Rethinking of Radar's Role: A Camera-Radar Dataset and Systematic Annotator via Coordinate Alignment

Yizhou Wang[1], Gaoang Wang[2], Hung-Min Hsu[1], Hui Liu[1,3], Jenq-Neng Hwang[1]
[1]University of Washington, Seattle, WA, USA
[2]Zhejiang University, Hangzhou, China
[3]Silkwave Holdings Limited, Hong Kong, China
{ywang26,hmhsu,huiliu,hwang}@uw.edu, gaoangwang@intl.zju.edu.cn

Abstract

Radar has long been a common sensor on autonomous vehicles for obstacle ranging and speed estimation. However, as a robust sensor to all-weather conditions, radar's capability has not been well-exploited, compared with camera or LiDAR. Instead of just serving as a supplementary sensor, radar's rich information hidden in the radio frequencies can potentially provide useful clues to achieve more complicated tasks, like object classification and detection. In this paper, we propose a new dataset, named CRUW[1], with a systematic annotator and performance evaluation system to address the radar object detection (ROD) task, which aims to classify and localize the objects in 3D purely from radar's radio frequency (RF) images. To the best of our knowledge, CRUW is the first public large-scale dataset with a systematic annotation and evaluation system, which involves camera RGB images and radar RF images, collected in various driving scenarios.

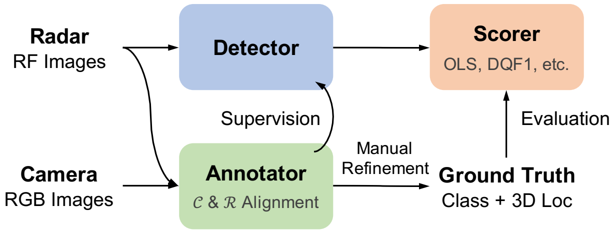


Figure 1. The pipeline of the radar object detection (ROD) task, including a detector, an annotator, and a scorer. The detector takes radar data as the only input and can predict object classes and 3D locations in adverse driving scenarios. The annotator manages to align detections between camera $\mathcal{C}$ and radar $\mathcal{R}$ coordinates to serve as a "teacher" during training the "student" detector. The ground truth of the testing set for evaluation purpose is further refined manually from annotator's results. Finally, a scorer, with defined evaluation metrics, is used on the testing set to evaluate the performance.

1. Introduction

Multi-modality data analytics is greatly involved in the autonomous or assisted driving systems [17, 36, 35] to improve the robustness of object perception [26, 31, 16, 37] in a variety of different driving scenarios. Among the common sensors, i.e., camera, LiDAR, radar, on the autonomous vehicles, the RGB images and point cloud data from cameras and LiDAR are relatively easy for human to understand since the semantic information they convey is obvious. For example, 2D and 3D bounding boxes are intuitive for human to annotate the objects from RGB images and LiDAR point clouds, respectively. Therefore, some large and well-labeled datasets [13, 8, 4, 3] have been released in the autonomous driving community to help develop and validate the machine learning algorithms. As an accurate 3D sensor for autonomous vehicles, LiDAR still faces the following critical limitations: 1) LiDAR is usually equipmental complex and computational expensive, so that not suitable for common industry use; 2) Laser transmitted by the LiDAR is not robust to occlusion or adverse weather scenarios. Radar, on the other hand, is a reliable and cost-efficient sensor capturing reliable 3D information even under adverse driving conditions, e.g., strong/weak lighting or bad weather. It is often used as supplement for other sensors due to its difficulty in parsing useful clues for semantic understanding. But this data unintuitiveness does not mean that radar has low potentials.

The use of the frequency modulated continuous wave (FMCW) radar in most autonomous driving solutions lies in its obstacle ranging and speed estimation. The capability of radar's semantic understanding, e.g., object classification and detection, has not been well-exploited. The fea-

[1] Dataset available at https://www.cruwdataset.org/.

sibility of achieving this semantic understanding owing to the hidden ***phase*** information inside the radio frequencies. Typically, radar's amplitude is commonly used to estimate the distance and speed of the obstacles, while the phase information is usually not well-utilized because of its "non-intuitiveness", making it difficult to be handled by the classical signal processing mechanisms. There are two kinds of data representations for the FMCW radar, i.e., radio frequency (RF) image and radar points (see examples in the supplementary document). RF image is a much denser and more informative data representation, which requires further processing to understand the contents, containing both amplitude and phase information, but the location and speed are implicit. Whereas radar points are a kind of handy representation, which are usually sparse (less than 5 points on a nearby car) [8, 12] and non-descriptive.

Therefore, to extract those hidden features from RF images for semantic understanding, some researchers start to take advantage of the recent deep convolution neural networks (CNNs) to explore the possibility of radar-only object detection [19, 10, 33], which usually require a large annotated dataset for training. However, radar data are very difficult to understand, making human annotations significantly expensive or sometimes impossible to obtain. Besides, due to the low angular resolution of the common FMCW radar sensors, resulting in unreliable object dimension information from radar; more specifically, the bounding boxes defined in camera-based object detection are rarely used in the RF images, especially when the absence of LiDAR. Consequently, people usually represent objects as ***points*** in the radar's bird's-eye view (BEV) coordinates instead [33]. These points are the reflection points of the radar signals from the obstacles in the radar's field of view (FoV). All of the above make the large-scale dataset, annotation and performance evaluation for object detection in the RF images very challenging while critically needed.

In this paper, we propose a platform, including a large-scale dataset, a systematic annotation system, and a set of evaluation metrics for ROD, as shown in Fig. 1. The proposed dataset contains about 400K synchronized camera-radar frames (3.5 hours) in various driving scenarios, i.e., parking lot, campus road, city street, and highway. As mentioned above, the radar data format is RF image with rich radio frequency information. The proposed object annotation method calculates the optimal camera-radar bilateral coordinate projection and aligns the detections between the two coordinates. Two different kinds of detections are fed into the system, including object detection from camera and peak detection from radar. Then, a coordinate alignment strategy is utilized based on the proposed bilateral coordinate projection between camera and radar, and the ground plane for each frame is optimized by the alignment cost. In order to evaluate the performance of ROD, which predicts objects as points, we introduce a point-based similarity metric, called object location similarity (OLS), to serve as a matching score between two points in an RF image. Based on OLS, to reasonably reflect the quality of the object detection results in the RF images, we introduce a series of evaluation metrics, considering object localization error, precision, and recall. Moreover, we define a new metric Detection Quality F1 Score (DQF1) that can jointly consider the above three metrics into a single metric to provide a comprehensive measurement of the detection performance. Unlike the widely used average precision (AP) and average recall (AR) defined for object detection tasks that emphasize classification, the proposed DQF1 focuses more on the localization accuracy.

Overall, the main contributions of this paper can be summarized as follows,

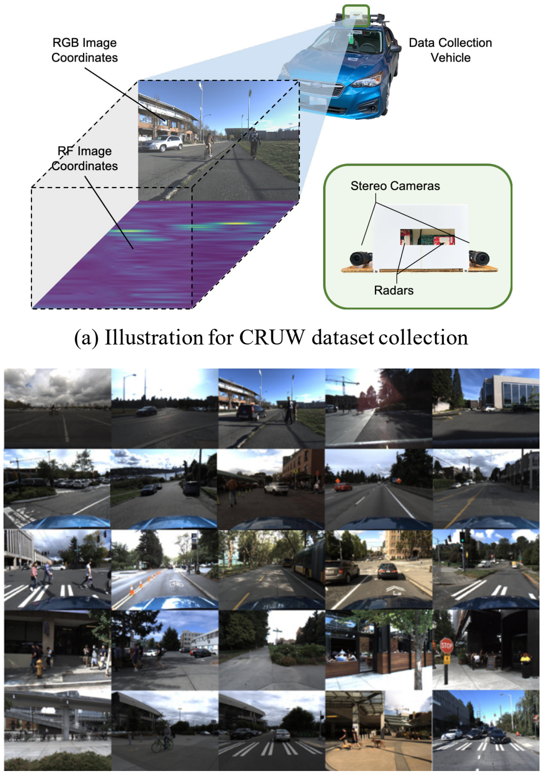
- A large-scale dataset with synchronized camera-radar frames in various driving scenarios, including RF images as the radar data format for radar semantic understanding tasks.
- An accurate and robust radar object annotator, that can systematically generate object labels for RF images, fusing the rich semantic information from a camera. It is also an object detector based on a camera-radar fusion manner in the normal driving scenarios.
- Derive the bilateral coordinate projection (BCP) between camera pixel coordinates and radar range-azimuth coordinates through the ground plane.
- Introduce a set of scoring metrics to evaluate the quality of ROD results comprehensively, including the proposed metrics, i.e., object location similarity (OLS) and detection quality F1 score (DQF1).

2. Related Works

Detectors. As mentioned in Section 1, radar-only object detection cannot reliably accomplish, especially the object-class identification, with sparse and non-descriptive radar points input. Therefore, most related research use the RF images as the input format. However, detecting objects from the radar RF data is very challenging because the inherent semantic information is not as obvious as that in the RGB images. Traditionally, peak detection algorithms are adopted to find the objects in the radar field of view (FoV), such as the widely used Constant False Alarm Rate (CFAR) detection algorithm [27]. A classifier is then appended to classify the object class [15, 5]. However, these algorithms usually result in a large number of *false positives* because it cannot distinguish the reflections of objects from that of background. Besides, it cannot reliably provide the object class. Moreover, one object may give multiple CFAR detections, which are confusing. Recently, some new techniques for radar data are proposed. Major *et al.* [19] propose an automotive radar based vehicle detection method trained by LiDAR. However, they only consider vehicles as the target object class, and the scenarios are mostly highways without noisy obstacles. Palffy *et al.* [24] propose a radar based, single-frame multi-class object detection method. However, they only consider the data from a single radar frame, which does not involve the object motion information. Wang *et al.* [34] propose the RODNet with temporal inception layers to capture temporal features of different lengths, an M-Net is used to merge and extract Doppler information from multiple chirps, and temporal deformable convolution is used to handle object relative motion.

Datasets. Datasets are important to validate the algorithms, especially for the deep learning based methods. Since the release of the first complete autonomous driving dataset, i.e., KITTI [13], larger and more advanced datasets are now available [3, 4, 8]. However, due to the hardware compatibility and less developed radar perception techniques, most datasets do not incorporate radar signals as a part of their sensor systems. Among the available radar datasets (summarized in Table 1), some of them [8, 20, 7, 29] consider radar data in the format of radar points that do not contain the useful Doppler and surface texture information of objects. Later, researchers start to focus on RF images as the radar data format. More specifically, some manage to collect a dataset with camera, radar and LiDAR, and annotate the objects as 3D bounding boxes based on the dense point cloud from LiDAR [19, 10]. Others consider the camera-radar solution without a LiDAR [23, 24], whose annotation format is usually in pixel or point level. However, most of the datasets with RF images are not publicly available except CARRADA [23]. But CARRADA only contains one simple and easy scenario, i.e., parking lot, and is not suitable for practical usage.

Annotation methods. When LiDAR is not available to provide reliable object annotations, many people manage to fuse the semantic information from cameras. There are some camera-based techniques that are helpful for the radar object annotation task. Camera-based object detection [26, 14, 9, 25, 18, 11] aims to detect every object with its class and precise bounding box location from RGB images. Besides, some recent works try to infer 3D information from the 2D RGB images. Some methods [21, 22] localize vehicles by estimating their 3D structures using a CNN. Others [30, 6] try to develop a real-time monocular structure-from-motion (SfM) system, taking into account different kinds of cues. However, the above methods only work for the vehicles, which can satisfy the rigid-body structure assumption. To overcome this limitation, Wang *et al.* [32] propose an accurate and robust object 3D localization system, based on the detected and tracked 2D bounding boxes of objects, which can work for most common moving objects in the road scenes, such as cars, pedestrians, and cyclists. The limitation is its inability to reliably estimate the 3D location of objects due to the inaccurate depth maps and non-negligible car sizes. After that, [33] proposes a probabilistic camera-radar fusion (CRF) algorithm to jointly consider the object 3D localization results from both camera and radar. But this kind of late fusion methods may introduce the errors from the early stages nor consider the correlation between different objects in the same frame. Recently, Ouaknine *et al.* [23] propose a semi-automatic annotation approach on radar data of some simple parking lot scenarios. However, its sensitivity to tracking and clustering makes it not robust in complex driving scenarios. Therefore, in this paper, we propose an annotation system to overcome the above issues.



(a) Illustration for CRUW dataset collection

(b) Different scenarios in CRUW dataset

Figure 2. The sensor platform and some sample driving scenarios.

3. CRUW Dataset

3.1. Sensor System and Data Description

The sensor platform contains a pair of stereo cameras [1] and two perpendicular 77GHz FMCW mmWave radar antenna arrays [2]. The sensors are assembled and mounted together as shown in Fig. 2 (a). Some configurations of our sensor platform are shown in Table 2.

Dataset	Year	Scale	Scenarios*	Radar Format	Classes	Anno Source	Anno Format	Public
nuScenes [8]	2019	5.5 hours	ML	Radar Points	23	LiDAR	3D Box	✓
Qualcomm [19]	2019	3 hours	HW	RF Images	1	LiDAR	3D Box	✗
Astyx HiRes2019 [20]	2019	546 frames	Urban	Radar Points	7†	LiDAR	3D Box	✓
RadarRobotCar [7]	2020	280 km	Urban	Radar Points	0	–	–	✓
CARRADA [23]	2020	21.2 min	PL	RF Images	3	Camera	Pixel	✓
Xsense.ai [10]	2020	34.2 min	HW	RF Images	1	LiDAR	3D Box	✗
RTCnet [24]	2020	1 hour	Urban	RF Images	3	Camera	Point	✗
RADIATE [29]	2020	3 hours	ML	Radar Points	7	LiDAR§	2D Box	✓
CRUW (Ours)	2021	3.5 hours	ML	RF Images	3	Camera	Point	✓

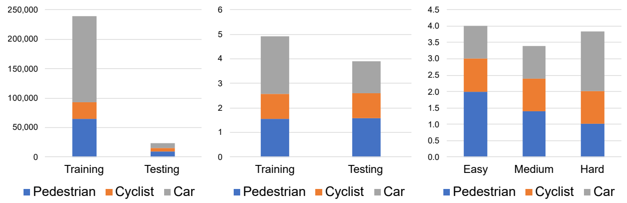
Table 1. Related datasets with radar data. *ML: multiple scenarios; HW: highway; PL: parking lot. †Significantly imbalanced object distribution where car is the majority class. §Details are not mentioned in the paper.

Camera	Value	Radar	Value
Frame rate	30 FPS	Frame rate	30 FPS
Pixels (W×H)	1440×1080	Frequency	77 GHz
Resolution	1.6 MP	# of transmitters	2
Field of View	93.6°	# of receivers	4
Stereo Baseline	0.35 m	# of chirps per frame	255
		Range resolution	0.23 m
		Azimuth resolution	∼15°

Table 2. Sensor Configurations for CRUW Dataset.

Scenarios	# of Seqs	# of Frames	Vision-Hard %
Parking Lot	124	106K	15%
Campus Road	112	94K	11%
City Street	216	175K	6%
Highway	12	20K	0%
Overall	464	396K	9%

Table 3. Driving scenarios statistics for CRUW dataset.



(a) # of objects in total

(b) # of objects per frame

(c) # of objects per frame in the testing set

Figure 3. Illustration for our CRUW dataset distribution. Here, (a)-(c) show the object distribution in the radar's FoV (0-25m, ±90°).

The proposed dataset contains 3.5 hours with 30 FPS (about 400K frames) of camera-radar data in different driving scenarios, including parking lot, campus road, city street, and highway. Some sample scenarios are shown in Fig. 2 (b). The data are collected in two different views, i.e., driver front view and driver side view, to validate different perspective views for autonomous or assisted driving. Besides, we also collect several vision-hard sequences of poor image quality, i.e., weak/strong lighting, blur, etc. These data are only used in the testing set for evaluation purpose.

3.2. Data Distribution

The data distribution is shown in Table 3 and Fig. 3. The object statistics in Fig. 3 only consider the objects within the radar's field of view (FoV), i.e., 0-25m, ±90°, based on the current hardware capability. There are about 260K objects in CRUW dataset in total, including 92% for training and 8% for testing. The average number of objects in each frame is similar between training and testing data.

The four different driving scenarios, i.e., parking lot, campus road, city street, and highway, are shown in Table 3 with the number of sequences, frames and vision-hard percentages. From each scenario, we randomly select several complete sequences as testing sequences, which are not used for training. Thus, the training and testing sequences are captured at different locations and different time. For the ground truth needed for evaluation purposes, 10% of the visible and 100% of the vision-hard data are human-labeled by manually refinement on the results of the annotator introduced in Section 4.

4. Radar Object Annotation System

In this section, our proposed radar object annotation system (Fig. 4) will be introduced. First, the input data, both from camera and radar, are pre-processed and fed into our systematic annotation system for initialization. Second, the bilateral coordinate projection is derived to connect the camera and radar coordinates. Then, the detection alignment strategy with the ground plane optimization is applied to accurately detect and localize the objects in RF images. Overall, the radar detections are aligned and clustered by different instances, and the final object annotations are the centers of the resulting clusters.

4.1. Notations

Coordinates. We first define two 2D coordinates used in our system: 1) Camera 2D pixel coordinates $(u, v) \in \mathcal{C}$; 2) Radar range-azimuth coordinates, i.e., bird's-eye view

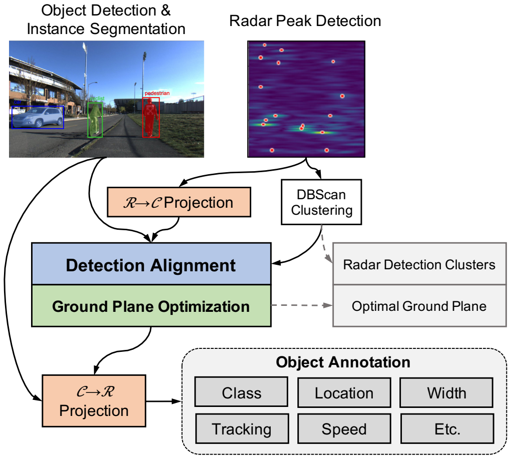


Figure 4. The framework of our proposed annotation system for automatic radar object annotation. The system takes detections inferred from both RGB images and RF images, and manages to align the detections between the RGB image coordinates $\mathcal{C}$ and radar range-azimuth coordinates $\mathcal{R}$ through the corresponding ground plane. After detections are aligned, the system can provide reliable object annotation taking advantages of camera information, including object attributes, e.g., class, location, reflection width, tracking, speed, etc., as well as radar detection clusters and the optimal ground planes.

(BEV), $(r, \theta) \in \mathcal{R}$; 3) 3D camera coordinates with the origin at the camera center $(x^c, y^c, z^c) \in \mathcal{W}^c$; and 4) 3D radar coordinates with the origin at the radar $(x^r, y^r, z^r) \in \mathcal{W}^r$. The system takes two different kinds of detections from both camera and radar, i.e., object detection from RGB images and peak detection from RF images, as the input, and these detections are defined within $\mathcal{C}$ and $\mathcal{R}$.

Ground plane parameters. We define the ground plane by three ground plane parameters, i.e., two rotation angles and one offset. The ground plane parameters for each frame can be represented as

$$\mathbf{g} = [\varphi, \gamma, h], \tag{1}$$

where φ and γ respectively denote the pitch and roll angles of the ground plane w.r.t. the z-axis of $\mathcal{W}^c$; h is the camera height. The illustration of these ground plane parameters is shown in Fig. 5.

4.2. Initialization

In this work, we use RF images to represent our radar signal reflections. They are in radar 2D range-azimuth coordinates $(r, \theta) \in \mathcal{R}$, where the θ-axis denotes azimuth (angle) and the r-axis denotes range (distance). Two different kinds of detections from camera and radar are used as the input to our system:

- Object detection and instance segmentation from RGB images using a Mask R-CNN object detector [14].

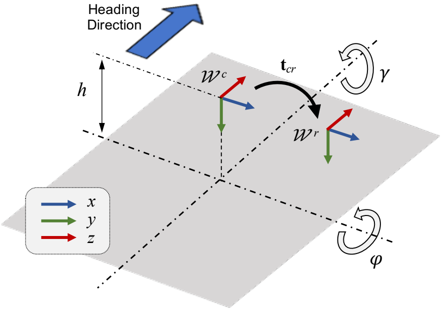


Figure 5. Illustration for 3D coordinates and ground plane parameters. The gray rectangle represents the ground plane. The camera and radar mounted on a vehicle are above the ground plane. $\mathbf{t}_{cr}$ is the translation vector from $\mathcal{W}^c$ to $\mathcal{W}^r$.

- Radar peak detection on RF images using the CFAR detection algorithm [27].

After the data and detections are prepared, we need to initialize the systematic annotation system. First, the ground plane parameters are initialized by the sensor calibration results. Specifically, for the CRUW dataset, $\varphi_0 = 4°$, $\gamma_0 = 0°$, $h_0 = 1.65$m. Second, the density-based spatial clustering of applications with noise (DBScan) clustering algorithm [28] is implemented on the CFAR detections to obtain the initial detection clusters.

4.3. Bilateral Coordinate Projection

Consider the ground plane parameters defined in Section 4.1, we derive the camera-radar bilateral coordinate projection (BCP) for any point on the ground plane. We start from projecting a point $(r, \theta) \in \mathcal{R}$ to $\mathbf{x}^c = (x^c, y^c, z^c)$.

We first do a polar to Cartesian coordinate transformation and transform to $\mathcal{W}^c$ by sensor calibration of the translation from camera to radar $\mathbf{t}_{cr} = [t_{cr,x}, t_{cr,y}, t_{cr,z}]^\top$,

$$\begin{aligned} x^c &= r\sin(\theta) + t_{cr,x}, \\ z^c &= r\cos(\theta) + t_{cr,z}. \end{aligned} \tag{2}$$

Here, we ignore the rotation between $\mathcal{W}^r$ and $\mathcal{W}^c$ since both sensors are well-calibrated with the same orientation.

To calculate y^c, we first consider the ground plane with only pitch rotation angle φ. Assuming a small φ, which is a valid assumption in driving scenarios, y^c_φ can be approximated by the similar triangles rule as

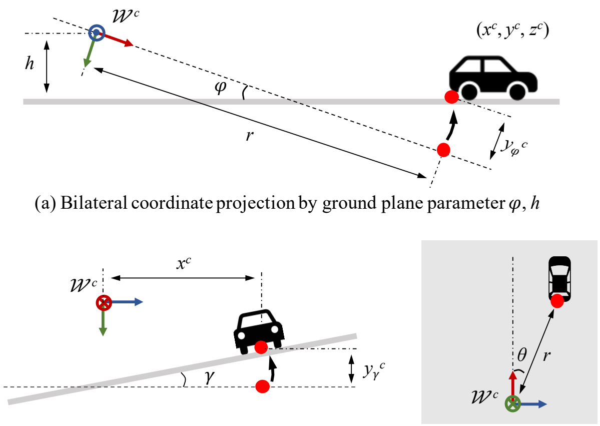
$$y^c_\varphi \approx h - r\sin(\varphi). \tag{3}$$

By taking the second rotation angle γ into consideration,

$$y^c_\gamma = x^c \tan(\gamma). \tag{4}$$

The illustrations for these two projections are shown in Fig. 6 (a) and (b). Therefore, the overall y^c can be represented as

$$y^c = y^c_\varphi - y^c_\gamma = h - r\sin(\varphi) - x^c\tan(\gamma). \tag{5}$$



(a) Bilateral coordinate projection by ground plane parameter φ, h

(b) Bilateral coordinate projection by ground plane parameter γ

(c) Radar's BEV

Figure 6. Geometric illustrations for our proposed bilateral coordinate projection. The projection is split into two steps, i.e., φ rotation and γ rotation, where the gray lines represent the ground planes. To clearly show the projection, these two steps are shown in two different perspectives, where (a) is based on the parameters φ and h; (b) is based on the parameter γ. (c) represents a car in the radar's BEV with range r and azimuth angle θ.

Finally, we project $\mathbf{x}^c \in \mathcal{W}^c$ to $\mathcal{C}$ by the camera intrinsic matrix $\mathbf{K}$. Put it together, we can project a point $(r, \theta) \in \mathcal{R}$ to $(u, v) \in \mathcal{C}$ by the following $\mathcal{R} \to \mathcal{C}$ projection, denoted as $P(\cdot)$,

$$[u, v]^\top = P(r, \theta), \tag{6}$$

where

$$\begin{aligned} u &= f_x \frac{r\sin(\theta) + t_{cr,x}}{r\cos(\theta) + t_{cr,z}} + c_x, \\ v &= f_y \frac{h - r\sin(\varphi) - (r\sin(\theta) + t_{cr,x})\tan(\gamma)}{r\cos(\theta) + t_{cr,z}} + c_y. \end{aligned} \tag{7}$$

Here, (f_x, f_y) represents the camera's focal length; (c_x, c_y) represents the camera center.

After the $\mathcal{R} \to \mathcal{C}$ projection is properly derived, we would like to consider the other direction, i.e., $\mathcal{C} \to \mathcal{R}$ projection. Since the projection in Eq. 6 is a 2D-to-2D projection without losing any degrees of freedom (DoF), it can be reversed. Therefore, we define the $\mathcal{C} \to \mathcal{R}$ projection as

$$[r, \theta]^\top = P^{-1}(u, v). \tag{8}$$

Here, the $\mathcal{C} \to \mathcal{W}^c$ projection can be derived as

$$\begin{aligned} x^c &= \frac{h\hat{x}}{\sqrt{1 + \hat{x}^2}\sin(\varphi) + \hat{x}\tan(\gamma) + \hat{y}}, \\ z^c &= \frac{h}{\sqrt{1 + \hat{x}^2}\sin(\varphi) + \hat{x}\tan(\gamma) + \hat{y}}, \end{aligned} \tag{9}$$

where

$$\begin{aligned} \hat{x} &= \frac{x^c}{z^c} = \frac{u - c_x}{f_x}, \\ \hat{y} &= \frac{y^c}{z^c} = \frac{v - c_y}{f_y}. \end{aligned} \tag{10}$$

4.4. Detection Alignment and Optimization

Based on the BCP, the connection between camera and radar is established through the ground plane parameters. To align these two coordinates, we come up with a detection alignment strategy by ground plane optimization.

Detection alignment cost. The alignment between $\mathcal{R}$ and $\mathcal{C}$ is challenging because they are both 2D coordinates with significantly different perspectives. However, we find that the object height is very useful information during the alignment. The object will have very different heights at different distances in the RGB image due to the perspective. Thus, the object height is a special cue of its 3D location. In the alignment, we use the average height for each object class and project all the CFAR peak detections to RGB images by $\mathcal{R} \to \mathcal{C}$ projection with the object height. Each CFAR detection will be projected as a line segment in the RGB image, named as CFAR line. We define the detection alignment cost for a CFAR detection i to be

$$\ell_i = \lambda_i \left(h_i - h_i^{mask}\right)^2 + (1 - \lambda_i)\left(h_i - h_i^{bbox}\right)^2, \tag{11}$$

where h_i is the height of the projected CFAR detection in the RGB image; h_i^{mask} and h_i^{bbox} are the heights of the object mask and bounding box near the projected CFAR line, respectively. ℓ_i is the combined cost from object mask and bounding box based on an adaptive weight λ_i, where

$$\lambda_i = e^{-\alpha z_i^c}, \tag{12}$$

which means λ_i is dependent on z_i^c. The intuition behind is that, for nearby objects, 2D object masks can accurately describe the height of the objects, especially for the cars, whereas bounding boxes fail to do that. While for faraway objects, bounding boxes can perform better than the unreliable masks. Here, α is a parameter to distinguish between nearby and faraway objects. During the experiment, we empirically find 10 meters is a good threshold, so that we set $\alpha = 0.06$ to make $\lambda = e^{-10\alpha} \approx 0.5$.

Ground plane optimization. After the detections are aligned between camera and radar, the ground plane parameters can be optimized accordingly. Here, we define the objective function as

$$\min_{\varphi,\gamma} \sum_t \sum_{i=1}^{n_{CFAR}} \left(v_{i,t} - v_{i,t}^{mask}\right)^2, \tag{13}$$

where n_{CFAR} is the number of CFAR detections in the frame t; $v_{i,t}$ is the vertical pixel location of the CFAR detection, i.e., v-axis of the lower endpoint of the CFAR line;

$v_{i,t}^{mask}$ is the bottom location of the object mask near the projected CFAR line.

When the ground plane is optimized for each frame, we can use the $\mathcal{C} \rightarrow \mathcal{R}$ projection to project the detections in the RGB image that are not aligned. Here, we call this procedure *supplementary projection*. Since the camera usually can detect more objects than radar, this projection can significantly improve the detection recall.

5. Point-based Detection Evaluation

As mentioned in Section 1, the common datasets with the FMCW radar [8, 7, 23, 29, 33] usually use points to represent detections, we propose an evaluation system for point-based detection without any bounding box.

5.1. Point-based Similarity

Before introducing the evaluation metrics, the matching strategy between the generated annotations and the ground truth needs to be explained. We propose the object location similarity (OLS) to represent the similarity between a detection i and a ground truth j to be

$$\text{OLS}(i,j) = \exp\left\{ \frac{-d_{ij}^2}{2(s_j \kappa_{cls})^2} \right\}, \tag{14}$$

where d is the distance (in meters) between the two points in an RF image; s is the object distance from the sensors; and κ_{cls} is a per-class constant that represents the error tolerance for class cls, which can be determined by the object average size of the corresponding class.

5.2. Scoring Metrics

The quality of the radar object detection includes two aspects: 1) The location accuracy of the object detections; 2) The object-class correctness in the radar's FoV.

Typically, to evaluate the location accuracy, the *mean absolute error* (MAE) is used to calculate the absolute localization error in meters. But MAE cannot reflect the false positives or false negatives, i.e., wrong or missing detections. Therefore, we also include *precision* and *recall* into our evaluation metrics.

Moreover, to better describe the quality of the detections, we define a new evaluation metric called *Detection Quality F1 Score* (DQF1), that aims to jointly consider MAE, precision and recall into one single number of measurement. Considering the F1 score that is frequently used to combine the precision and recall, we define DQF1 as

$$\text{DQF1} = \frac{2}{n_{det} + n_{gt}} \sum_{j=1}^{n_{gt}} \sum_{i=1}^{n_{det}} \delta_{i,j} \cdot \text{OLS}(i,j), \tag{15}$$

where n_{det} is the number of detections; n_{gt} is the number of ground truth; $\delta_{i,j}$ is a binary flag to illustrate whether the i-th detection is matched with the j-th ground truth. Note that we use the summation of OLS to replace the number of true positives, so that the object localization accuracy is also involved. Since OLS is between 0 and 1, DQF1 is also a metric between 0 and 1.

6. Baseline Evaluation

6.1. Evaluation on Radar Object Detection

We first evaluate the radar-only object detection performance of the state-of-the-art method, called RODNet [33, 34], using our proposed evaluation system. Vanilla, Hourglass (HG), and Full are three different network configurations mentioned in the original paper. The evaluation scores are shown in Table 4. The comprehensive metrics for radar object detection include AP, AR, and the proposed DQF1. AP and AR used in the original paper are classical metrics for object detection, while DQF1 focuses more on localization accuracy instead of classification.

It is obvious that the localization accuracy of the RODNet is very promising since it considers radar data as the only input. Therefore, the DQF1 scores are also a little bit higher than the original AP and AR, which shows the property of DQF1 that emphasize more on localization. Overall, with AP, AR, and DQF1, we can have an overall impression of the point-based detection performance from both classification and localization.

6.2. Radar Object Annotation Comparison

We use our proposed annotation system to automatically generate object annotations on CRUW dataset and evaluate the annotation quality by the evaluation metrics mentioned in Section 5.2, on the selected crowded training set. Note that the MAE is evaluated by both mean and standard deviation of the localization errors.

We compare the proposed annotation system with a camera-only (CO) object 3D localization method [32] and CRF algorithm [33]. Comparing with CO, the MAE achieved by our proposed annotator can be decreased by as much as 40% after taking radar into consideration. Besides, our system can achieve over 90% for both the precision and recall. Overall, the DQF1 score of our system outperforms CO [32] by about 15%. Comparing with CRF [33], the MAE and precision are similar but the recall of CRF is much worse, which means that CRF has a number of false negatives. These false negatives may be due to the large errors from the results of the CO method. Overall, our annotator improves the DQF1 performance by around 6%, compared with CRF. The qualitative results are shown in Fig. 7, where the objects are accurately annotated even if they are not detected by the radar.

We also use different time window sizes t in the ground plane optimization (Eq. 13) and evaluate the performance

Method	Scenario	MAE	Precision	Recall	**AP**	**AR**	**DQF1**
RODNet (Vanilla) [33]	Overall	0.31 (±0.26)	95.90%	78.03%	74.29%	77.85%	81.02%
	Parking Lot	0.26 (±0.19)	98.29%	87.76%	85.33%	86.76%	89.33%
	Campus Road	0.42 (±0.30)	89.49%	53.02%	42.67%	49.03%	56.03%
	City Street	0.48 (±0.39)	88.88%	73.42%	59.79%	67.23%	71.15%
RODNet (HG) [33]	Overall	0.31 (±0.23)	96.02%	88.56%	83.76%	85.62%	86.64%
	Parking Lot	0.26 (±0.16)	98.26%	96.94%	93.60%	94.98%	93.63%
	Campus Road	0.40 (±0.26)	92.16%	68.76%	50.34%	57.23%	70.28%
	City Street	0.48 (±0.39)	91.53%	81.27%	64.54%	70.47%	75.55%
RODNet (Full) [34]	Overall	0.31 (±0.25)	95.93%	88.86%	85.98%	87.86%	87.82%
	Parking Lot	0.27 (±0.21)	98.49%	97.98%	95.79%	96.85%	94.62%
	Campus Road	0.36 (±0.26)	92.08%	69.40%	57.06%	62.08%	73.62%
	City Street	0.49 (±0.37)	91.59%	76.37%	62.83%	70.41%	74.65%

Table 4. Performance evaluation using our proposed scoring metrics for a radar-only object detection method (RODNet) on the ***testing*** set under different driving scenarios.

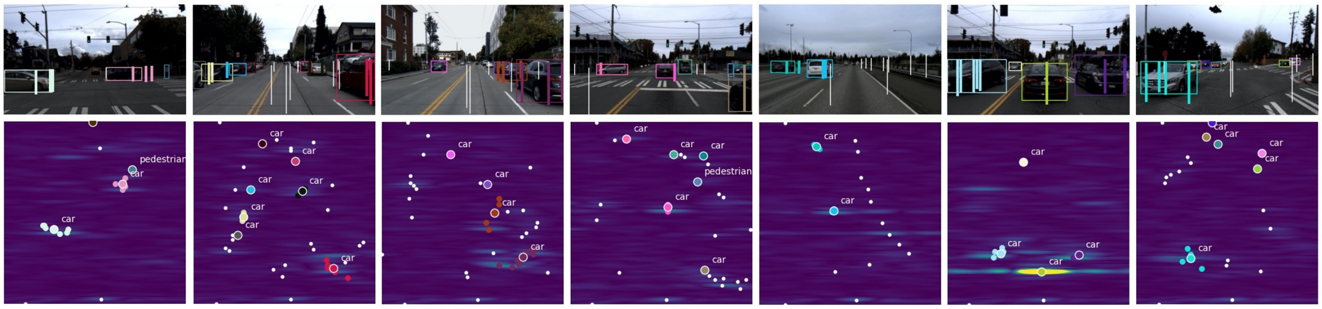


Figure 7. Qualitative results for our proposed annotation system in various driving scenarios. The upper row shows the RGB images with the detected bounding boxes from Mask R-CNN and the projected CFAR detections (vertical lines). The lower row shows the RF images with the CFAR detections (dots) and the final object annotations. The colors of the detections illustrate the detection alignment results, whereas the outliers, i.e., background detections, are presented using the white lines and dots in RGB and RF images, respectively.

Methods	MAE	Precision	Recall	**DQF1**
CO [32]	1.21 (±1.05)	81.10%	96.11%	55.16%
CRF [33]	0.68 (±0.72)	93.02%	70.11%	64.14%
Ours	0.72 (±0.78)	90.57%	95.35%	**70.36%**

Table 5. Performance evaluation of different annotation methods on the ***selected training*** set (campus road and city street).

Window t	MAE	Precision	Recall	**DQF1**
1	0.69 (±0.77)	90.78%	91.03%	65.39%
5	0.71 (±0.79)	90.70%	92.88%	66.22%
10	0.72 (±0.80)	91.06%	93.47%	67.33%
50	0.72 (±0.78)	90.57%	95.35%	**70.36%**
100	0.73 (±0.79)	90.34%	95.87%	70.35%

Table 6. The performance using different time window sizes evaluated on the ***selected training*** set (campus road and city street).

to choose a good time window t. The results are shown in Table 6. Besides, with larger window size t, the recall increases and the DQF1 score gradually converges. Overall, the best DQF1 score is achieved at $t = 50$.

7. Conclusion

In this paper, we proposed a novel radar object detection platform for adverse driving scenarios, including a large-scale dataset, annotation and evaluation system. This platform is potentially valuable to the autonomous driving community for the deep learning based radar semantic understanding tasks, e.g., detection, segmentation, tracking, etc. It is also an inspiration for a new autonomous vehicle solution using a camera-radar sensor system for all-weather conditions.

Acknowledgement

This research work was partially supported by CMMB Vision – UWECE Center on Satellite Multimedia and Connected Vehicles. The authors would also like to thank the colleagues and students in Information Processing Lab (IPL) at the University of Washington for their help and assistance on the dataset collection, processing, and annotation works.

References

[1] Flir systems. https://www.flir.com/. 3

[2] Texas instruments. http://www.ti.com/. 3

[3] Apollo scape dataset. http://apolloscape.auto/, 2018. 1, 3

[4] Waymo open dataset: An autonomous driving dataset. https://www.waymo.com/open, 2019. 1, 3

[5] A. Angelov, A. Robertson, R. Murray-Smith, and F. Fioranelli. Practical classification of different moving targets using automotive radar and deep neural networks. *IET Radar, Sonar Navigation*, 12(10):1082–1089, 2018. 2

[6] Junaid Ahmed Ansari, Sarthak Sharma, Anshuman Majumdar, J Krishna Murthy, and K Madhava Krishna. The earth ain't flat: Monocular reconstruction of vehicles on steep and graded roads from a moving camera. In *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 8404–8410. IEEE, 2018. 3

[7] Dan Barnes, Matthew Gadd, Paul Murcutt, Paul Newman, and Ingmar Posner. The oxford radar robotcar dataset: A radar extension to the oxford robotcar dataset. In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, Paris, 2020. 3, 4, 7

[8] Holger Caesar, Varun Bankiti, Alex H. Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuscenes: A multimodal dataset for autonomous driving. *arXiv preprint arXiv:1903.11027*, 2019. 1, 2, 3, 4, 7

[9] Zhaowei Cai and Nuno Vasconcelos. Cascade r-cnn: Delving into high quality object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6154–6162, 2018. 3

[10] Xu Dong, Pengluo Wang, Pengyue Zhang, and Langechuan Liu. Probabilistic oriented object detection in automotive radar. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 102–103, 2020. 2, 3, 4

[11] Kaiwen Duan, Song Bai, Lingxi Xie, Honggang Qi, Qingming Huang, and Qi Tian. Centernet: Keypoint triplets for object detection. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 6569–6578, 2019. 3

[12] Di Feng, Christian Haase-Schütz, Lars Rosenbaum, Heinz Hertlein, Claudius Glaeser, Fabian Timm, Werner Wiesbeck, and Klaus Dietmayer. Deep multi-modal object detection and semantic segmentation for autonomous driving: Datasets, methods, and challenges. *IEEE Transactions on Intelligent Transportation Systems*, 2020. 2

[13] Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun. Vision meets robotics: The kitti dataset. *The International Journal of Robotics Research*, 32(11):1231–1237, 2013. 1, 3

[14] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017. 3, 5

[15] S. Heuel and H. Rohling. Two-stage pedestrian classification in automotive radar systems. In *2011 12th International Radar Symposium (IRS)*, pages 477–484, Sep. 2011. 2

[16] Hung-Min Hsu, Yizhou Wang, and Jenq-Neng Hwang. Traffic-aware multi-camera tracking of vehicles based on reid and camera link model. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 964–972, 2020. 1

[17] Jesse Levinson, Jake Askeland, Jan Becker, Jennifer Dolson, David Held, Soeren Kammel, J Zico Kolter, Dirk Langer, Oliver Pink, Vaughan Pratt, et al. Towards fully autonomous driving: Systems and algorithms. In *2011 IEEE Intelligent Vehicles Symposium (IV)*, pages 163–168. IEEE, 2011. 1

[18] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017. 3

[19] Bence Major, Daniel Fontijne, Amin Ansari, Ravi Teja Sukhavasi, Radhika Gowaikar, Michael Hamilton, Sean Lee, Slawomir Grzechnik, and Sundar Subramanian. Vehicle detection with automotive radar using deep learning on range-azimuth-doppler tensors. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, 2019. 2, 3, 4

[20] Michael Meyer and Georg Kuschk. Automotive radar dataset for deep learning based 3d object detection. In *2019 16th European Radar Conference (EuRAD)*, pages 129–132. IEEE, 2019. 3, 4

[21] Arsalan Mousavian, Dragomir Anguelov, John Flynn, and Jana Kosecka. 3d bounding box estimation using deep learning and geometry. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7074–7082, 2017. 3

[22] J Krishna Murthy, GV Sai Krishna, Falak Chhaya, and K Madhava Krishna. Reconstructing vehicles from a single image: Shape priors for road scene understanding. In *2017 IEEE International Conference on Robotics and Automation (ICRA)*, pages 724–731. IEEE, 2017. 3

[23] Arthur Ouaknine, Alasdair Newson, Julien Rebut, Florence Tupin, and Patrick Pérez. Carrada dataset: Camera and automotive radar with range-angle-doppler annotations. *arXiv preprint arXiv:2005.01456*, 2020. 3, 4, 7

[24] Andras Palffy, Jiaao Dong, Julian FP Kooij, and Dariu M Gavrila. Cnn based road user detection using the 3d radar cube. *IEEE Robotics and Automation Letters*, 5(2):1263–1270, 2020. 3, 4

[25] Joseph Redmon and Ali Farhadi. Yolov3: An incremental improvement. *arXiv preprint arXiv:1804.02767*, 2018. 3

[26] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, pages 91–99, 2015. 1, 3

[27] Mark A Richards. *Fundamentals of radar signal processing*. Tata McGraw-Hill Education, 2005. 2, 5

[28] Erich Schubert, Jörg Sander, Martin Ester, Hans Peter Kriegel, and Xiaowei Xu. Dbscan revisited, revisited: why and how you should (still) use dbscan. *ACM Transactions on Database Systems (TODS)*, 42(3):1–21, 2017. 5

[29] Marcel Sheeny, Emanuele De Pellegrin, Saptarshi Mukherjee, Alireza Ahrabian, Sen Wang, and Andrew Wallace. Radiate: A radar dataset for automotive perception. *arXiv preprint arXiv:2010.09076*, 2020. 3, 4, 7

[30] Shiyu Song and Manmohan Chandraker. Joint sfm and detection cues for monocular 3d localization in road scenes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3734–3742, 2015. 3

[31] Gaoang Wang, Yizhou Wang, Haotian Zhang, Renshu Gu, and Jenq-Neng Hwang. Exploit the connectivity: Multi-object tracking with trackletnet. In *Proceedings of the 27th ACM International Conference on Multimedia*, pages 482–490, 2019. 1

[32] Yizhou Wang, Yen-Ting Huang, and Jenq-Neng Hwang. Monocular visual object 3d localization in road scenes. In *Proceedings of the 27th ACM International Conference on Multimedia*, pages 917–925. ACM, 2019. 3, 7, 8

[33] Yizhou Wang, Zhongyu Jiang, Xiangyu Gao, Jenq-Neng Hwang, Guanbin Xing, and Hui Liu. Rodnet: Radar object detection using cross-modal supervision. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 504–513, 2021. 2, 3, 7, 8

[34] Yizhou Wang, Zhongyu Jiang, Yudong Li, Jenq-Neng Hwang, Guanbin Xing, and Hui Liu. 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. 3, 7, 8

[35] Hao Yang, Chenxi Liu, Meixin Zhu, Xuegang Ban, and Yinhai Wang. How fast you will drive? predicting speed of customized paths by deep neural network. *IEEE Transactions on Intelligent Transportation Systems*, 2021. 1

[36] Hao Yang, Chenxi Liu, Meixin Zhu, Wei Sun, and Yinhai Wang. Hybrid data-fusion model for short-term road hazardous segments identification based on the acceleration and deceleration information. In *International Conference on Transportation and Development 2020*, pages 313–326. American Society of Civil Engineers Reston, VA, 2020. 1

[37] Hao Frank Yang. *Novel Traffic Sensing Using Multi-Camera Car Tracking and Re-Identification (MCCTRI)*. PhD thesis, 2020. 1