Multi-class Road User Detection with 3+1D Radar in the View-of-Delft Dataset

Palffy, Andras; Pool, Ewoud; Baratam, Srimannarayana; Kooij, Julian; Gavrila, Dariu

DOI
10.1109/LRA.2022.3147324

Publication date
2022

Document Version
Final published version

Published in
IEEE Robotics and Automation Letters

Multi-Class Road User Detection With 3+1D Radar in the View-of-Delft Dataset

Andras Palffy, Ewoud Pool, Srimannarayana Baratam, Julian F. P. Kooij, and Dariu M. Gavrila

***Abstract*—Next-generation automotive radars provide elevation data in addition to range-, azimuth- and Doppler velocity. In this experimental study, we apply a state-of-the-art object detector (PointPillars), previously used for LiDAR 3D data, to such 3+1D radar data (where 1D refers to Doppler). In ablation studies, we first explore the benefits of the additional elevation information, together with that of Doppler, radar cross section and temporal accumulation, in the context of multi-class road user detection. We subsequently compare object detection performance on the radar and LiDAR point clouds, object class-wise and as a function of distance. To facilitate our experimental study, we present the novel View-of-Delft (VoD) automotive dataset. It contains 8693 frames of synchronized and calibrated 64-layer LiDAR-, (stereo) camera-, and 3+1D radar-data acquired in complex, urban traffic. It consists of 123106 3D bounding box annotations of both moving and static objects, including 26587 pedestrian, 10800 cyclist and 26949 car labels. Our results show that object detection on 64-layer LiDAR data still outperforms that on 3+1D radar data, but the addition of elevation information and integration of successive radar scans helps close the gap. The VoD dataset is made freely available for scientific benchmarking at https://intelligent-vehicles.org/datasets/view-of-delft/.**

***Index Terms*—Object detection, segmentation and categorization; data sets for robotic vision; automotive radars.**

I. Introduction

RADARS are often used in intelligent vehicles because they are relatively robust to weather and lighting conditions, have excellent range sensitivity, and can directly measure objects' radial velocities at a reasonable cost. Traditional automotive radars (*2+1D radars*) output a sparse point cloud of reflections called *radar targets*. Each point has two spatial dimensions, range r and azimuth α, and a third dimension referred to as Doppler, which is the radial velocity v_{rel} of the target relative to the ego-vehicle [1]. In recent years, developments in both radar technology and proposed algorithms have made it possible to use these radars for road user detection [2]–[6]. Despite these improvements, the sparsity of point clouds provided by traditional automotive radars is still a bottleneck in object detection research. Due to their small number of points, it is challenging to regress accurate 2D bird's-eye view (BEV) bounding boxes, especially for smaller objects such as pedestrians. Furthermore, the lack of elevation information (i.e., the height of the points) makes it nearly impossible to infer the height and vertical offset of objects. Unlike LiDAR based detectors, most 2+1D radar based object detection methods do not regress bounding boxes either in 2D (BEV) or in 3D, but instead perform semantic or instance segmentation of the 2+1D radar point clouds [3], [5], [7]–[10]. Bounding box regression on sparse radar point clouds remains challenging since the objects usually only have a few points on them, providing little spatial information on the exact location and extent of the true bounding box. The latest improvement in automotive radar technology, *3+1D radars* may help to overcome these limitations. Unlike traditional automotive radars, 3+1D radars have three spatial dimensions: range, azimuth, and elevation, while still providing Doppler as a fourth dimension. They also tend to provide a denser point cloud [11]. With the additional elevation information and increased density, 3+1D radar point clouds are somewhat reminiscent of LiDAR point clouds. Therefore, these radars may be better suited for multi-class 3D bounding box regression, and it is intuitive to apply object detection networks developed for LiDAR data to them. Nonetheless, 3+1D radars have only been used for the single-class car detection task [12], [13], not for pedestrian, cyclist, or multi-class detection tasks. We see two possible reasons for this. First, the object detection networks regularly used for LiDAR input were not designed with the Doppler dimension in mind, and it is unclear how best to incorporate this additional information. Furthermore, the measured Doppler

Fig. 1. Example scene from the View-of-Delft (VoD) dataset. Our recordings contain camera images, LiDAR point clouds (shown here as lines of small dots), and 3+1D radar data (shown as large dots), along with accurate localization information and 3D bounding box annotations (cyclist/pedestrian class labels are colored red/green).

Manuscript received September 9, 2021; accepted January 10, 2022. Date of publication February 1, 2022; date of current version March 8, 2022. This letter was recommended for publication by Associate Editor A. Valada and Editor M. Vincze upon evaluation of the reviewers' comments. This work was supported by the Dutch Science Foundation NWO-TTW, within the SafeVRU project (nr. 14667). *(Corresponding author: Dariu M. Gavrila.)*

The authors are with the Intelligent Vehicles Group, TU Delft, 2628 CD Delft, Netherlands (e-mail: d.m.gavrila@tudelft.nl).

This letter has supplementary downloadable material available at https://doi.org/10.1109/LRA.2022.3147324, provided by the authors.

Digital Object Identifier 10.1109/LRA.2022.3147324

values depend on the direction in which the object is located, so many data augmentation techniques often applied to LiDAR point clouds are not suitable for radar ones. Second, while many datasets contain several thousand 3D bounding box annotations for multiple classes on LiDAR data [14]–[16], the only publicly available detection dataset [11] with 3+1D radar data has only ∼500 frames, with fewer than 40 annotations for pedestrians or cyclists, and thus, it is not suitable for multi-class object detection.

In this experimental study, we apply a state-of-the-art object detector (PointPillars [17]), commonly used for LiDAR 3D data, to such 3+1D radar data. We incorporate the Doppler information, and explore how it influences the detection performance. Furthermore, we investigate how elevation information and the use of past radar scans (i.e. temporal information) increase road user detection performance. We also discuss what kind of data augmentation methods are applicable to 3+1D radar data. Finally, we compare our best radar based object detection method with a PointPillars network operating with LiDAR data, and examine the two sensors' performance and capabilities as a function of class and distance.

To facilitate our experimental study, we introduce the View-of-Delft[1] (VoD) dataset, a multi-sensor automotive dataset for multi-class 3D object detection, see Fig. 1.

II. Related Work

A. 2+1D Radar Based Multi-Class Object Detection

Traditional automotive radars have been used for multi-class road user detection in various ways, such as using clustering algorithms [2], [7], convolutional neural networks (CNNs) [3], [4], [22], or point cloud processing neural networks [5], [6]. The sparsity of the point cloud provided by 2+1D radars is one of the largest bottlenecks of the radar perception domain. Furthermore, the lack of elevation information renders the inference of objects' height nearly impossible. Researchers attempted to overcome these challenges and obtain more information in various ways, e.g.: by merging multiple frames over time [5], [22], [23], using multiple radars [24], using low-level radar data [3], [4], [23], or fusing radar with other sensor modalities [25]–[28]. Nevertheless, there is no 2+1D radar based method that performs multi-class 3D bounding box regression. Instead, most existing methods perform semantic or instance segmentation of the radar point cloud, i.e. they assign a class label (and potentially an object id) to each radar target individually [3], [5], [7]–[10].

B. 3+1D Based Multi-Class Object Detection

Only few works have used 3+1D radars for object detection. In [29] the authors applied such a sensor to build a static 3D occupancy map of highway and parking lot scenes after filtering out dynamic targets. Afterward, the map is semantically segmented by image segmentation networks into the street, curbstone, fence, barrier, or parked car classes. Currently, the only publicly available automotive detection dataset that contains 3+1D radar data is the Astyx dataset [11]. Despite the small size of the dataset (∼500 frames), the authors have successfully used it to perform 3D car detection by fusing radar and camera with the AVOD fusion network [12]. Furthermore, they also compared this radar-camera fusion with LiDAR-camera fusion, although the LiDAR sensor had only 16 layers. Finally, [13] used the combination of two spatially separated low-resolution 3+1D radars to detect vehicles by a novel neural network called RP-net, containing several Pointnet layers. To the best of our knowledge, 3+1D radars have neither been used for multi-class road user detection before, nor have they been compared to high-end LiDAR sensors.

C. The Use of Doppler

Doppler has been exploited in various ways before. Its most trivial use is to distinguish static and dynamic objects after ego-motion compensation. E.g., while some research only keeps static radar targets [29]–[31], others use the Doppler information to keep only moving reflections to detect dynamic objects [3], [23], [32]. After first clustering the radar point cloud to generate object proposals, basic statistical properties (mean, deviation, etc.) of the velocity spectrum can been used for classification [2], [7]. [5] presented in an ablation study that adding Doppler as an input channel to a Pointnet++ network significantly improves semantic segmentation. [3] showed that the (relative) velocity distribution contains valuable class information which can be exploited for multi-class road user detection. With multiple radar targets originating from the same object, it is also possible to regress the 2D velocity vector (and thus, orientation) of the object using the targets' measured radial velocities as samples at different azimuth angles, as [33] showed for cars and [34] for bikes. Thus, it has been shown that the Doppler dimension can be beneficial in 3D object detection in two ways: 1) classification, as classes may have distinct velocity patterns [3], [5], and 2) in orientation estimation, as the general velocity (moving direction) of an object is highly correlated with its orientation [33], [34]. Despite its advantages, in the few works that used a 3+1D radar sensor, Doppler was either ignored [12], used to filter static radar targets [29], or used as an additional input channel in a point cloud processing network without ego-motion compensation [13]. Although Doppler has been shown to be beneficial for multi-class road user detection using traditional 2+1D automotive radars, 3+1D radars have only been used for single-class vehicle detection in the literature [13].

D. Radar Datasets

Recently, several automotive datasets containing radar data were published for various tasks such as localization [35], [36], object classification [37], or scene understanding with a stationary radar sensor [38]. In this section, we focus on detection datasets that contain realistic recordings from a moving ego-vehicle. To be suitable for multi-class road user detection tasks with radar (either pure radar or sensor fusion), we argue that an automotive dataset should meet the following requirements: 1) use a next-generation 3+1D radar to provide both elevation and Doppler information, 2) equip high-end sensors from the other modalities as well, i.e., a high definition camera and a 64-layer LiDAR, 3) provide annotations for the objects that include their

[1]Named after the famous painting by Johannes Vermeer (pun intended)

TABLE I
COMPARISON OF PUBLICLY AVAILABLE RADAR DETECTION DATASETS WITH SENSORS USED, TYPE OF ANNOTATION, AND THE NUMBER OF VEHICLE (SUM OF *CAR*, *TRUCK*, AND *BUS*), *PEDESTRIAN*, AND *CYCLIST* ANNOTATIONS (INDIVIDUAL ANNOTATIONS/UNIQUE INSTANCES, WHERE UNIQUE OBJECT ID IS AVAILABLE)

Name	Radar data	Camera	LiDAR	Size	Vehicles	Pedestrians	Cyclists	Annotation
RadarScenes (2021) [18]	4×2+1D, front/side	mono	✗	832k frames	326636/3889	128197/1529	61051/268	point-wise
CRUW (2021) [19]	2×2+1D, front/side	stereo	✗	396k frames	23330/-	31980/-	13347/-	2D position
RADIATE (2020) [20]	1×2D, surround	stereo	32 layers	3 hours	185810/-	10970/-	499/-	2D bboxes
Zendar (2020) [21]	1×2+1D, front	mono	16 layers	4780 frames	11300/-	0/-	0/-	2D bboxes
nuScenes (2019) [15]	5×2+1D, surround	6×mono	32 layers	400k frames	598849/-	217913/-	7331/-	3D bboxes
Astyx (2019) [11]	1×3+1D, front	mono	16 layers	546 frames	3087/-	39/-	11/-	3D bboxes
View-of-Delft (2021)	**1×3+1D, front**	stereo	**64 layers**	8693 frames	27273/429	26587/380	10800/183	**3D bboxes**

Top/bottom sections are datasets with radars providing 2D/3D spatial coordinates.

extent and orientation (2D or 3D bounding boxes), and 4) should have reasonable number of annotations for the most important urban road users: pedestrians, cars, and cyclists.

Table I gives an overview of the currently available radar detection datasets according to these requirements. It can be seen that both the RadarScenes [18] and CRUW [19] datasets contain 2+1D radar and camera data, and have large number of annotations for all three main classes. Unfortunately, they do not provide LiDAR data or bounding box annotations. Furthermore, in RadarScenes, only the moving objects are annotated. The RADIATE dataset [20] contains radar, camera, and LiDAR data along with 2D BEV bounding box annotations for all three classes. It was collected using a mechanically rotating 2D radar, which provides a 360° dense image of the environment, but does not output Doppler or elevation information. The Zendar dataset [21] provides Synthetic Aperture Radar (SAR) data using a 2+1D radar. Unfortunately, it only has annotations for the car class. The nuScenes dataset [15] contains data from all three sensor modalities, and they provide a large number of 3D bounding box annotations. However, the output of the equipped 2+1D radar sensors is considered too sparse for radar-only detection methods by some in the research community [1], [18], and the used LiDAR sensor has only 32 layers. Finally the Astyx dataset [11] is the only one to use a 3+1D radar, and it also contains data from a camera and a 16-layer LiDAR. Unfortunately, its limited size (∼500 frames) and highly imbalanced classes (e.g., only 39/11 pedestrians/cyclists annotations) make it ill-suited for multi-class object detection research. In conclusion, no existing publicly available dataset satisfies all the requirements.

E. Contributions

Our main contributions are as follows:

1) We examine road user detection with 3+1D radar using PointPillars [17], a state-of-the-art multi-class 3D object detector commonly used for LiDAR. We investigate the importance of different features of the radar point cloud in an ablation study, including Doppler, *RCS*, and the elevation information that traditional 2+1D automotive radars cannot provide.
2) We compare radar based to LiDAR based detection by training and testing on the same traffic scenes. We show that currently point cloud based detection on dense LiDAR still outperforms detection on radar. However, we also find that the performance gap can be reduced when radar data includes elevation information, and when multiple radar scans are temporally integrated. Additionally, the detection benefits from Doppler measurements, which are unique to radar.
3) We publish the View-of-Delft (VoD) dataset, a novel multi-sensor automotive dataset for multi-class 3D object detection, consisting of calibrated and synchronized LiDAR, camera, and radar data recorded in real-world traffic situations with annotations for both static and moving road users. The View-of-Delft dataset is the largest dataset containing 3+1D radar recordings with ∼20 times as many annotated frames as the Astyx dataset [11], and it is the only public dataset containing camera, (any kind of) radar, and 64-layer LiDAR data at the same time. Although this work focuses on radar-only methods, the dataset is also suitable for sensor fusion, camera-only, or LiDAR-only methods due to this sensor arrangement, and could be useful for researchers interested in cluttered urban traffic.

III. DATASET

In this section, we present the View-of-Delft dataset, including the sensor setup used and the annotations provided.[2] The dataset was recorded while driving with our demonstrator vehicle [39] through *campus*, *suburb* and *old-town* locations in the city of Delft (The Netherlands). Recordings were selected with a preference for scenarios containing vulnerable road users (VRU-s), i.e., pedestrians and cyclists.

A. Measurement Setup and Provided Data

We recorded the output of the following sensors: a ZF FR-Gen21 3+1D radar (see Table II for specifications, ∼13 Hz) mounted behind the front bumper, a stereo camera (1936 × 1216 px, ∼30 Hz) mounted on the windshield, a Velodyne HDL-64 S3 LIDAR (∼10 Hz) scanner on the roof, and the ego vehicle's odometry (filtered combination of RTK GPS, IMU, and wheel odometry, ∼100 Hz). All sensors were jointly calibrated following [40]. See Fig. 2 for a general overview of the sensor setup.

[2]The VoD dataset, including its annotations for the training and validation sets, will be made freely available at https://intelligent-vehicles.org/datasets/view-of-delft/ to academic and non-profit organizations for non-commercial, scientific use. The test set annotations will be withheld.

TABLE II
NATIVE ACCURACY AND RESOLUTION ALONG THE FOUR DIMENSIONS OF OUR RADAR SENSOR CONFIGURATION

	range	velocity	azimuth	elevation
Accuracy	≤0.02 m	0.01 m/s	0.15°	0.3°
Resolution	≤0.2 m	0.1 m/s	1.5°	1.5°

On-board signal processing provides further resolution gains.

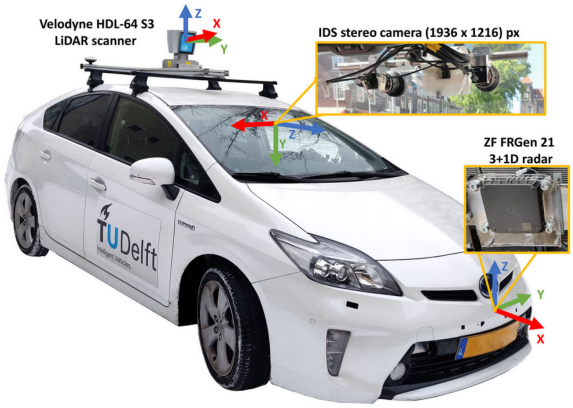


Fig. 2. Recording platform. Our Toyota Prius 2013 platform is equipped with a stereo camera setup, a rotating 3D LiDAR sensor, a ZF FRGen 21 3+1D radar, and a combined GPS/IMU inertial navigation system.

We provide the dataset in synchronized "frames" similar to [14], consisting of a LiDAR point cloud, a rectified mono-camera image, a radar point cloud, and a transformation describing the odometry. Timestamps of the LiDAR sensor were chosen as lead, and we chose the closest camera, radar and odometry information available (maximum tolerated time difference is set to 0.05 seconds). The frames are sequential in time with 10 Hz (after synchronization) and they are organized into clips with an average length of ∼40 seconds. The LiDAR and radar point clouds are ego-motion compensated, both for ego-motion between the capture of LiDAR/radar and camera data, and for ego-motion during the scan (i.e., one full rotation of the LiDAR sensor). Our dataset follows the popular KITTI dataset [14] both in the defined coordinate systems (see Fig. 2) and in the file structure. The main advantage of this choice is that several open-source toolkits and detection methods are directly applicable to our dataset. In addition to this synchronized version of our dataset, we also make the "raw" asynchronous recorded data available, including all radar scans at 13 Hz, and rectified camera images at 30 Hz from both the left and right cameras. This can benefit researchers seeking richer temporal data for detection, tracking, prediction, or other tasks.

B. Annotation

Any object of interest (static or moving) within 50 meters of the LiDAR sensor and partially or fully within the camera's field of view (horizontal FoV: ±32°, vertical FoV: ± 22°) was annotated[3] with a six degree of freedom (6 DoF) 3D bounding box. 13 object classes were annotated, see Table III for their object count. For each object, we also annotated the level of occlusion for two types of occlusions ("spatial" and "lighting") and an activity attribute ("stopped," "moving," "parked," "pushed," "sitting"). Furthermore, same physical objects were assigned unique object ids over frames to make the dataset suitable for tracking and prediction tasks. Annotation instructions with detailed descriptions of the classes and attributes will be shared along with the dataset.

[3]Annotation was done by https://understand.ai, a subsidiary of DSpace.

IV. METHODOLOGY

This work uses PointPillars [17] as the baseline state-of-the-art multi-class object detector. While PointPillars is typically trained on LiDAR data, we instead train it on 3+1D radar point clouds. In this section we detail the available features of the radar input, and describe how to encode Doppler. We also discuss data augmentation techniques and describe temporal merging of multiple radar scans.

A. 3+1D Radar Point Clouds and Doppler Encoding

The 3+1D radar outputs a point cloud with spatial, Doppler and reflectivity channels for each scan, giving a total of five features for each point: r range, α azimuth, θ elevation, v_{rel} relative radial velocity, and *RCS* reflectivity. Since most point cloud based object detectors use Cartesian coordinates, we also transform the radar point cloud: $p = [x, y, z, v_{rel}, RCS]$, where p denotes a point, and x, y, z are the three spatial coordinates with x and y axes pointing forward and left respectively w.r.t. the vehicle, see Fig. 2. *Compensated radial velocity* is a signed scalar value denoted by v_r, describing the ego-motion compensated (i.e. absolute) radial velocity of the point. To obtain it, we perform ego-motion compensation for v_{rel} by eliminating the motion of the sensor that comes from both the translational and rotational movement of the ego-vehicle. Examples of such encoding of Doppler for multi-class object detection include [3] and [5]. v_r was used as additional decoration for the radar points and it was normalized feature-wise to have zero-mean and unit standard deviation.

B. Accumulation of Radar Point Clouds

We experiment with incorporating multiple radar scans in our object detector similar to what [15] has been done for LiDAR and [5] for 2+1D radar data. Aside from the advantage of richer point clouds, merging also provides temporal information, which may help object detectors not only in localization but in classification as well. Accumulation is implemented by transforming point clouds from previous scans to the coordinate system of the last scan and appending a scalar time id denoted by t to each point indicating which scan it originates from. E.g., a point from the current scan has a $t = 0$, while a point from the third most recent scan has a $t = -2$. The encoder includes this time id as an extra decoration for the radar points. Note that a "scan" is not the same as a "frame" defined in Section III. While radar point clouds in the frames are synchronized with the LiDAR sensor, here we merge

TABLE III
DATASET STATISTICS: NUMBER OF ANNOTATED OBJECTS (TOP), NUMBER OF UNIQUE OBJECTS (MIDDLE), AND PERCENTAGE OF MOVING OBJECTS (BOTTOM), PER CLASS. THE RATIOS COMPARED TO THE WHOLE DATASET ARE GIVEN IN BRACKETS

	car	**pedestrian**	**cyclist**	**rider**	**unused bicycle**	**bicycle rack**	**human depiction**	**moped or scooter**	**motor**	**other**	$\sum$
# objects	26949 (21.9%)	26587 (21.6%)	10800 (8.8%)	12809 (10.4%)	24933 (20.3%)	12025 (9.8%)	370 (0.3%)	5403 (4.4%)	629 (0.5%)	2601 (2.1%)	123106
# unique obj	423 (22.6%)	380 (20.3%)	183 (9.8%)	222 (11.8%)	372 (19.8%)	156 (8.3%)	10 (0.5%)	80 (4.3%)	13 (0.7%)	36 (1.9%)	1875
% moving	7.2%	73.2%	96.1%	95.5%	0.7%	0.0%	0.0%	10.7%	59.9%	34.5%	37.4%

(The "other" column combines the classes *ride_other*, *vehicle_other*, *truck*, and *ride_uncertain*).

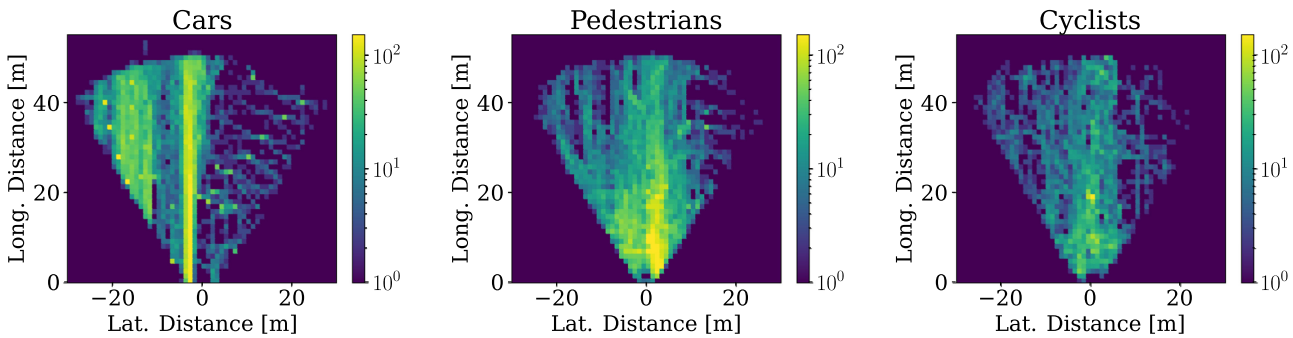


Fig. 3. Overall spatial distribution of cars, pedestrians, and cyclists in the dataset as a log plot. The ego-vehicle is positioned at (0, 0), looking upwards. Each pixel corresponds to one square meter area. Darkest blue means zero annotation.

the last scans received from the radar independently of other sensors.

C. Data Augmentation

Not every data augmentation method used in LiDAR research is directly applicable to radar point clouds since the v_r measured by the radar should remain correlated with the angle at which the object is observed. The same object with the same kinematics (speed and direction) would be detected with different velocity measurements at a different azimuth or elevation angle, i.e., after being translated during augmentation. Similarly, it is not possible to rotate the ground truth bounding boxes and the points within them locally (around their vertical axis), as this changes the radial component of the object velocity in an unknown way. Finally, rotating the radar point cloud around the sensor (e.g., around its vertical axis) does not affect the measured relative radial velocities. However, this is not true for the ego-motion compensated radial velocities, since the compensation uses the angles between the motion vector of the radar and the direction of the objects. Therefore, commonly used augmentation methods such as translation and rotation of the point cloud or rotation of the ground truth boxes can even be detrimental in the case of radar point clouds. However, mirroring the point cloud to the longitudinal axis and scaling are applicable, as the (absolute) observation angles of radar points do not change. Note that augmentation by scaling is only valid if the origin is the radar sensor itself.

V. EXPERIMENTS

We consider object detection performance on three object classes: *car*, *pedestrian* and *cyclist*. The spatial distributions of these classes are shown in Fig. 3. Unlike [3][5][18][23], we considered both static and moving objects in our experiments. We split the dataset into a training, validation, and testing set in a ratio of 59%/15%/26% such that frames from the same clip will only be present in one split. The clips are assigned to splits such that the number of annotations (both static and moving) of the three main classes (cars, pedestrians, and cyclists) are proportionally distributed among the splits.

We use two performance measures following the KITTI benchmark [14]: Average Precision (AP) and Average Orientation Similarity (AOS). For AP, we calculate the intersection over union (IoU) of the predicted and ground truth bounding boxes in 3D, and require a 50% overlap for *car*, and 25% overlap for *pedestrian* and *cyclist* classes as in [14]. Mean AP (mAP) and mean AOS (mAOS) are calculated by averaging class-wise results. We report results for two regions: 1) the entire annotated region (camera FoV up to 50 meters) and 2) a more safety-relevant region called "Driving Corridor," defined as a rectangle on the ground plane in front of the ego-vehicle as $[-4\,m < x < +4\,m,\ z < 25\,m]$ in camera coordinates.

In our experiments, we will refer to several sensor data and feature combinations: *PP-LiDAR* is PointPillars trained on LiDAR data, with the 4 typically used input features: spatial coordinates and intensity. This method will serve as a baseline for our radar-LiDAR comparison experiment. *PP-radar* is also a PointPillars network, but trained on 3+1D radar data with all 5 features, using spatial coordinates, reflectivity, and Doppler. In contrast, *PP-radar* $(no\ X)$ has the feature X removed and is trained only with 4 features. Finally, *PP-radar (N scans)* is a *PP-radar* using N accumulated radar scans as described in Subsection IV-B. The implementation is built on OpenPCDet [41]. All networks are trained in a multi-class fashion.

TABLE IV
RESULTS FOR ALL TESTED METHODS ON THE ENTIRE ANNOTATED AREA AND WITHIN THE "DRIVING CORRIDOR" ONLY

Method	Features	Entire annotated area					In Driving Corridor				
		Car	Pedestrian	Cyclist	mAP	mAOS	Car	Pedestrian	Cyclist	mAP	mAOS
PP-radar (no elevation)	x, y, RCS, v_r	32.4	28.8	34.6	31.9	25.1	68.2	44.2	63.6	58.6	50.1
PP-radar (no Doppler)	x, y, z, RCS,	35.6	21.3	30.4	29.1	22.1	67.3	31.0	58.7	52.3	41.0
PP-radar (no RCS)	x, y, z, v_r	33.9	33.1	42.7	36.6	30.3	66.8	45.3	**67.2**	59.8	55.6
PP-radar	x, y, z, RCS, v_r	**35.9**	**34.9**	**43.1**	**38.0**	**30.5**	**74.1**	**47.8**	67.1	**63.0**	**56.8**
PP-radar (3 scans)	x, y, z, RCS, v_r, t	44.4	40.4	**54.2**	46.3	39.1	78.4	56.9	**76.6**	70.6	67.1
PP-radar (5 scans)	x, y, z, RCS, v_r, t	**44.8**	**42.1**	54.0	**47.0**	**39.6**	**78.8**	**59.2**	76.1	**71.4**	**68.2**
PP-LiDAR (LiDAR)	x, y, z, $intensity$	75.6	55.1	55.4	62.1	49.4	90.8	71.4	82.5	81.6	70.3

Top: Ablation study of radar features. Middle: Study of temporal information. Bottom: LiDAR based detector. Bold face highlights best radar results per section. All class-specific columns involve AP calculated with a 3D IoU (0.5 for car, 0.25 for pedestrian/cyclist).

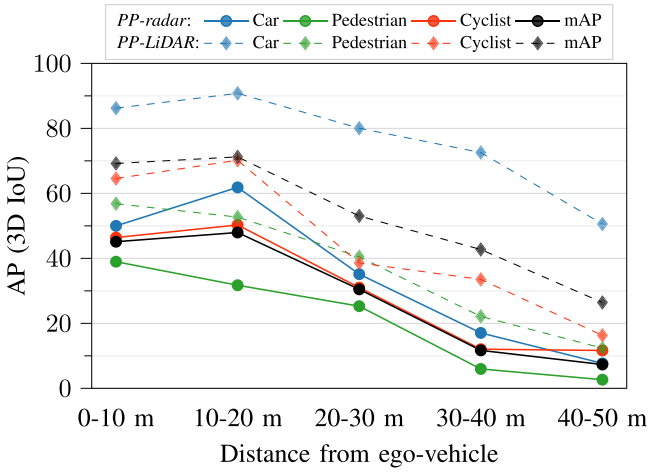


Fig. 4. Performance of *PP-LiDAR* (dashed, diamond) and *PP-radar* (solid, circles) over distance for each class (3D IoU=0.5 for car, IoU=0.25 for pedestrian/cyclist).

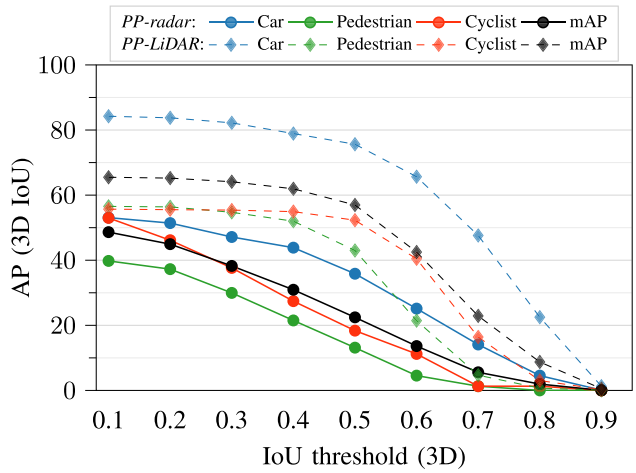


Fig. 5. Performance of *PP-LiDAR* (dashed, diamond) and *PP-radar* (solid, circles) with different 3D IoU thresholds.

A. *Ablation Study:* PP-radar

See Table IV for the performances of the various PointPillars networks in our ablation study, for the entire coverage area and within the "Driving Corridor" region. The results show that removing the Doppler information (*PP-radar (no Doppler)*) significantly degrades performance for the two VRU classes (pedestrian: 34.9 vs. 21.3, cyclist 43.1 vs. 30.4 for the entire annotated area). Furthermore, it hampers the orientation estimation overall (mAOS: 30.5 vs. 22.1). The results also show that removing either elevation information or *RCS* (i.e. *PP-radar (no elevation)* or *PP-radar (no RCS)*) hurts the performance (mAP: 38.0 vs. 31.9 vs. 36.6 for the entire annotated area). Finally, we examined whether including radar targets from previous scans to provide temporal information makes a significant difference. We trained and evaluated two additional networks using points from the last three and five scans, respectively, to create *PP-radar (3 scans)* and *PP-radar (5 scans)*. Adding further scans increased the overall performance (mAP: 38.0 vs. 47.0 for single/five scans) and improved orientation estimation (mAOS: 30.5 vs. 39.6 for single/five scans).

Examples of correct and incorrect detections by *PP-radar* are shown on Figs. 6 and 7 for all road user classes.

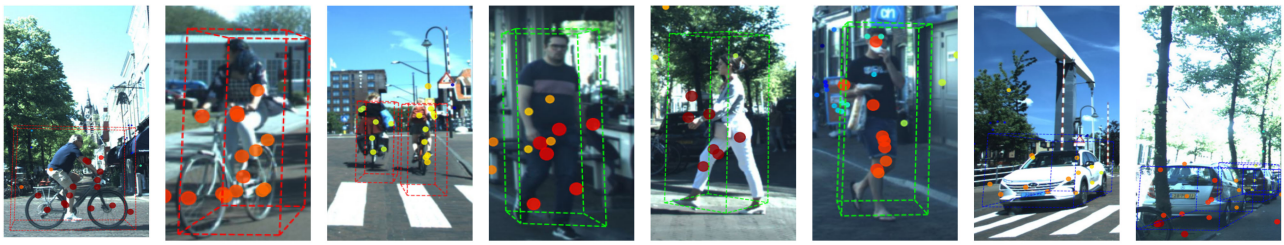
B. *Performance Comparison:* PP-radar *vs.* PP-LiDAR

We subsequently compare the object detection performance of *PP-radar* and *PP-LiDAR*, see Table IV. *PP-LiDAR* outperformed *PP-radar* in all three classes by a clear margin (mAP: 62.1 vs. 38.0). The relative performance gap decreases when we consider only the "Driving Corridor" region (mAP: 81.6 vs. 63.0). Fig. 4 provides performance as a function of distance. See next section for an interpretation of these results. Fig. 5 shows performance as a function of required IoU overlap. An interesting trend that can be seen is that the performance of radar drops off earlier than LiDAR at higher IoU thresholds. This suggests that radar correctly detects and classifies many objects but has difficulty determining their exact 3D position, which hampers overall performance.

On average, *PP-radar* inference took 40% less time than *PP-LiDAR* inference (7.8 ms vs. 12.9 ms on average measuring only the feed-forward step).

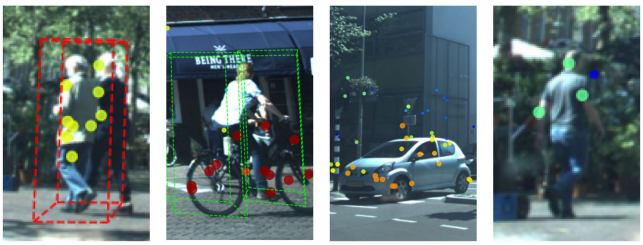
VI. DISCUSSION

In general, object detection performance will be determined by multiple factors: the number of 3D points lying on a particular object of the target class, their individual positional accuracy,



(a) (b) (c) (d) (e) (f) (g) (h)

Fig. 6. Examples of correctly detected objects by *PP-radar* projected to the image plane. Car/pedestrian/cyclist detections are shown as blue/green/red bounding boxes. Dots are radar targets colored by distance from the sensor.



(a) (b) (c) (d)

Fig. 7. Examples of incorrect detections by *PP-radar*: (a) merged smaller objects (two pedestrians are detected as a single cyclist, (b) larger objects split into smaller ones (one cyclist is detected as two pedestrians), (c) strong reflections and clutter nearby (metal poles and high curbs) and (d) distant objects with too few reflections (far away pedestrian).

their spatial configuration and additional attributes (e.g. velocity), their saliency vs. objects of the non-target class, and lastly, the size of the training set.

All radar based methods using Doppler performed best on the *cyclist* class. In contrast to pedestrians, and especially cars, the vast majority of cyclists in the dataset are moving, see Table III. The circular motion of the wheels and pedalling, plus the highly reflective metal frame near the center causes a clear and distinctive reflection pattern that radar can more reliably detect. On the *car* class the radar methods performed more poorly relative to the large size of these objects. This can be explained by the few moving cars in the dataset, and by the fact that many are parked on the other side of the road or canal at larger distances (see Fig. 3), and thus have few reflections. Fig. 4 confirms that nearby cars are detected better. When focusing on just the safety-critical "Driving Corridor" region in front of the vehicle, radar performs considerably better for all classes, see Table IV. This performance is more relevant for driver assistance or automated driving.

The comparison of *PP-LiDAR* and *PP-radar* showed that the former has clearly higher overall performance. This can be attributed to the much higher point density of the specific type of 64-layer LiDAR sensor used (average number of points in the annotated area: LiDAR: 21344, radar: 216). Also the high viewpoint of the LiDAR sensor, on the roof of the car, benefits object detection performance as there is less pronounced occlusion. The radar sensor comes, however, with clear advantages in terms of cost and ease of packaging. Accumulating multiple radar scans was shown to yield substantial performance improvements. This is because of the increased point density, but presumably also because the past scans provide temporal information, which can help classification (change in Doppler signature over time is class-specific, e.g., swinging limbs). Thus using multiple scans closes the relative performance gap to LiDAR somewhat. Compromising on object detection performance might be acceptable if, as a result of the much lower point cloud density, embedding on special hardware (with certain memory and processing limitations) becomes possible. Further improvements of radar resolution and target extraction (i.e., peak finding), and/or the availability of low-level data (e.g. radar cube [3]) could further improve object detection.

VII. Conclusion

We performed an experimental study on multi-class road user detection (PointPillars) on 64-layer 3D LiDAR data and 3+1D radar data. In ablation studies, we showed that the addition of elevation data (as in a next-generation automotive radar) clearly increases object detection performance (from 31.9 to 38.0 mAP). Doppler information remains essential for radar based object detection as its removal would greatly degrade performance (mAP 38.0 vs. 29.1). *RCS* information helps too (mAP 38.0 vs. 36.6 if removed).

Results indicate that object detection on 64-layer LiDAR data still substantially outperforms that on 3+1D radar data, when using the same PointPillars model (mAP 62.1 vs. 38.0). However, accumulating successive radar scans closes the gap to LiDAR to some degree (mAP 62.1 vs. 47.0 for five radar scans) especially in the "Driving Corridor" (mAP 81.6 vs. 71.4 for five radar scans).

For our experimental study, we introduced the View-of-Delft (VoD) dataset, a multi-sensor dataset for multi-class 3D object detection, consisting of calibrated, synchronized, and annotated LiDAR, camera, and 3+1D radar data. It is the largest dataset containing 3+1D radar recordings, suitable to facilitate future research on radar-only, camera-only, LiDAR-only, or fusion methods for object detection and tracking.

REFERENCES

[1] F. Engels, P. Heidenreich, M. Wintermantel, L. Stcker, M. Al Kadi, and A. M. Zoubir, "Automotive radar signal processing: Research directions and practical challenges," *IEEE J. Sel. Topics Signal Process.*, vol. 15, no. 4, pp. 865–878, Jun. 2021.

[2] O. Schumann, C. Wöhler, M. Hahn, and J. Dickmann, "Comparison of random forest and long short-term memory network performances in classification tasks using radar," *Sensor Data Fusion: Trends Solutions, Appl.*, pp. 1–6, 2017.

[3] A. Palffy, J. Dong, J. F. P. Kooij, and D. M. Gavrila, "CNN based road user detection using the 3D radar cube," *IEEE Robot. Automat. Lett.*, vol. 5, no. 2, pp. 1263–1270, Apr. 2020.

[4] R. Pérez, F. Schubert, R. Rasshofer, and E. Biebl, "Single-frame vulnerable road users classification with a 77 GHz FMCW radar sensor and a convolutional neural network," in *Proc. Int. Radar Symp.*, 2018, pp. 1–10.

[5] O. Schumann, M. Hahn, J. Dickmann, and C. Wöhler, "Semantic segmentation on radar point clouds," in *Proc. Int. Conf. Inf. Fusion*, 2018, pp. 2179–2186.

[6] A. Danzer, T. Griebel, M. Bach, and K. Dietmayer, "2D car detection in radar data with PointNets," in *Proc. IEEE Int. Conf. Intell. Transp. Syst.*, 2019, pp. 61–66.

[7] R. Prophet *et al.*, "Pedestrian classification with a 79 GHz automotive radar sensor," in *Proc. Int. Radar Symp.*, 2018, pp. 1–6.

[8] F. Nobis, F. Fent, J. Betz, and M. Lienkamp, "Kernel point convolution LSTM networks for radar point cloud segmentation," *Appl. Sci.*, vol. 11, no. 6, 2021, Art. no. 2599.

[9] A. Cennamo, F. Kaestner, and A. Kummert, "A neural network based system for efficient semantic segmentation of radar point clouds," *Neural Process. Lett.*, vol. 53, no. 5, pp. 3217–3235, 2021.

[10] O. Schumann, J. Lombacher, M. Hahn, C. Wöhler, and J. Dickmann, "Scene understanding with automotive radar," *IEEE Trans. Intell. Veh.*, vol. 5, no. 2, pp. 188–203, Jun. 2020.

[11] M. Meyer and G. Kuschk, "Automotive radar dataset for deep learning based 3D object detection," in *Proc. Eur. Radar Conf.*, 2019, pp. 129–132.

[12] M. Meyer and G. Kuschk, "Deep learning based 3D object detection for automotive radar and camera," in *Proc. Eur. Radar Conf.*, 2019, pp. 133–136.

[13] K. Bansal, K. Rungta, S. Zhu, and D. Bharadia, "Pointillism: Accurate 3D bounding box estimation with multi-radars," in *Proc. Conf. Embedded Networked Sensor Syst.*, 2020, pp. 340–353.

[14] A. Geiger, P. Lenz, C. Stiller, and R. Urtasun, "Vision meets robotics: The KITTI dataset," *Int. J. Robot. Res.*, vol. 32, no. 11, pp. 1231–1237, 2013.

[15] H. Caesar *et al.*, "nuScenes: A multimodal dataset for autonomous driving," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 11618–11628.

[16] P. Sun *et al.*, "Scalability in perception for autonomous driving: Waymo open dataset," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 2443–2451.

[17] A. H. Lang, S. Vora, H. Caesar, L. Zhou, J. Yang, and O. Beijbom, "Pointpillars: Fast encoders for object detection from point clouds," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 12689–12697.

[18] O. Schumann *et al.*, "RadarScenes: A real-world radar point cloud data set for automotive applications," in *Proc. Int. Conf. Inf. Fusion*, 2021, pp. 1–8.

[19] Y. Wang, Z. Jiang, Y. Li, J.-N. Hwang, G. Xing, and H. Liu, "RODNet: A real-time radar object detection network cross-supervised by camera-radar fused object 3D localization," *IEEE J. Sel. Topics Signal Process.*, vol. 15, no. 4, pp. 954–967, 2021.

[20] M. Sheeny, E. De Pellegrin, S. Mukherjee, A. Ahrabian, S. Wang, and A. Wallace, "RADIATE: A radar dataset for automotive perception in bad weather," in *Proc. IEEE Int. Conf. Robot. Automat.*, 2021, pp. 1–7.

[21] M. Mostajabi, C. M. Wang, D. Ranjan, and G. Hsyu, "High resolution radar dataset for semi-supervised learning of dynamic objects," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, 2020, pp. 450–457.

[22] N. Scheiner *et al.*, "Seeing around street corners: Non-line-of-sight detection and tracking in-the-wild using Doppler radar," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 2065–2074.

[23] A. Angelov, A. Robertson, R. Murray-Smith, and F. Fioranelli, "Practical classification of different moving targets using automotive radar and deep neural networks," *Inst. Eng. Technol. Radar Sonar Navigation*, vol. 12, no. 10, pp. 1082–1089, 2018.

[24] C. Diehl, E. Feicho, A. Schwambach, T. Dammeier, E. Mares, and T. Bertram, "Radar-based dynamic occupancy grid mapping and object detection," in *Proc. IEEE Int. Conf. Intell. Transp. Syst.*, 2020, pp. 1–6.

[25] V. John and S. Mita, "RVNet: Deep sensor fusion of monocular camera and radar for image-based obstacle detection in challenging environments," *Lecture Notes Comput. Sci.*, vol. 11854, pp. 351–364, 2019.

[26] V. John, M. K. Nithilan, S. Mita, H. Tehrani, R. S. Sudheesh, and P. P. Lalu, "SO-Net: Joint semantic segmentation and obstacle detection using deep fusion of monocular camera and radar," *Lecture Notes Comput. Sci.*, vol. 11994, pp. 138–148, 2020.

[27] L. Wang, T. Chen, C. Anklam, and B. Goldluecke, "High dimensional frustum PointNet for 3D object detection from camera, LiDAR, and radar," in *Proc. IEEE Intell. Veh. Symp.*, 2020, pp. 1621–1628.

[28] F. Nobis, M. Geisslinger, M. Weber, J. Betz, and M. Lienkamp, "A deep learning-based radar and camera sensor fusion architecture for object detection," *Sensor Data Fusion: Trends Solutions, Appl.*, 2019, pp. 1–7.

[29] R. Prophet, A. Deligiannis, J.-C. Fuentes-Michel, I. Weber, and M. Vossiek, "Semantic segmentation on 3D occupancy grids for automotive radar," *IEEE Access*, vol. 8, pp. 197917–197930, 2020.

[30] J. Lombacher, M. Hahn, J. Dickmann, and C. Wöhler, "Potential of radar for static object classification using deep learning methods," in *Proc. IEEE MTT-S Int. Conf. Microw. Intell. Mobility*, 2016, pp. 1–4.

[31] J. Lombacher, K. Laudt, M. Hahn, J. Dickmann, and C. Wöhler, "Semantic radar grids," in *Proc. IEEE Intell. Veh. Symp.*, 2017, pp. 1170–1175.

[32] A. Palffy, J. F. P. Kooij, and D. M. Gavrila, "Occlusion aware sensor fusion for early crossing pedestrian detection," in *Proc. IEEE Intell. Veh. Symp.*, 2019, pp. 1768–1774.

[33] D. Kellner, M. Barjenbruch, K. Dietmayer, J. Klappstein, and J. Dickmann, "Instantaneous lateral velocity estimation of a vehicle using Doppler radar," in *Proc. Int. Conf. Inf. Fusion*, 2013, pp. 877–884.

[34] P. Held, D. Steinhauser, A. Kamann, A. Koch, T. Brandmeier, and U. T. Schwarz, "Normalization of micro-Doppler spectra for cyclists using high-resolution projection technique," in *Proc. IEEE Int. Conf. Veh. Electron. Saf.*, 2019, pp. 1–6.

[35] G. Kim, Y. S. Park, Y. Cho, J. Jeong, and A. Kim, "MulRan: Multimodal range dataset for urban place recognition," in *Proc. IEEE Int. Conf. Robot. Automat.*, 2020, pp. 6246–6253.

[36] D. Barnes, M. Gadd, P. Murcutt, P. Newman, and I. Posner, "The oxford radar RobotCar dataset: A radar extension to the oxford RobotCar dataset," in *Proc. IEEE Int. Conf. Robot. Automat.*, 2019, pp. 6433–6438.

[37] J. Bai, L. Zheng, S. Li, B. Tan, S. Chen, and L. Huang, "Radar transformer: An object classification network based on 4D MMW imaging radar," *Sensors*, vol. 21, no. 11, 2021, Art. no. 3854.

[38] A. Ouaknine, A. Newson, J. Rebut, F. Tupin, and P. Pérez, "CARRADA dataset: Camera and automotive radar with range- angle- Doppler annotations," in *Proc. Int. Conf. Pattern Recognit.*, 2021, pp. 5068–5075.

[39] L. Ferranti *et al.*, "SafeVRU: A research platform for the interaction of self-driving vehicles with vulnerable road users," in *Proc. IEEE Intell. Veh. Symp.*, 2019, pp. 1660–1666.

[40] J. Domhof, J. F. P. Kooij, and D. M. Gavrila, "A joint extrinsic calibration tool for radar, camera and lidar," *IEEE Trans. Intell. Veh.*, vol. 6, no. 3, pp. 571–582, Sep. 2021.

[41] *OpenPCDet Develop. Team*, "OpenPCDet: An open-source toolbox for 3D object detection from point clouds," 2020. [Online]. Available: https://github.com/open-mmlab/OpenPCDet