driving: The waymo open motion dataset. Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 9710–9719.
234. Lin, Y.C.; Gu, M.X.; Lin, C.H.; Lee, T.S. Deep-Learning Based Decentralized Frame-to-Frame Trajectory Prediction Over Binary Range-Angle Maps for Automotive Radars. *IEEE Transactions on Vehicular Technology* **2021**, *70*, 6385–6398.
235. Kunert, M. The EU project MOSARIM: A general overview of project objectives and conducted work. 2012 9th European Radar Conference. IEEE, 2012, pp. 1–5.
236. Alland, S.; Stark, W.; Ali, M.; Hegde, M. Interference in automotive radar systems: Characteristics, mitigation techniques, and current and future research. *IEEE Signal Processing Magazine* **2019**, *36*, 45–59.
237. Oyedare, T.; Shah, V.K.; Jakubisin, D.J.; Reed, J.H. Interference Suppression Using Deep Learning: Current Approaches and Open Challenges. *arXiv preprint arXiv:2112.08988* **2021**.