RaDICaL: A Synchronized FMCW Radar, Depth, IMU and RGB Camera Data Dataset With Low-Level FMCW Radar Signals

Teck-Yian Lim, Spencer A. Markowitz, and Minh N. Do

Abstract*—Within the autonomous driving community, millimeter-wave frequency-modulated continuous-wave (FMCW) radars are not used to their fullest potential. Classical, hand-designed target detection algorithms are applied in the signal processing chain and the rich contextual information is discarded. This early discarding of information limits what can be applied in algorithms further downstream. In contrast with object detection in camera images, radar has thus been unable to benefit fully from data-driven methods. This work seeks to bridge this gap by providing the community with a diverse, minimally processed FMCW radar dataset that is not only RGB-D (color and depth) aligned but also synchronized with inertial measurement unit (IMU) measurements in the presence of ego-motion. Moreover, having time-synchronized measurements allow for verification, automated or assisted labelling of the radar data, and opens the door for novel methods of fusing the data from a variety of sensors. We present a system that could be built with accessible, off-the-shelf components within a $1000 budget and an accompanying dataset consisting of diverse scenes spanning indoor, urban and highway driving. Finally, we demonstrated the ability to go beyond classical radar object detection with our dataset with a classification accuracy of 85.1% using the low-level radar signals captured by our system, supporting our argument that there is value in retaining the information discarded by current radar pipelines.

***Index Terms*—Radar, FMCW, sensor-fusion, autonomous driving, dataset, RGB-D, object detection, odometry.**

I. Introduction

In COMPARISON to visible light and the lasers used by lidar systems, millimeter-wave (mmWave) FMCW radars use wavelengths that are much larger than fog, dust, and other particles present in adverse driving conditions that limit visibility. This longer wavelength allows the radar signals to easily penetrate or diffract around such particles, allowing mmWave radars to function as a robust, all-weather sensor [1]–[4].

While recent published works in autonomous driving attempt to incorporate radars, the input from the radar consists only of points with velocity, retaining little information from the raw measurements [5]–[8]. In these sources, we see methods to increase the number of points such as integrating over time and using inputs from multiple sensors. In contrast, lidar provides a much denser point cloud than radars and thus see more use in sensor fusion works.

The use of radars, however, should not be limited by these sparse point cloud. The sparse points returned from the commercial radar packages are the results of statistical object detection algorithms (CFAR) [9], of which the goal is to detect strong radio reflectors in the scene, with no intention of capturing the semantic meaning of the objects. As a result, the rich information of the reflected radar signal is discarded. Therefore, current published work within the autonomous driving community often does not exploit the capabilities of radar to its fullest potential. In our work, we seek to remove this limitation, furthermore, within a very small budget of $1000.

In particular, our contributions include:

1) Modularized RGB-D-Radar Architecture: We present the modularized and expandable design of a raw frequency-modulated continuous wave (FMCW) + RGB-D (color and depth) system, ready to be integrated into robotics projects. Our system is built using simple, off-the-shelf components that cost less than $1000 to assemble. Our system is also designed with the purpose of functioning as an additional module for existing autonomous driving data-collection platform that may include lidars and other sensors.

2) Radar-Camera Alignment Method: We present an approach to automatically capture and label complex millimeter wave radar signatures with the help of a calibrated RGB-D camera.

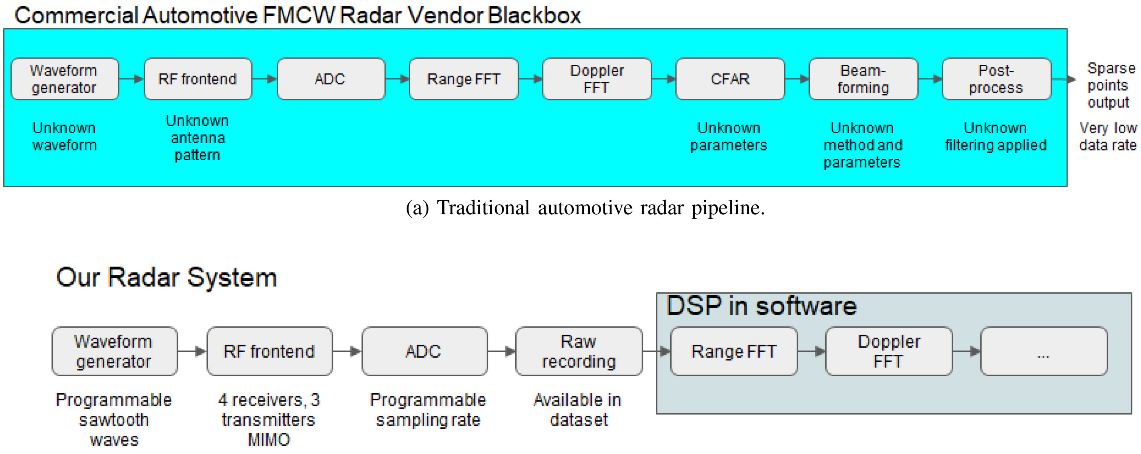
3) Indoor and Street Scenes Dataset: As part of our work, we publicly share[1] a novel dataset consisting of both indoor and outdoor scenes. The indoor scenes include a variety of rooms with different quantities of people and static clutter. The outdoor scenes contain people walking and running together with vehicles driving along a suburban street as well as scenes recorded from a moving vehicle in a myriad of environments. For the scenes recorded from a moving car, inertial measurement

Manuscript received September 4, 2020; revised January 5, 2021; accepted February 6, 2021. Date of publication February 23, 2021; date of current version June 3, 2021. The guest editor coordinating the review of this manuscript and approving it for publication was Dr. Philipp Heidenreich. *(Teck-Yian Lim and Spencer A. Markowitz contributed equally to this work.) (Corresponding author: Teck Yian Lim.)*

The authors are with the University of Illinois at Urbana-Champaign, Champaign, IL 61801-3028 USA (e-mail: tlim11@illinois.edu; samarko2@illinois.edu; minhdo@illinois.edu).

Digital Object Identifier 10.1109/JSTSP.2021.3061270

[1] All data used in this publication, including preprocessing code and associated documentation, is available at https://doi.org/10.13012/B2IDB-3289560_V1 Code snippets for using the dataset and code documentation can be found on the authors' website.



(a) Traditional automotive radar pipeline.

(b) Our raw radar pipeline.

Fig. 1. A comparison of commercially available automotive FMCW radar systems with our raw radar dataset. We record the unprocessed ADC readings from the antennas and differ processing to a later stage. With off-the-shelf components, we can achieve a data rate of up to 325 Mbps for the radar alone. This is in contrast with the heavily processed and sparse point targets of commercially available radar packages. Furthermore, commercially available radar packages are often black-boxes that contain non-trivial optimizations to radar waveform and antenna pattern design.

unit (IMU) data is also provided. To the best of our knowledge, there are no other public datasets that offer raw radar measurements, let alone aligned RGB-D images, IMU measurements, and projected object labels generated with the help of recent advances in RGB image object detection.

4) Rich Object Detection With Radar: Finally, we show the capabilities of our dataset in two applications. One for semantic object detection, in contrast with classical radar object detection, and micro-Doppler exploitation assisted by RGB-D pose estimation.

A. *Paper Organization*

The rest of our paper is organized as follows, Section II covers recent works involving FMCW radars and their related use in autonomous driving. Section III provides a brief overview of FMCW radar signal processing, its classical signal processing chain and our proposed changes to it to enable better exploitation by deep neural nets. Section IV describes our system architecture, capabilities and limitations. Section V describes the scenarios and configuration in which we operated our system. Finally, in Section VI, we demonstrate the capability of our dataset and system with high resolution velocity estimation and temporal synchronization task and a small object radar signature classification task.

II. Related Works

Autonomous driving application is the impetus behind sensor fusion research. In a recently published dataset, nuScenes [10], automotive radar is listed as one of the available sensors. However, the radar processing pipeline (Fig. 1) discards all of the semantic information and only provides sparse point clouds. In PointPillars [11], the authors proposed a method to convert point clouds into pseudo-images, allowing the use of convolutional neural networks for object detection. However, due to the sparsity of the points that commercial radar packages return, with some manufactureres even having an upper limit of 64 points [12], [13], such an approach is unlikely to provide quality detection results. While there seems to be a significant amount of research that attempts to utilize radar information, the sparsity of information available from existing radar setups greatly constrains its effectiveness in early stage processing, forcing many systems to incorporate radar in later stage post-processing.

In two much more recent datasets [14], [15], lower level radar signals are available. The radar however, differs significantly from the solid state radars typically found in vehicle systems. This radar functions similarly to a lidar and is mechanically spun at 4 Hz and 400 angle bins are sampled per revolution [16]. Only range-magnitude measurements are available at each azimuth and the entire field-of-view is not observed simultaneously. A row in a frame in this dataset is equivalent to 1 transmitter and 1 receiver using a single chirp and after taking the magnitude of the range FFT in our radar setup (mathematical details are provided in Sec. III-A).

The closest related work is RF-Pose [17] and FusionNet [18], where minimally processed radar signals are used to produce object detections with the help of a deep neural net. Authors of RF-Pose [17] demonstrated empirically the possibilities of minimally processed FMCW radars when coupled with today's advances in deep learning. Using a 16x16 2D array of antennas, they demonstrated the capability of predicting a human's pose with a customized FMCW radar system configured for short range detection. The radar used by the authors had a wavelength of approximately 5 cm, thus falling below the Rayleigh criterion [19], [20] for humans, causing the reflection to be highly specular. Whereas in our work and in [18], a wavelength of approximately 3.8 mm is used. This enables better spatial resolution and lower specularity at the cost of poorer penetration through thick building materials. In comparison with FusionNet [18], we include depth information with an RGB-D sensor, enabling an accurate projection of the camera image into world coordinates, rather than relying on a planar road assumption. Lastly, and most importantly, our system is easily reproducible, using off-the-shelf components and a widely available commercial radar platform [21], with a total cost below $1000.

III. Primer on FMCW Radar Signal Processing

There exists extensive literature about radar signal processing since its invention the 1930 s. We do not intend for this section, nor is it possible, to be comprehensive, but we hope to cover the basics in sufficient detail to enable the effective use of our novel dataset. We limit our discussions mainly to the type and characteristics of our radar system.

Radars work on a simple idea: send out a radio signal and wait for an echo. The time it takes for the echo to arrive is directly proportional to the distance of the reflecting object. A simple manifestation of this concept is a *pulse radar*. Transmission occurs for an instant, followed by a period of waiting for echoes. Mathematically, the transmitted signal at any instant, t, can be defined as:

$$S_{tx} = A_{tx}(t)\cos(2\pi f_c t + \phi_0). \tag{1}$$

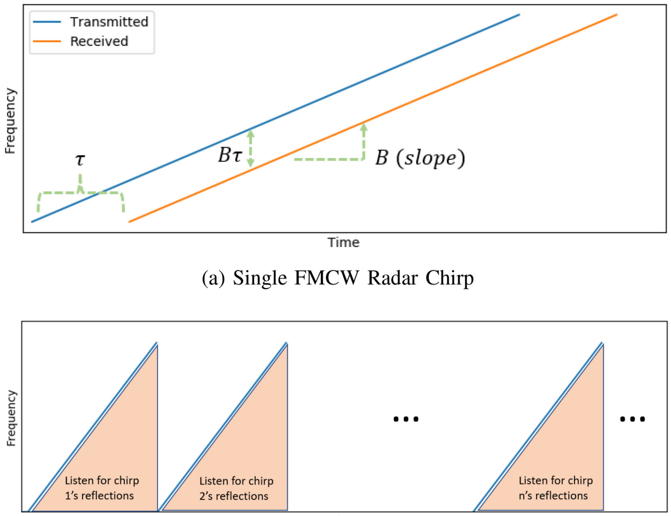
Where A_{tx} is a constant transmit amplitude when the radar is transmitting and zero otherwise, f_c is the transmission frequency, and ϕ_0 is the starting phase. Without loss of generality, we can assume that the starting phase is 0, and we will drop the term for clarity of notation, and only reintroduce it when its value is no longer negligible.

In addition to being able to estimate a target's range from its reflection's time delay, its velocity can be determined from the frequency shift of its reflection due to a phenomenon known as the Doppler effect. Because the transmit frequency is constant, a target with no radial velocity will reflect the signal at the same frequency that was transmitted, while a moving target will induce some measurable Doppler shift of that frequency. Although simple in terms of operating principles, due to the speed of light, pulse radars are blind at short ranges (below 1 km). While not an issue for long-range applications (e.g. aircraft, ships), this makes them of limited use where the range is small.

Consequently, for small working ranges in automotive and small robot applications, pulse radars cannot be used. For these applications, Frequency Modulated Continuous Wave (FMCW) radars are a better fit as they allow for very short working ranges. Similar to the pulse radar, we receive a time-delayed and Doppler-shifted version of the transmitted signal. In contrast to the pulse radar, however, both the transmitter and receiver are on simultaneously, mitigating the blindness at very short ranges. Additionally, FMCW radars transmit a signal, often referred to as a chirp, whose frequency changes with time:

$$S_{tx}(t) = A_{tx}(t)\cos(2\pi(f_c + f_\tau(t))t). \tag{2}$$

Where f_c is the starting frequency, and $f_\tau(t)$ is a function describing how the frequency changes over time. One possible waveform for a single chirp is a sawtooth wave (in frequency-time), with one period as:


Transmitted
Received
Frequency
Time
τ
Bτ
B (slope)
Frequency
Listen for chirp 1's reflections
Listen for chirp 2's reflections
...
Listen for chirp n's reflections
...
Time

(a) Single FMCW Radar Chirp

(b) Consecutive FMCW Radar Chirps

Fig. 2. For the scenario with one target, the reflected waveform is a time delayed version of the transmitted signal. With prior knowledge of the slope B, the time delay τ can be deduced from the frequency of the low-pass-filtered signal.

$$S_{tx}(t) = A_{tx}(t)\cos\left(2\pi\left(f_c t + \frac{B}{2}t^2\right)\right), \quad \text{for } 0 \leq t < T \tag{3}$$

Where B is the slope of the rate of change in frequency. For the rest of the discussion, we assume that we are working with a sawtooth wave.

A. Estimating Range With FMCW Radars

The reflected waveform is a delayed version of the transmitted wave as shown in Fig. 2(a). Again, by measuring this delay, denoted τ, we can compute the radial distance of the object from the radar. At the receiver, a mixer (multiplier) mixes the reflected signal with the transmitted signal. Next, this signal passes through a low-pass filter and is sampled by an ADC. At any instant, we can describe the signal as:

$$S_{rx}(t) = A_{rx}(t)\cos(\alpha t)\cos(\beta t) \tag{4}$$

Where, A_{rx} is the received amplitude, α is the frequency that is being transmitted and β is that of the reflected signal. Using the product to sum identity, we can see that:

$$S_{rx}(t) = \frac{A_{rx}(t)}{2}(\cos(\alpha - \beta)t + \cos(\alpha + \beta)t) \tag{5}$$

In this form, we see that there are two frequency components in the received signal, one of much lower frequency than the transmitted waveform and one of very high frequency. After low-pass filtering, we are left with a low frequency signal which demands a far lower performance ADC than what the original GHz-band signal would have required.

Since the slope is known, we can determine the distance by leveraging its relationship to the time delay, slope, and frequency:

$$\tau = \frac{2\,d}{c_0}, f = B\tau \tag{6}$$

$$d = \frac{c_0 \cdot f}{2B}, \tag{7}$$

where c_0 is the speed of light in free space.

Since the mixed signal gives us a frequency difference, all we have to do is perform an FFT over the entire chirp, and the (frequency) location of the (amplitude) peak is directly proportional to the range of the target. In FMCW radar literature, this is often referred to as the "intermediate frequency," "beat frequency" or the IF signal.

B. Estimating Doppler

With a sawtooth wave, there is no way to disentangle frequency shifts that are due to a non-zero relative velocity. It is treated as measurement noise for low-velocity targets. If this is not the case, a different waveform might be a more suitable choice, such as a triangular waveform as in [22].

While we are unable to resolve the velocity of a target from a single chirp, if we look across multiple chirps as depicted in 2(b), the relative velocity can be recovered. Recall that we are assuming that the velocity of the target is small, and its range does not change significantly over several chirps. Numerically this results in FFTs with peaks at the same frequency bin. While unable to be resolved as different distances, this small displacement manifests as a phase shift.

Suppose two chirps are sent T_c seconds (usually in the order of microseconds) apart. Recall that the IF signal is a sinusoid:

$$A_{rx}(t)\cos(2\pi f t + \phi_0). \tag{8}$$

If the object is stationary, the phase term of the first chirp will be identical to that of the second chirp. However, if there is a small, non-zero relative velocity, this slight change in distance will result in a phase delay between the closely spaced chirps. Using a typical configuration of $f_0 = 77\,\mathrm{GHz}$, with a slope of $B = 30\,\mathrm{MHz}/\mu\mathrm{s}$ and $T_c = 40\,\mu\mathrm{s}$ between chirps, a vehicle traveling at speeds of $v = 18\,\mathrm{m/s}$ (40 mph) will be displaced by $\Delta d = 0.72$ mm. This displacement is smaller than the wavelength and this will manifest as a phase change of

$$\Delta\phi = \frac{2\pi \cdot 2\Delta d}{\lambda}, \tag{9}$$

where the factor of 2 in front of Δd accounts for the effective change in radar wave traveling distance through a round trip.

Rearranging and dividing by the time between chirps, T_c, we obtain the relationship between the phase difference and the velocity of the target:

$$v = \frac{\lambda\Delta\phi}{4\pi T_c} \tag{10}$$

Velocities that result from phase shifts of greater than $\pm\pi$ will be aliased, or could also result in range bin migration. A workaround for such situations is discussed in Section III-D.

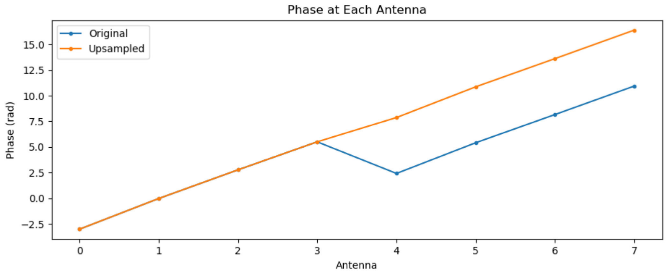


Fig. 3. Using [26] with two transmitting and four receiving antennas (creating a total of 8 virtual antennas), we see that the discontinuity in the phase due to TDM-MIMO can be compensated for.

Numerically, the phase difference can be obtained by performing an FFT across chirps. The number of chirps and the period between the chirps determines the velocity resolution.

In a practical FMCW radar system, N chirps are sent and processed as a group in order to determine the velocity of the target. We call this sequence of N chirps a frame, also commonly referred to as the coherent processing interval (CPI), and this is the basic unit of FMCW radar signal just as an image is the basic unit of a camera.

C. Estimating Angle of Arrival

Finally, with multiple receiving antennas, we can estimate angle of arrivals using the same principle as Doppler estimation as mentioned in the preceding section. However, instead of using multiple chirps in time, we compute the FFT across multiple antennas. In practice, especially for a small linear array, using the FFT directly results in a very low resolution and noisy range-azimuth heatmap. More sophisticated beamforming algorithms, *e.g.* MVDR [23] and MUSIC [24], can instead be applied but the details are beyond the scope of this discussion.

D. Mimo

Since angular resolution is related to the spatial diversity of the receiving antenna array, it is advantageous to have as many receiving antennas as possible. However, space and computation often constrain the number of receiving antennas too heavily to achieve fine angular resolution. One solution to this is time division multiplexing (TDM) which leverages multiple transmitting antennas along with a uniform linear array (ULA) of receiving antennas. By transmitting identical chirps successively from two adjacent antennas and approximating the transmit times as the same, it is possible to create instances of *virtual antennas* and subsequently increase angular resolution. This method is thoroughly described in [25].

Making the assumption that successive chirps are transmitted simultaneously can cause errors in the phase when a detected object has a nonzero velocity as seen in Fig. 3. This stems from the motion of the target that occurs between the two transmission times which can cause a discontinuity in the phase. One easy way to correct for this is described in [26] and suggests that before taking the FFT along the Doppler axis, one should upsample the radar frame along that axis such that the two different transmitters alternate. For example, for two transmit antennas $Tx_{0:1}$ and four receiving antennas $Rx_{0:3}$, one can upsample as follows:

$$\begin{bmatrix} Tx_0Rx_{0:3} & \mathbf{0} \\ \mathbf{0} & Tx_1Rx_{0:3} \\ Tx_0Rx_{0:3} & \mathbf{0} \\ \mathbf{0} & Tx_1Rx_{0:3} \\ \vdots & \vdots \end{bmatrix} \tag{11}$$

where $Tx_nRx_{0:3}$ is the data from all four antennas associated with the nth transmitter.

Another undesirable result of TDM-MIMO is that if each frame is constrained to N chirps, increasing the number of transmitting antennas reduces the number of chirps per transmitting antennas which in turn reduces the maximum unambiguous velocity. One can overcome this reduction by observing if there is a still a discontinuity in the phase due to TDM-MIMO even after the phase correction described above. As outlined in [27], for a system with two transmitting antennas, there is a residual phase jump of $\pm\pi$ if the detected velocity is actually within $[-2f_{D,\text{Max}}, -f_{D,\text{Max}}]$ or $[f_{D,\text{Max}}, 2f_{D,\text{Max}}]$, respectively, where $f_{D,\text{Max}}$ is the maximum unambiguous velocity.

E. Suggested Further Reading

Many aspects of FMCW radar signal processing are beyond the scope of this paper. The topics covered in this section pertain to our radar antenna configuration, MIMO mode, and waveform selection. We encourage readers to refer to other literature such as [28] in order to gain a deeper understanding of FMCW radars or radars in general.

Open implementations of the methods discussed in this section are available in OpenRadar [29], a library that we used heavily in our work.

IV. Modularized Hardware and Software

A. Sensors Overview

Our setup consists of an RGB-D camera and a 4-Rx 3-Tx 77 GHz mmWave radar as photographed in Fig. 4. While the radar in our setup has 3 transmitters, with one suitable for elevation estimation, we did not enable the elevation transmitter in this dataset so as to improve our maximum unambiguous velocity estimation and velocity resolution. The RGB-D sensor module is an Intel RealSense D435i that includes an IMU.

B. System Architecture

Our data collection system is implemented on top of the Robot Operating System (ROS) [30], so as to allow for integration into ROS based autonomous driving systems.

Our radar consists of 2 hardware components, the radar front-end with a Texas Instruments single chip radar and the data acquisition card that streams the radar measurements to a computer over Ethernet. In a typical setup, the single chip radar acquires the raw radar signals and processes it with the traditional FMCW radar pipeline as described in the previous section, which results in sparse points for objects with statistically significant returns. To retain the raw signals, we run a bare minimum real-time firmware on the radar chip that does not perform any signal processing. This results in a far higher data rate than a typical setup and requires us to transfer the data over a high-bandwidth Ethernet connection which is provided by the data acquisition board. Different physical requirements (i.e. max range, range resolution, max Doppler, Doppler resolution, frame rate, etc.) can result in different data rates. While our selection of hardware components allow a maximum of 600 Mbps, we found that 375 Mbps is a more manageable rate. Our datasets work within these limits to ensure minimal packets are lost by the data recording host. The RGB-D camera is controlled by the Realsense SDK. We use it as is.

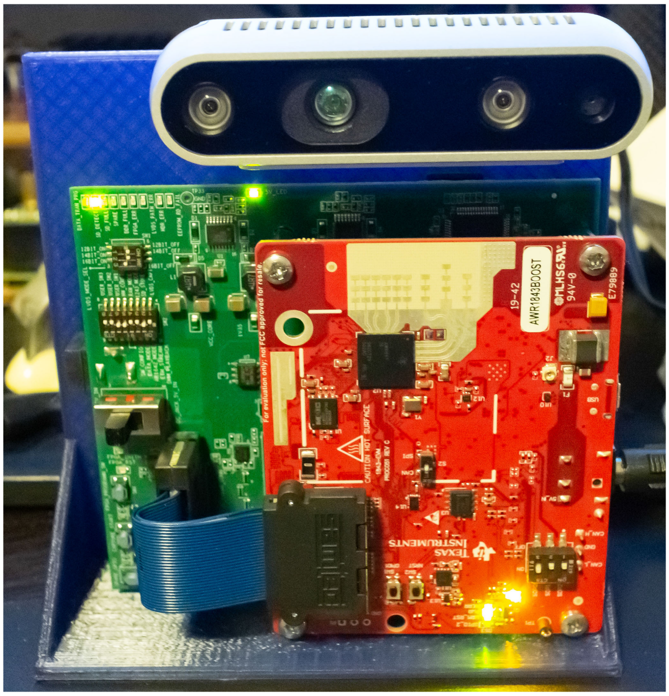


Fig. 4. Our FMCW radar with a RealSense D435 RGB-D camera mounted above it. The RealSense camera provides 1280 × 720 RGB images at 30 fps, an aligned depth map at the same resolution. Optionally, we can also record stereo near-infrared images at 640 × 480, however this is not included in our published dataset. The FMCW radar consists of 3 transmitting antennas and 4 receiving antennas. The receiving antennas are spaced $\lambda/2$ wavelength apart and the transmitters are spaced 2λ apart. This provides us with a virtual array of 8 antennas.

C. Temporal Alignment Between the Radar and RGB-D Camera

Depending on the radar configuration, the radar and the camera may run at a different frame rate. Instead of an external clock trigger, we allow the sensors to be triggered independently by their own internal clocks. We maintain high resolution time-stamps of each data unit received from the individual sensors on the recording system. To this end, we implemented a custom firmware on the radar that does not perform any radar signal processing, but send out the raw signals immediately after

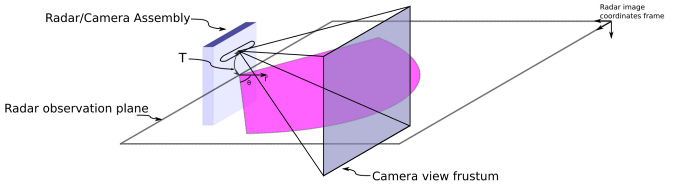


Fig. 5. Coordinate frames of radar and camera. For spatial calibration of the sensors, we seek to estimate the rotation and translation, T, between the camera frame and the radar frame.

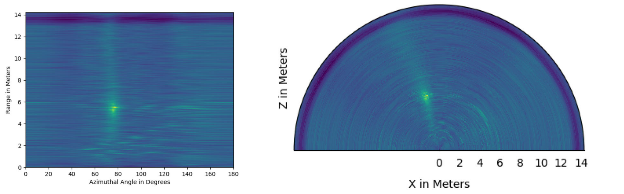


Fig. 6. Polar radar 'image' to Cartesian transform. After range processing and beamforming, we obtain a polar image(right). In the calibration process, we transform this polar image into Cartesian coordinates. This image is the bird's-eye view of the scene in front of the radar and pixel coordinates is directly proportional to the physical distances of reflectors in the scene.

sampling. As for the RealSense camera, high resolution timestamps were readily available in the library provided.

D. Sensor Spatial Calibration

Calibration, the spatial transformation between the camera image and the radar observation, is largely similar to that of calibrating the transforms between multiple camera views, with some modifications, as the radar is a 'camera' with peculiar imaging properties.

Firstly, the appearance of objects in camera and in radar is significantly different. Thus, the typical approach of performing feature detection and key-point matching will fail. In order to tackle this problem, we collected calibration sequences consisting of objects that are easily identifiable both in radar and in camera. Our chosen object is a radar corner reflector, the radar equivalent of a retro-reflecting mirror in the visible spectrum. This object will produce a strong and well localized return in the radar's measurement, thus easily interpretable in the radar heat-map. Visually, this is a large, silver octahedron, and we use the midpoint of this object when computing correspondence with radar observations. We use only one reflector in each frame so that we do not have to disambiguate the source of the returns. Furthermore, we collected the calibration sequence outdoors in an open space to avoid potential problems due to multi-path effects. We opted for outdoor data collection as opposed to using an RF anechoic chamber which might be prohibitively expensive for groups that do not have easy access to such facilities. Next, we set the reflector on the ground for a short period of time so as to remove the need of precise temporal calibration for spatial calibration. Finally, we assigned point correspondence manually from the camera image to the point with the strongest return in the radar image over multiple frames.

Secondly, the radar 'image' after beamforming is not a typical camera image, but an image in polar coordinates (Fig. 6). As we are observing objects on the ground and the height of our radar is relatively fixed, we do not expect variations in height to result in significant changes to radial range. Thus we model the cartesian projected radar 'image' as an orthographic projection onto the horizontal plane (xz-plane, Fig. 5). Thus we will require some modification to the typical approach used in multiple-view geometry.

In our calibration, we picked the camera frame, denoted $\mathbf{x}_c = [x \quad y \quad 1]^T$, as the reference frame *i.e.*:

$$\lambda_c \mathbf{x}_c = \Lambda_c [\mathbf{I}|\mathbf{0}]\mathbf{W}, \tag{12}$$

where λ_c is a normalizer, Λ_c is the camera's intrinsic matrix, $\mathbf{W}$ the homogenous world coordinates, and $[\mathbf{I}|\mathbf{0}]$ is the camera's extrinsic matrix with the identity matrix and the zero vector as its rotation and translation respectively. On the contrary, we model the radar as having a rotation $\mathbf{\Omega}$ and translation $\mathbf{T}$ as in:

$$\mathbf{x}_r = \Lambda_r [\mathbf{\Omega}|\mathbf{T}]\mathbf{W}, \tag{13}$$

where $\mathbf{x}_r$ is the pixel coordinates in the Cartesian radar heat-map, and Λ_r is the intrinsic matrix of the radar which can be expressed as:

$$\Lambda_r = r \begin{bmatrix} 1 & 0 & 0 & -1 \\ 0 & 0 & 1 & 0 \\ 0 & -1 & 0 & 1 \end{bmatrix} \tag{14}$$

Here, $r = r_{\max}/r_{res}$ is a scale factor capturing the max range, $r_{\max}$, and range resolution, r_{res}, of the radar configuration in use. With this projection, we lose the y dimension of the world coordinates and x and z are scaled according to the range of the radar. Note that in contrast with perspective projection, we simply drop the 3 rd coordinate for orthographic projection. Furthermore, by setting both $\mathbf{x}_c$ and $\mathbf{x}_r$ as functions of $\mathbf{W}$, we are able to work directly with the projection of world coordinates into the radar frame simplifying the task of finding point correspondences.

Because the radar is unable to detect the elevation of a given target, we project the world coordinates determined from (12) onto a plane by setting the vertical component to a constant value for all $\mathbf{W}$ in (13).

With point correspondence and projection taken care of, we can proceed with the multiple-view estimation of the fundamental matrix, $\mathbf{F} = \Lambda_r [\mathbf{\Omega}|\mathbf{T}]$ by solving:

$$\begin{aligned} &\mathbf{x}_r = \mathbf{FW} \\ &\quad \iff \mathbf{x}_r \times \mathbf{FW} = 0 \\ &\quad \iff \begin{bmatrix} 0 & -\mathbf{W}^T & z\mathbf{W}^T \\ \mathbf{W}^T & 0 & -x\mathbf{W}^T \\ -z\mathbf{W}^T & x\mathbf{W}^T & 0 \end{bmatrix} \begin{pmatrix} \mathbf{F}_1 \\ \mathbf{F}_2 \\ \mathbf{F}_3 \end{pmatrix} = 0 \end{aligned} \tag{15}$$

Since a single 2D/3D correspondence described in (15) only has two linearly independent equations, at least 6 correspondences are needed for a minimal solution. To determine such a solution, one can solve the following equation using least squares:

$$\begin{bmatrix} 0^T & \mathbf{W}_1^T & -y_1\mathbf{W}_1^T \\ \mathbf{W}_1^T & 0^T & -x_1\mathbf{W}_1^T \\ \cdots & \cdots & \cdots \\ 0^T & \mathbf{W}_n^T & -y_n\mathbf{W}_n^T \\ \mathbf{W}_n^T & 0^T & -x_n\mathbf{W}_n^T \end{bmatrix} \begin{pmatrix} \mathbf{F}_1 \\ \mathbf{F}_2 \\ \mathbf{F}_3 \end{pmatrix} = 0, \text{for } n \geq 6 \tag{16}$$

For the purpose of our dataset, obtaining the least squares estimate of the fundamental matrix is sufficient. We can now project objects and points observed in the camera image to the radar image with

$$\mathbf{x}_r = \mathbf{F}\Lambda_c^{-1}\lambda_c\mathbf{x}_c \tag{17}$$

While the fundamental matrix, $\mathbf{F}$, can be decomposed into its rotation and translation components [31], [32], it is beyond the scope of our work. As the radar image is an orthographic projection, the solution to this system will be ambiguous for translations in the vertical(y) direction.

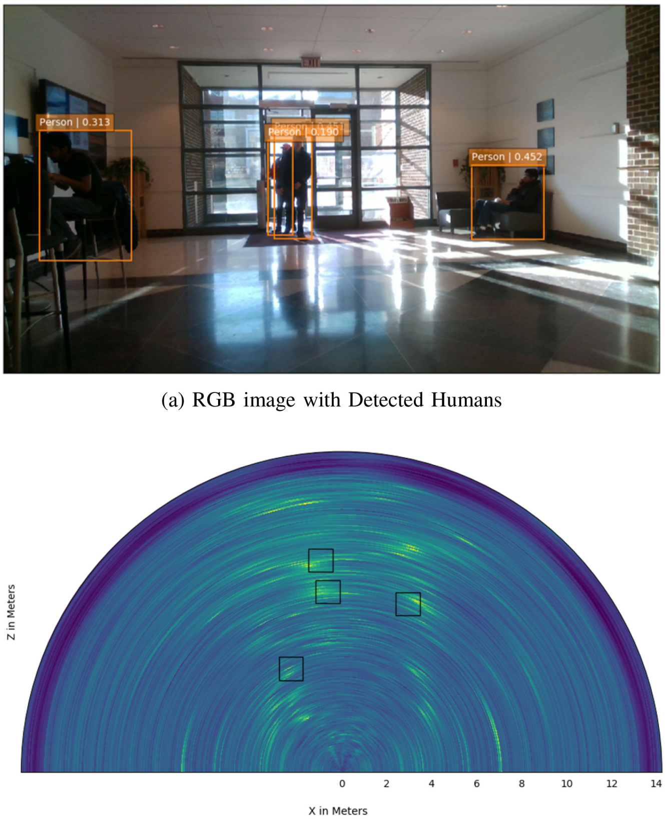
Calibration for different radar configurations: Our chosen depth camera has a maximum depth of approximately 10 m. Therefore, we perform our parameter estimation with the radar configured to match this range. The results from our estimation can be seen in Fig. 7. Here, the bounding boxes in the range-azimuth frame are positioned by first determining the location of the humans in the RGB frame using an off-the-shelf neural network. From there, we set x_c in (17) to points on the perimeter of those RGB bounding boxes and compute the locations of the projected bounding boxes in the radar frame.

Depending on the radar configuration parameters, the maximum range, range resolution, and azimuth resolution can change significantly. These changes can be computed from the radar profile and should be incorporated into the radar intrinsic matrix according to (14) when projecting points from the camera frame. Since the relative position of the camera and the radar are fixed (excluding unavoidable mechanical vibrations when moving the data collection setup) we only need to perform the calibration once for different radar configurations. Concretely, this is a scale operation on the projected points $\mathbf{x}_r$, with scale factor

$$\lambda_{r_1,r_2} = r_{2,max}/r_{1,max}, \tag{18}$$

where $r_{\cdot,max}$ is the maximum range of radar configuration 1 or 2, with radar configuration 1 being the calibrated configuration.

For depth beyond reliable range of the RGB-D camera: While scaling various elements in the radar's intrinsic matrix will suffice for targets within the 10 m range of the depth camera, to fully take advantage of the radar's range and the camera's high resolution, one must compute the location of the bounding boxes in the radar frame without the use of the depth camera. One such method is to first perform some object detection algorithm in the range-azimuth frame using an algorithm like CFAR. Then, compute the azimuthal angle of the targets in the RGB frame. This is done by first finding the difference in the horizontal coordinates of the targets, x_c^{target} from the horizontal component of the camera center, x_c^{center}. Then, using the pinhole camera model with a focal length f_x, the projected angle in the radar frame is just



(a) RGB image with Detected Humans

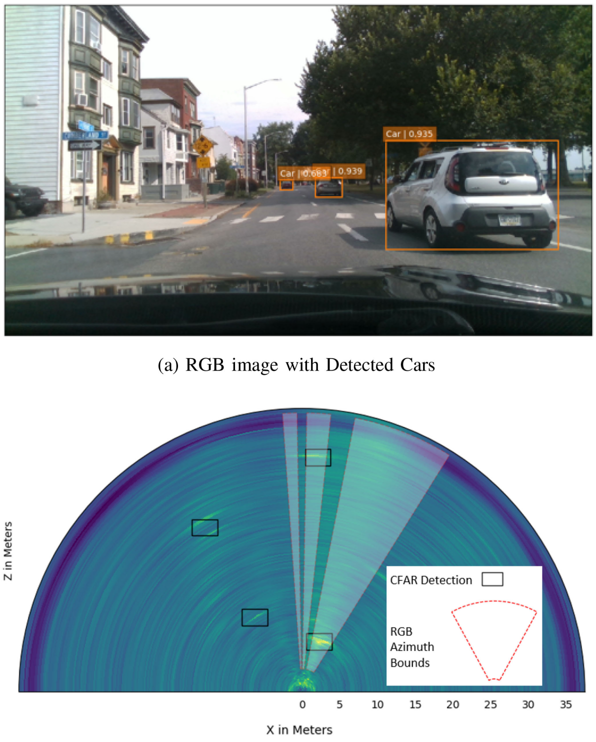
(b) Range-Azimuth Plot with Predicted Bounding Boxes

Fig. 7. In Fig. (a), we see the results of feeding an RGB image into an off-the-shelf deep neural network that detects humans. Using the calibration methodology discussed in Section IV-D and the depth measured at the center of the RGB bounding box, we are able to predict the location of those same targets within the associated range-azimuth plot in Fig. (b).

$$\theta_x = \tan^{-1}\left(\frac{x_c^{target} - x_c^{center}}{f_x}\right) \tag{19}$$

To simplify the procedure, here it is appropriate to assume the translation and rotation between the radar and camera are negligible. Subsequently, one can determine which objects detected in the range-azimuth frame correspond to which targets in the RGB frame by measuring the distance or overlap between the projected angle and the detected object in the radar. Similar to the scenario in which the depth camera could be used, we demonstrate the efficacy in Fig. 8 by first using a neural network to compute bounding boxes in the RGB frame. Using (19), we create bounds in the horizontal axis by projecting points on the box's perimeter. From there, we match the detections.

For the purpose of sanity checks, we've also provided simple scenes consisting of one to a few targets. Furthermore, we've also provided simple scenes containing a radar reflector.



(a) RGB image with Detected Cars

(b) Range-Azimuth Plot with Azimuthal Bounds and CFAR Detections

Fig. 8. In Fig. (a), we see the results of feeding an RGB image into an off-the-shelf deep neural network that detects cars. Using (19), we are able to project those bounding boxes into azimuthal bounds in the range-azimuth plot in Fig. (b). The two detected objects farthest to the right exhibit convincing overlap with those bounds. The distant car in the left lane is detected in the RGB image but is out of the detectable range of the radar configuration. The detected objects on the left of the radar frame are caused by static clutter.

E. System Constraints, Scaling Out and Real-Time Considerations

Our choice of radar hardware has the maximum capacity of streaming 600 Mbps of raw sensor data. In practice, we recommend working with lower rates (below 325 Mbps) to ensure that recording software can keep up with the amount of generated data. Together with the RGB-D images and IMU data, we generate about 12 GB of data per minute.

V. Dataset Description

Because radar heatmaps are very different from what humans see, it is very difficult for humans to generate ground truth labels without special training and only given a radar heatmap. Our dataset provides means of assisted labelling with well synchronized RGB frames. The dataset consists of a number of different radar chirp configurations that are described in Table I as well as multiple types of scenes as described in Table II.

TABLE I
Data Subsets and the Physical Properties of the Associated Radar Configuration. *The Maximum Velocity Can Be Extended Twofold Using [26] and [27].

Profile Name	Max Range (m)	Range Res (m)	Max Velocity* (m/s)	Velocity Res (m/s)	FPS	Data rate (Mbit/s)
indoor	14.24	0.047	4.86	0.30	30	77
outdoor30	37.47	0.20	15.43	0.48	30	98
outdoor60	62.45	0.97	23.02	0.36	30	65
highRes	14.24	0.047	2.78	0.043	22.2	229

A. Calibration Dataset

The calibration dataset contains scenes with few targets and minimal static clutter to cut down on multi-path. For the indoor radar chirp configuration, the scenes consist of a the radar mounted to the ground with a radar-reflector placed at many locations on the ground within the field of view in order to test a variety of angles and radial distances from the camera and radar. Calibration of the relative transforms between the camera view and radar observations was performed using this dataset. It should be noted that the maximum detectable range for the indoor radar configuration is only slightly greater than that of the depth camera. This allowed us to incorporate the depth camera in our calibration. In our own calibration experiments using hand labeled camera and radar data and (16), we were able to achieve an MSE of $.054\ \mathrm{m}^2$ for targets within 8 m. While we do provide our own calibration parameters, we share this sequence for others to reproduce our calibration or to design novel calibration methods.

For the outdoor radar configuration that can detect targets at ranges greater than 30 meters, we provide multiple scenes with human and automotive targets in environments with minimal static clutter. Although the depth camera is not able to accurately see much more than 10 meters, it can still prove useful when the targets are at closer ranges. Additionally, because the environment is largely empty, it will be relatively easy to associate targets between the camera images and the range-azimuth plots at longer ranges.

B. Indoor Scenes

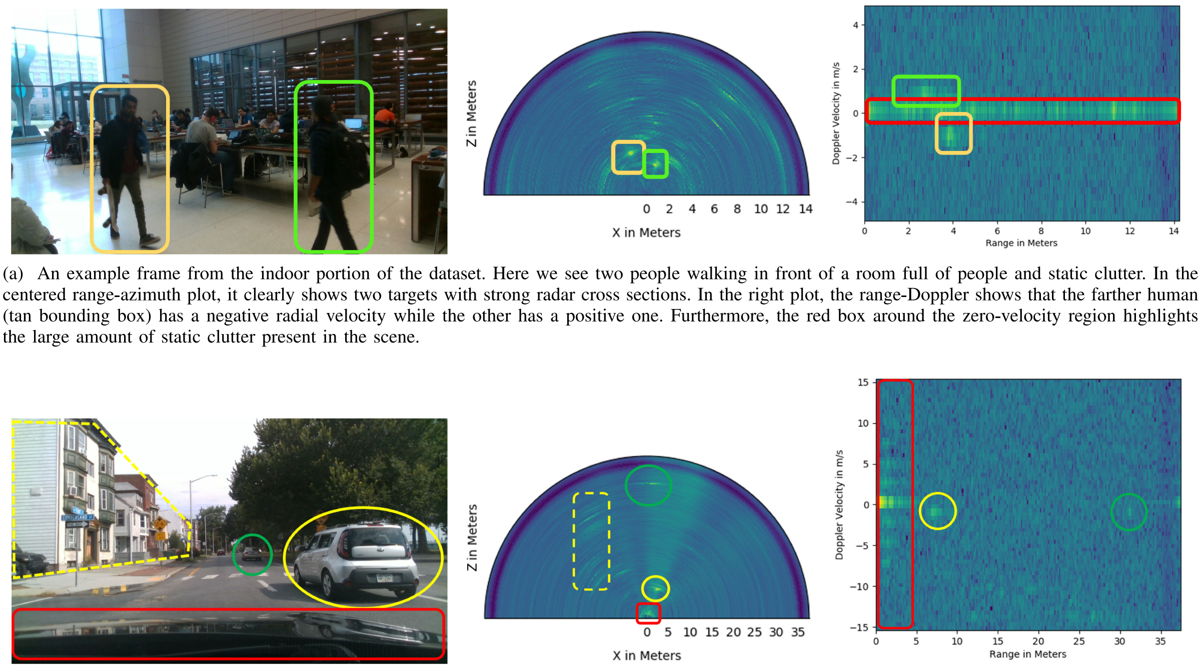
For the indoor scenes, we configured the radar to work at a comparable range to our depth camera. The indoor scenes consist of lobbies of campus buildings as well as a small eating area. These scenes also include a variety of levels of human activity. An example of a crowded scene is shown in Fig. 9(a).

Indoor scenes present many unique challenges to processing the data. First, due to the geometry and material makeup of the enclosure, there will be observable multi-path. Moreover, in many indoor settings, there is a high presence of static clutter including, but not limited to, furniture and decorations. Lastly, in indoor settings, human targets often interact in close proximity to each other, a problem that is not as prominent in outdoor driving settings. These characteristics of indoor scenes demand further algorithm development in the areas of detection, segmentation, and other related fields of deep learning. The provided dataset offers an opportunity to engineers to further develop algorithms and deep learning architectures that can improve detection and segmentation with radar data in these challenging environments.

TABLE II

NUMBER OF FRAMES IN EACH RADAR CONFIGURATION AND SCENE SUBTYPES. DEPTH MEASUREMENTS FROM OUR DEPTH CAMERA IS ONLY RELIABLE UP TO 10 m, THUS SOME ARE EXCLUDED FROM THE ALIGNED FRAMES IN THE OUTDOOR SCENES IN THE LONG RANGE CONFIGURATION. THE IMU MEASUREMENTS WERE TAKEN AT A MUCH HIGHER RATE THAN THE RGB AND RADAR SENSORS AND HAVE THEIR OWN TIME STAMPS AVAILABLE FOR DEAD-RECKONING PURPOSES

Scene Type	Radar frames	Aligned RGB	Depth available	IMU available
Calibration(indoor)	28,245	yes	yes	no
Indoors(indoor)	141,724	yes	yes	no
Parking lot stationary (outdoor30)	64,390	yes	partial*	no
Highway(outdoor60)	19,337	yes	no	yes
Urban/Suburban(outdoor30)	101,468	yes	partial*	yes
Urban/Suburban(outdoor60)	33,494	yes	partial*	yes
Single Human Walking(hihgRes)	5,223	yes	yes	no



(a) An example frame from the indoor portion of the dataset. Here we see two people walking in front of a room full of people and static clutter. In the centered range-azimuth plot, it clearly shows two targets with strong radar cross sections. In the right plot, the range-Doppler shows that the farther human (tan bounding box) has a negative radial velocity while the other has a positive one. Furthermore, the red box around the zero-velocity region highlights the large amount of static clutter present in the scene.

(b) An example frame from the outdoor portion of the dataset. In the centered range-azimuth plot, two cars, the buildings on the left, and the hood of the car being driven are all visible. In the range-Doppler plot, we can see that the car containing the radar is moving faster than the two within range since both of their radial velocities have negative magnitudes. It should be noted that between 0 and 3 meters, the engine of the car produces a very pronounced signature in the range-Doppler. A closer look reveals discrete velocities resulting from the engine's mechanics.

Fig. 9. Sampling of RGB frames and their associated range-azimuth and range-Doppler plots.

C. Outdoor Scenes

We provide two subsets of outdoor scenes with different radar configurations. One with shorter range, but better resolutions (in both range and Doppler), and another with longer range and large maximum velocity, but with poorer resolution. We envision the radar on a vehicle to be able to change its configuration adaptively depending on the scene it's in.

The outdoor scenes consist largely of data collected from inside a car that was driving on a road. The roads driven are a myriad of neighborhood, suburban, highways and city roads. The radar/camera setup was placed in two different positions inside the car as seen in Fig. 10:

- On top of the front dashboard on the passenger side looking in front of the car.
- In the second row of seats looking out the passenger side window.

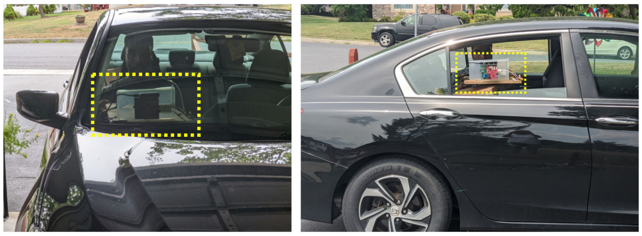


Fig. 10. The two ways in which the radar system was mounted in the moving vehicle for the scenes on the road. In both cases, there was a window in front of the system. For the data from the front facing radar, there are reflections caused by the front of the car, namely the hood and engine.

Both views offer unique perspectives of the road. The front view, as shown in Fig. 9(b), can see oncoming traffic, incoming obstacles, street signs, guardrails when present, as well as the reflections from the engine and hood of the car. The side view offers views of traffic in adjacent lanes, off-road targets such as humans and infrastructure, as well as guardrails when present. While vision/radar sensors are often often seen mounted on a car's roof or exterior, it is also not uncommon to have those sensors mounted inside the cockpit of the car. Furthermore, such a mounting enables the retrofitting of older cars with smart sensors for collision avoidance [33].

Additionally, there are scenes that were taken from a moving wheelbarrow on the road. This unique part of the dataset offers closer and longer views of both cars and humans.

D. High Doppler Resolution

Lastly, our dataset provides synchronized RGB-D and radar data that aims to capture the finer micro-Doppler features of human motion. To accomplish this and stay within the constraints of our hardware and ROS setup, the configuration increases the Doppler resolution of the indoor configuration while lowering the frame rate slightly. Using this configuration with walking humans reveals the cyclic nature of human motion including one's arms, legs, and even torso.

VI. Rich Radar Object Detection and RGB-D Early Fusion

With raw ADC measurements and good synchronization with an RGB-D camera that's easy for humans to interpret and label, we present several opportunities to take radar object detection to the next level.

A. Human Tracking With Radar and Depth

In order to highlight the utility of synchronized radar and RGB-D measurements, we implemented a simple human tracker in radar and compared the results we obtained from the depth camera. In order to make the best use of the depth measurements, and not to amplify small errors through computation, the human walked directly in front of the radar system so that the measured depth roughly corresponded to the range. To detect the location of the human target, we first computed the range-Doppler spectrum and then subsequently used CFAR to find the cluster that corresponded to the human. Within the target's cluster, we took a few of the points with the highest complex magnitudes and averaged their range and Doppler-velocity values for each radar frame.

To determine the depth using the RGB-D camera, applied OpenPose [34] on the RGB frames to find the locations of each of the prominent body-parts. From there, we took the location of the torso and found its depth using the corresponding depth image. As seen in Fig. 11, the data from the two sensors are aligned very well. Furthermore, the velocity results highlight one of the advantages of using radar over traditional RGB-D sensors. Not only does the radar exhibit far less noise than the depth camera, but it also shows the slight variations in the torso's velocity due to each step.

B. CFAR+: Beyond CFAR Object Detection

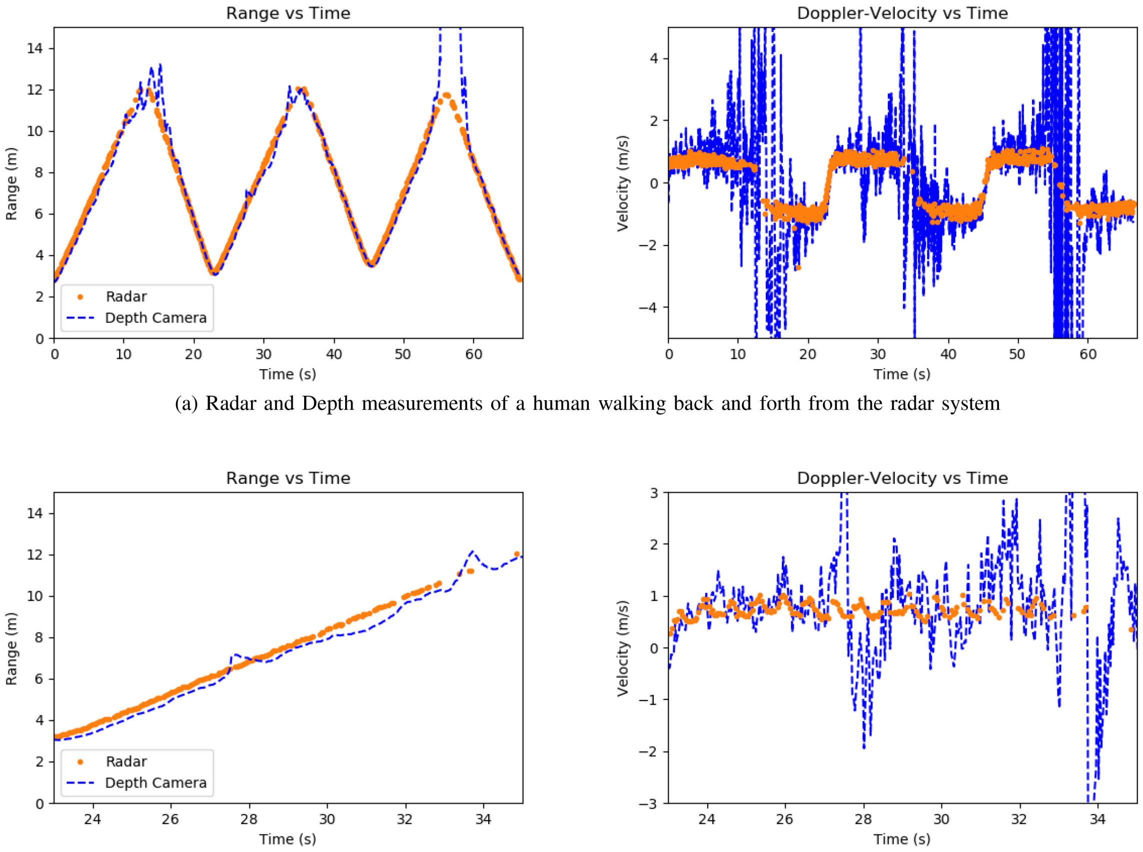
While CFAR, as employed by commercially available radar pipelines, can provide a list of statistically significant radar reflectors as objects, it is unable to differentiate between object classes. With a temporally and spatially aligned RGB-D camera, we can overcome the barriers of labeling otherwise uninterpretable radar signatures either by employing human labelers to label objects in RGB or by automatically applying RGB object detectors on the RGB image, and projecting these labels onto the radar frame, thereby allowing us to inspect the radar signatures of objects in their environment. With such labels, we can go beyond CFAR object detection in radars and create object detectors by applying data-driven object detection methods that were successfully applied to RGB images in a similar method as proposed in [18].

However, object detection networks are known to have difficulty in detecting small objects [35]. Several recent works specifically seek to address this problem. With our indoor radar configuration, assuming that a human occupies the space of a 70 cm square, we should see a radar signature of approximately 12 pixels in the range dimension. While RGB image object detector methods might prove useful for larger objects like cars as in [18], similar performance might be impossible for smaller objects such as humans or bicycles.

In contrast to RGB images, the radar domain has two advantages. First, physical object sizes correspond directly to sizes in the Cartesian projected image. Therefore, we do not need to handle multiple object sizes for the same object class in the same radar configuration. Next, we can assume that our objects of interest will show up as CFAR targets if they reflect the wavelength in use. This allows us to use CFAR detections as object proposals, followed by a classifier on the patch surrounding the CFAR target, assuming that humans have a different radar signature than other statistically significant radar reflectors present in the environment.

To investigate this, we performed our experiments on the indoor dataset, where there are much more radar reflectors compared to an open, outdoor environment. For each RGB image, depth, and radar tuple, we ran a recent state of the art object detector, EfficientDet D6 [36], to obtain a list of bounding-boxes in the camera frame of objects detected as 'person'. These boxes are projected to the radar frame using the calibration parameters described in Sec. IV-D. Next, we applied cell-averaging CFAR on the range-azimuth radar heatmap to obtain a list of point targets. For each point target, we checked if it was enclosed by a projected bounding box. If it was, we extracted a small patch, 32×32 (1.5 m × 1.5 m), from the cartesian projected heat-map around the point target as a class positive training. Points that were not within a bounding box were labeled as environmental reflectors. A sampling of such patches can be seen in Fig. 12. We observed that the neighborhood around CFAR detections corresponding to the 'person' object is qualitatively different from environmental reflectors, thus it is reasonable to assume that we can create an object classifier for such patches.

1) Preparing the Classifier Dataset: We generated our training dataset automatically using projected bounding boxes



(a) Radar and Depth measurements of a human walking back and forth from the radar system

(b) Magnified portion of the human walking away from the radar system

Fig. 11. Plots that show the range/depth and velocity of the torso of a human target. The depth measurements were smoothed using a linear Kalman filter. Despite the smoothing, the depth camera still exhibits a high presence of noise which is conveyed in the velocity plots. On the contrary, in the bottom right plot, the radar's velocity measurements are precise enough to detect the small fluctuations due to the stepping motion of the human.

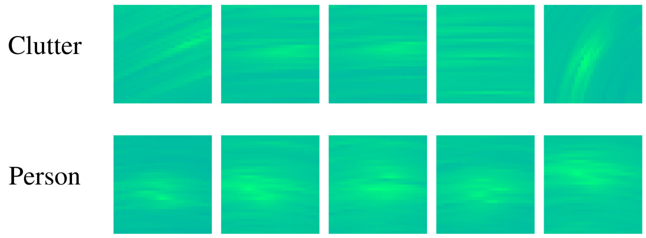


Fig. 12. Randomly selected samples of 32×32 patches surrounding CFAR detections points in the indoor dataset. The top row shows detections that were not enclosed by projected bounding boxes and the second row shows patches where detections were enclosed by a projected 'person' bounding box. We observe that the radar signatures are qualitatively different from strong reflectors present in the environment.

predicted by the EfficientDet-D6 RGB object detector. As there are much more environmental reflectors than there are people in our dataset, we balance the classes in the training dataset by randomly dropping the environment clutter patches so that we end up with a class balanced dataset. Finally, as our dataset consists of video frames, to ensure that our validation dataset is not too similar to the training set, we selected distinct video sequences instead of random frames from the entire collection. Our resultant training dataset consists of 143617 person examples and clutter examples of each, whereas our validation dataset consists of 16108 examples of each.

2) Radar Classification Network: We implemented a small network of three 7×7 separable convolutional layers with ELU activations [40], followed by 2 dense layers with 256 hidden units and ELU activations as our classifier network. We also compared our results with several well-known image classifier architectures, [37]–[39] on the generated classifier dataset. For all networks, we used the Adam optimizer [41], with learning rate 1×10^{-3}, $\beta_1 = 0.9$ and $\beta_2 = 0.999$ and trained with a batch size of 128. Finally we applied random left/right flipping as data augmentation and used early stopping as regularization. We did not use other methods for data augmentation use in image classification as they do not have physical meaning in the radar heatmap domain. Finally, for ResNet50 and VGG16, we applied l_2 weight decay of 1×10^{-5} [38] and 5×10^{-4} [39] respectively, as described in the original papers. No weight decay was applied for MobileNet as recommended in the original paper [37]. We trained each network for 200 epochs and evaluate every epoch, retaining the best performing model on the validation set. All networks were trained from scratch with Glorot uniform initilization [42] as implemented in TensorFlow. The performance of each network network on our dataset is shown in Table III.

TABLE III
PERFORMANCE OF OUR RADAR SIGNATURE CLASSIFICATION NETWORK IN COMPARISON WITH SEVERAL MODERN RGB CLASSIFICATION NETWORK ARCHITECTURES

Network	Train Acc.(%)	Val Acc.(%)	Parameter count
Ours (linear)	95.4	83.1	**1.7M**
Ours (log)	92.7	80.0	**1.7M**
MobileNetv2[37]	93.6	**85.1**	2.23M
ResNet50[38]	94.4	84.08	23.5M
VGG16[39]	49.9	50.0	33.6M

3) Discussion of Classifier Performance: We found that good performance on natural images does not translate to good performance on the Cartesian-projected radar heatmap patches. In contrast to architectures with good performance on natural images, we found that shallower networks with larger filter sizes can give performance comparable to very deep networks. Moreover, using only 3×3 filters, and relying on pooling and depth to increase the receptive field sizes resulted in networks that failed to converge. We also found that while applying a logarithmic scale to the heatmaps resulted in better images for human interpretation, it did not help in deep network performance. The relative small size of the dataset might result in networks with large number of parameters, like ResNet50 and VGG16, to overfit, thus resulting in poor performance. In the case of overfitting, we would expect very high training accuracy but poor validation accuracy. This, however, is not observed in our experiments, and we speculate that intuition and priors [43] that arise due to the network achitectures for natural image classifiers might not apply for our data.

VII. CONCLUSION

We demonstrated baseline results and presented scenarios where modern advances in deep learning could help in getting richer object detection from automotive FMCW radars. We encourage fellow researchers to beat us on our baseline, improve on early stage preprocessing, and design novel network architectures suited for object detection and fusion using previously unavailable low-level radar signals.

On top of providing our dataset, we've also included our hardware BOM and design files which can be 3D printed. We also encourage groups with sufficient resources to build their own system and collect additional data.

Lastly, we also demonstrated the flexibility of a software configurable FMCW radar with our dataset. No hardware changes were required to allow usage where signal requirements were significantly different. A next step in development would be to allow object detection algorithms to actively reconfigure the radar for better performance.

A. Future Work

While we demonstrated a successful use of deep neural networks much earlier in the radar signal processing chain, the work done is not exhaustive. We strongly believe that much better results can be achieved with a more in-depth exploration, for example, transfer learning from pretrained networks, more object classes, and alternative network architectures. Improvements to our simplistic RGB to radar matching methodology could also be explored. Next, while available in the dataset, we did not make use of IMU data, which could be useful in situations where single frame depth measurements are not reliable. Finally, our placement of the radar is within the vehicle cockpit, which may not be ideal for all situations. Other mounting locations of our sensor system could also be explored.

ACKNOWLEDGMENT

The authors would like to thank Texas Instruments for their support for this work.

REFERENCES

[1] C. C. Chen, "Attenuation of electromagnetic radiation by haze, fog, clouds, and rain," RAND CORP MONICA SANTA CA, Tech. Rep., 1975.

[2] K. Garcia, M. Yan, and A. Purkovic, "Robust traffic and intersection monitoring using millimeter wave sensors," Tech. Rep., Texas Instrum., Tech. Rep., 2018.

[3] N. Balal, G. A. Pinhasi, and Y. Pinhasi, "Atmospheric and fog effects on ultra-wide band radar operating at extremely high frequencies," *Sensors*, vol. 16, no. 5, 2016, Art. no. 751.

[4] Y. Golovachev, A. Etinger, G. A. Pinhasi, and Y. Pinhasi, "Millimeter wave high resolution radar accuracy in fog conditions-theory and experimental verification," *Sensors*, vol. 18, no. 7, 2018, Art. no. 2148.

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

[6] O. Schumann, M. Hahn, J. Dickmann, and C. Wöhler, "Semantic segmentation on radar point clouds," in *Proc. IEEE 21st Int. Conf. Inf. Fusion*, 2018, pp. 2179–2186.

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

[8] S. Chadwick, W. Maddetn, and P. Newman, "Distant vehicle detection using radar and vision," in *Proc. IEEE Conf. Robot. Automat.*, 2019, pp. 8311–8317.

[9] M. A. Richards, *Fundamentals of Radar Signal Processing*. New York, NY, USA: Tata McGraw-Hill Educ., 2005.

[10] H. Caesar *et al.*, "Nuscenes: A multimodal dataset for autonomous driving," 2019, *arXiv:1903.11027*.

[11] 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. 12 697–12705.

[12] ESR DelphiDatasheet, 2011. [Online]. Available: http://delphi.com.

[13] Aptiv Electronically Scanning RADAR, 2020. [Online]. Available: https://autonomoustuff.com/products/aptiv-esr-2-5-24v

[14] 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.*, 2020, pp. 6433–6438.

[15] 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.

[16] CTS350 Navtech, 2019. [Online]. Available: https://navtechradar.com/clearway-technical-specifications/compact-sensors/?pfstyle=wp

[17] M. Zhao *et al.*, "Through-wall human pose estimation using radio signals," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 7356–7365.

[18] T. Y. Lim *et al.*, "Radar and camera early fusion for vehicle detection in advanced driver assistance systems," *NeurIPS Mach. Learn. Auton. Driving Workshop*, 2019.

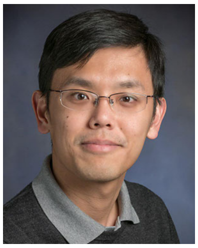
[19] "Remote sensing notes," 1999. [Online]. Available: http://wtlab.iis.u-tokyo.ac.jp/wataru/lecture/rsgis/rsnote/cp3/cp3-4.htm

[20] P. Beckmann and A. Spizzichino, "The scattering of electromagnetic waves from rough surfaces," Norwood, MA, USA: Artech House, Inc., 1987.

[21] "Texas instruments: millimeter wave (mmWave) Sensors," [Online]. Available: https://www.ti.com/sensors/mmwave-radar/overview.html

[22] P. Koivumäki *et al.*, "Triangular and ramp waveforms in target detection with a frequency modulated continuous wave radar," Master's thesis, School Elect. Eng., Aalto Univ., Espoo, Finland, 2017.

[23] J. Capon, “High-resolution frequency-wavenumber spectrum analysis,” *Proc. IEEE Proc. IRE*, vol. 57, no. 8, pp. 1408–1418, Aug. 1969.
[24] R. Schmidt, “Multiple emitter location and signal parameter estimation,” *IEEE Trans. Antennas Propag.*, vol. 34, no. 3, pp. 276–280, Mar. 1986.
[25] S. Rao, “MIMO radar application report,” Texas Instruments, 2018.
[26] J. Bechter, F. Roos, and C. Waldschmidt, “Compensation of motion-induced phase errors in TDM MIMO radars,” *IEEE Microw. Wireless Compon. Lett.*, vol. 27, no. 12, pp. 1164–1166, Dec. 2017.
[27] F. Roos, J. Bechter, N. Appenrodt, J. Dickmann, and C. Waldschmidt, “Enhancement of doppler unambiguity for chirp-sequence modulated TDM-MIMO radars,” in *Proc. IEEE MTT-S Int. Conf. Microw. Intell. Mobility*, 2018, pp. 1–4.
[28] S. Sun, A. P. Petropulu, and H. V. Poor, “Mimo radar for advanced driver-assistance systems and autonomous driving: Advantages and challenges,” *IEEE Signal Process. Mag.*, vol. 37, no. 4, pp. 98–117, Jul. 2020.
[29] E. Pan, J. Tang, D. Kosaka, R. Yao, and A. Gupta, “OpenRadar,” 2019. [Online]. Available: https://github.com/presenseradar/openradar
[30] S. A. I. L. *et al.* “Robotic operating system.” [Online]. Available: https://www.ros.org
[31] R. Hartley and A. Zisserman, *Multiple View Geometry in Computer Vision*. Cambridge, U.K.: Cambridge Univ., Press, 2003.
[32] S. J. Prince, *Computer Vision: Models, Learning, and Inference*. Cambridge, U.K.: Cambridge Univ., Press, 2012.
[33] “Mobileye 8 connect - collision avoidance system,” [Online]. Available: https://www.mobileye.com/us/fleets/products/mobileye-8-connect/
[34] Z. Cao, G. Hidalgo Martinez, T. Simon, S. Wei, and Y. A. Sheikh, “OpenPose: Realtime multi-person 2d pose estimation using part affinity fields,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 1, pp. 172–186, Jan. 2021.
[35] G. Chen *et al.*, “A survey of the four pillars for small object detection: Multiscale representation, contextual information, super-resolution, and region proposal,” *IEEE Trans. Syst., Man, Cybern.*: Syst., to be published, doi: 10.1109/TSMC.2020.3005231.
[36] M. Tan, R. Pang, and Q. V. Le, “EfficientDet: Scalable and efficient object detection,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 10 781–10790.
[37] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, “Mobilenetv2: Inverted residuals and linear bottlenecks,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 4510–4520.
[38] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778.
[39] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” 2014, *arXiv:1409.1556*.
[40] D.-A. Clevert, T. Unterthiner, and S. Hochreiter, “Fast and accurate deep network learning by exponential linear units (elus),” 2015, *arXiv:1511.07289*.
[41] D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” 2014, *arXiv:1412.6980*.
[42] X. Glorot and Y. Bengio, “Understanding the difficulty of training deep feedforward neural networks,” in *Proc. 13th Int. Conf. Artif. Intell. Stat.*, 2010, pp. 249–256.
[43] D. Ulyanov, A. Vedaldi, and V. Lempitsky, “Deep image prior,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 9446–9454.



Teck-Yian Lim received the B.Eng degree in electrical and electronics engineering from Nanyang Technological University and the M.Sc degree in electrical and computer engineering from the University of Illinois at Urbana-Champaign (UIUC). He is currently working toward the Ph.D. degree under the guidance of Prof. Minh N. Do with UIUC. He is concurrently a Senior Member of Technical Staff with DSO National Laboratories, Singapore. His primary research interests include image and audio signal processing, FMCW radars, computer vision and sensor fusion with applications in autonomous and guided systems.

Spencer A. Markowitz received the B.S. degree in electrical engineering from the University of Illinois. He is a Graduate Student with the Electrical and Computer Engineering Department, the University of Illinois at Urbana-Champaign. His primary research interests include FMCW radar, computer vision, object tracking, and deep learning.

Minh N. Do (Fellow, IEEE) was born in Vietnam in 1974. He received the B.Eng. degree in computer engineering from the University of Canberra, Australia, in 1997, and the Dr.Sci. degree in communication systems from the Swiss Federal Institute of Technology Lausanne (EPFL), Switzerland, in 2001.

Since 2002, he has been on the Faculty with the University of Illinois at Urbana-Champaign (UIUC), where he is currently the Thomas and Margaret Huang Endowed Professor in signal processing & data science with the Department of Electrical and Computer Engineering, and hold affiliate appointments with the Coordinated Science Laboratory, the Beckman Institute for Advanced Science and Technology, Department of Bioengineering, and the Department of Computer Science. In 2020–2021, he is on leave from UIUC to serve as the Vice Provost of VinUniversity in Vietnam.

He received a Silver Medal from the 32nd International Mathematical Olympiad in 1991, University Medal from the University of Canberra in 1997, Doctorate Award from the EPFL in 2001, CAREER Award from the National Science Foundation in 2003, Xerox Award for Faculty Research from UIUC in 2007, and Young Author Best Paper Award from IEEE in 2008. He was an Associate Editor for the IEEE TRANSACTIONS ON IMAGE PROCESSING, and a member of several IEEE Technical Committees on Signal Processing. He was elected as an IEEE Fellow in 2014 for his contributions to image representation and computational imaging. He has contributed to several tech-transfer efforts, including as a Co-Founder and CTO of Personify and Chief Scientist of Misfit.