UNIVERSITY OF CALIFORNIA SAN DIEGO

Reliable Sensing for Automation in Adverse Conditions

A dissertation submitted in partial satisfaction of the
requirements for the degree Doctor of Philosophy

in

Computer Science

by

Kshitiz Bansal

Committee in charge:

Professor Dinesh Bharadia, Chair
Professor Manmohan Chandrekar
Professor Aaron Schulman
Professor Zhouwen Tu
Professor Xinyu Zhang

2024

The Dissertation of Kshitiz Bansal is approved, and it is acceptable in quality and form for publication on microfilm and electronically.

University of California San Diego

2024

DEDICATION

This doctoral thesis is dedicated to:

My Family,

For their unwavering support, love, and encouragement throughout this challenging journey. Your belief in me has been my constant source of inspiration.

My Advisors and Mentors,

For their guidance, expertise, and invaluable insights that have shaped my research and academic growth. Your dedication to excellence has been a beacon of inspiration.

My Friends,

For the laughter, camaraderie, and understanding that have lightened the load during the long hours of research and writing. Your friendship has been a source of joy and motivation.

This work is a testament to the collective support and inspiration I have received from those closest to me. Thank you for being an integral part of this academic journey.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

I extend my sincere gratitude to my mentors and friends, who have been unwavering sources of motivation throughout my dissertation journey.

Special thanks to Professor Dinesh Bharadia, my advisor, who introduced me to the fascinating world of radars, providing me with purpose and specialization. Over the span of five years, I've gained valuable technical, managerial, and marketing skills under his guidance. Professor Bharadia not only taught me how to critically analyze research papers but also played a pivotal role in my academic growth.

I would also like to acknowledge the continuous support of my committee members: Professor Manmohan Chandraker, Professor Aaron Schulman, Professor Xinyu Zhang, and Professor Zhouwen Tu. Their constructive feedback significantly contributed to the improvement of my research. The courses I took under each of them laid the essential technical foundation for my work. Special thanks to my undergraduate mentors, Professor Manjunath and Professor Nikhil Karamchandani, whose recommendations led me to the opportunity of pursuing a Ph.D. at UCSD. Gratitude to the CSE department for its vibrant atmosphere and the ECE department for organizing engaging activities, especially the coffee hours.

A heartfelt appreciation goes to the WCSNG lab, which has been my support system and feels like a second family. As one of the initial members, I've witnessed its growth alongside my own. The friendships and collaborations forged here are invaluable, and I'm grateful for the lifelong connections.

I want to recognize the pivotal role played by my collaborators – Keshav Rungta, Siyuan Zhu, Manideep Dunna, Sanjeev Ganesh, Gautham Reddy, Jason Pham, Satyam Shrivastava, Junyi Xu, Ish Jain, and Rohith Vennam. Beyond being excellent research companions, each has imparted skills and lessons that shaped me as a researcher. The countless hours spent in the lab running experiments and brainstorming ideas have resulted in enduring friendships and cherished moments.

My deepest appreciation goes to my family, my constant source of motivation and trust.

To my parents, sister, grandfather, and Himanshi – their unwavering support has empowered me to overcome challenges. A special mention to my roommates Shaurya and Shreyam, who provided a home away from home and kept me grounded. Lastly, heartfelt thanks to my Friday Board Game night friends – Maulik, Anne, Rohan, Rajat, Agrim, and Michael – for the joyous times shared together.

Chapter 1, in full, is a reprint of the material as it appears in "Pointillism: Accurate 3D Bounding Box Estimation with Multi-Radars" by Kshitiz Bansal, Keshav Rungta and Dinesh Bharadia, which appears in the 18th ACM Conference on Embedded Networked Sensor Systems (SenSys '20), November 16–19, 2020, Virtual Event, Japan. ACM, New York, NY, USA. The dissertation author was the primary investigator and author of this paper.

Chapter 2, in full, is currently being prepared for submission for publication of the material by Kshitiz Bansal, Keshav Rungta and Dinesh Bharadia. The dissertation author was the primary investigator and author of this paper.

Chapter 3, in full, is a reprint of the material as it appears in "R-fiducial: Millimeter Wave Radar Fiducials for Sensing Traffic Infrastructure." by Kshitiz Bansal, Manideep Dunna, Sanjeev Anthia Ganesh, Eamon Patamasing and Dinesh Bharadia, which appears in 2023 IEEE 97th Vehicular Technology Conference (VTC2023-Spring). IEEE, 2023. The dissertation author was the primary investigator and author of this paper.

Chapter 4, in full, is a reprint of the material as it appears in "SHENRON-Scalable, High fidelity and EfficieNt Radar simulatiON" by Kshitiz Bansal, Gautham Reddy and Dinesh Bharadia, which appears in IEEE Robotics and Automation Letters (2023). The dissertation author was the primary investigator and author of this paper.

My co-authors have kindly approved the inclusion of the aforementioned publications in my dissertation.

VITA

2013–2017	Bachelor of Technology, Indian Institute of Technology, Bombay
2017–2018	Samsung Research Institute, Delhi
2018–2024	Research Assistant, University of California San Diego
2024	Doctor of Philosophy, University of California San Diego

PUBLICATIONS

"Pointillism: Accurate 3D Bounding Box Estimation with Multi-Radars" The 18th ACM Conference on Embedded Networked Sensor Systems (SenSys '20), November 16–19, 2020, Virtual Event, Japan. ACM, New York, NY, USA

"Radsegnet: A reliable approach to radar camera fusion." arXiv preprint arXiv:2208.03849 (2022).

"R-fiducial: Millimeter Wave Radar Fiducials for Sensing Traffic Infrastructure." 2023 IEEE 97th Vehicular Technology Conference (VTC2023-Spring). IEEE, 2023.

"SHENRON-Scalable, High fidelity and EfficieNt Radar simulatiON." IEEE Robotics and Automation Letters (2023).

ABSTRACT OF THE DISSERTATION

Reliable Sensing for Automation in Adverse Conditions

by

Kshitiz Bansal

Doctor of Philosophy in Computer Science

University of California San Diego, 2024

Professor Dinesh Bharadia, Chair

This dissertation addresses the critical role of perception in autonomous navigation and driving systems, proposing a paradigm shift by focusing on radar as the primary sensor. Traditional light-based sensors, such as cameras and lidars, face challenges in adverse weather conditions, impacting the reliability of autonomous systems. The research explores four key aspects of radar-based sensing: high-quality perception, semantic understanding, contextual information, and scalable deployment.

The first aspect introduces a multi-radar system that mitigates specularity and sparsity issues in radar point clouds, enhancing perception quality. Spatial averaging techniques filter out noise, and a deep learning system accurately converts radar data into bounding boxes for objects.

Semantic understanding is tackled by a novel sensor fusion system, integrating radar point clouds with missing texture and color information from cameras. A unique fusion philosophy breaks feature dependence, ensuring robust performance even in adverse conditions.

Recognizing the importance of environmental context, the dissertation addresses challenges in identifying traffic infrastructure using radar alone. Specialized millimeter-wave backscatter tags, affixed to infrastructure, enhance radar visibility. The Van-Atta array amplifies reflections, RF switches modulate backscattered signals, and unique gold codes enable precise localization without modifying radar hardware.

To facilitate deployment, an open-source radar simulation framework is introduced, simulating high-fidelity Multiple Input Multiple Output (MIMO) radar data using lidar point clouds and camera images. This framework allows researchers to efficiently evaluate algorithms with effectiveness comparable to real-world data, enabling rapid iterations across the radar parameter space.

The comprehensive exploration of multi-radar systems, radar-camera fusion, smart infrastructure augmentation, and the innovative radar simulation framework contributes to advancing autonomous systems, particularly in challenging environmental conditions. The dissertation aims to overcome current limitations and chart a path for future breakthroughs in autonomous perception technologies.

Introduction

Autonomous navigation and driving systems have emerged as transformative technologies with the potential to redefine transportation and enhance safety. Central to the effectiveness of these systems is the pivotal role played by perception, which serves as the bedrock for informed decision-making in dynamic environments. However, the reliance on traditional light-based sensors, such as cameras and lidars, has presented a formidable challenge—these sensors often falter in adverse weather conditions, hindering the reliability of autonomous systems.

Against this backdrop, this dissertation explores a paradigm shift in perception technology, focusing on the integration of radar as the primary sensor. Radar systems offer a promising solution, demonstrating resilience in all-weather scenarios and providing accurate depth and velocity measurements. The research presented herein delves into four fundamental aspects of radar-based sensing: high-quality perception, semantic understanding, contextual information and scalable deployment.

First we propose implementation of a multi-radar system that provides high quality perception. In this work, we tackle the well known problem of specularity and sparsity in radar point clouds, that limits its usage for complex tasks like imaging for autonomous driving. Using a strategic combination of multiple low-resolution radars, we are able to create a high quality perception region, free from the adverse effects of specularity and sparsity. Furthermore, we show how using spatial averaging techniques we can filter out the harmful noise caused by multipath or clutter which may result into false positives. Finally we, present a deep learning based system that can use this high quality point cloud and convert them into accurate bounding boxes for cars.

Accurately determining the location and size of objects through bounding boxes is valuable, but understanding the semantic information of these objects, such as identifying a car as a car or a human as a human, is equally crucial. Radars lack texture and color information, making it challenging to achieve semantic understanding. To address this limitation, we propose a novel sensor fusion system that integrates a camera to supplement radar point clouds with missing texture and color information. This approach enhances the overall semantic understanding of objects detected by the radar.

Existing camera-radar fusion systems exhibit various shortcomings. Some approaches involve using the camera's front view to project radar point clouds onto the camera plane. However, this method proves suboptimal in instances of occlusions, as points from occluded objects may be incorrectly projected onto the front object. To address this issue, alternative approaches combine both the camera's front view and the radar's bird's eye view, extracting separate features from each. Unfortunately, this approach introduces a significant problem of strong feature dependence in both views. Adverse weather or occlusion affecting the camera sensor can lead to a drastic decline in the overall system's performance. In our work, we propose a new fusion philosophy by independently extracting information from each modality before merging it. This strategy breaks the feature dependence during extraction, preventing a sharp decline in system performance when one sensor is compromised due to adverse weather or occlusion.

High-quality perception and semantic understanding are fundamental to radar-based sensing. However, in many cases, additional environmental context is essential for tasks such as navigation. For instance, recognizing traffic infrastructure and roadside signs in challenging conditions like fog or dust plays a vital role in a perception system. Nevertheless, several challenges arise when attempting to identify this context using only radar. The small size of signage results in minimal signal reflection back to the radar. Furthermore, this small signal can easily be overshadowed by larger reflections from more substantial objects like cars. Additionally, it becomes crucial to differentiate one sign from another, even when they exhibit very similar

radar signatures. These challenges collectively make it extremely challenging to identify traffic infrastructure using radar alone.

This dissertation addresses the challenge of identifying traffic infrastructure using a solution involving specialized millimeter-wave backscatter tags. These tags can be easily affixed to the infrastructure, providing enhanced visibility for radar detection. The Van-Atta array incorporated into these tags amplifies the reflection directed towards the radar. To distinguish the tag's reflection from background objects, RF switches modulate the backscattered signal with a distinctive signature. This unique signature aids in recognizing the presence of the tag. Moreover, the utilization of specialized gold codes assigns a distinct identity to each tag, enabling the simultaneous operation of multiple tags. Notably, our tags offer precise localization within the environment without requiring any additional modifications to the radar hardware.

The aforementioned systems provide a robust sensing solution based on radar, applicable both indoors and outdoors for perception in adverse conditions. However, a significant challenge lies in deploying these systems for various applications. Radar involves a vast parameter set, encompassing measurement timings and sensor placements, and each parameter set results in a distinct configuration. Determining the most optimal set for a specific application is a daunting task for radar engineers. Addressing this challenge, the dissertation introduces an open-source radar simulation framework, a transformative tool for researchers in the field.

This framework efficiently simulates high-fidelity Multiple Input and Multiple Output (MIMO) radar data using lidar point clouds and camera images. The key contribution lies in its capability to generate simulated data for algorithm evaluation with a level of effectiveness comparable to real-world data. Furthermore, the framework allows for rapid iterations across the extensive parameter space of the radar, assisting researchers in identifying optimal parameters for diverse applications. The integration of simulation and real-world applicability propels radar perception and sensor fusion research into new realms of efficiency and effectiveness.

The following chapters intricately explore multi-radar systems, radar-camera fusion, smart infrastructure augmentation, and an innovative radar simulation framework. Together,

these components contribute to the progress of autonomous systems, particularly in navigating challenging environmental conditions. By undertaking these efforts, this dissertation aims not only to tackle current limitations but also to chart a course for future breakthroughs in the field of autonomous perception.

Chapter 1

Pointillism

1.1 INTRODUCTION

Autonomous vehicles require a high-quality geometric perception [111, 26, 85, 32] of the scene in which they are navigating, even in adverse weather conditions. Most of the existing computer vision algorithms and data-driven techniques rely on high resolution, multi-channel LiDARs [6, 101, 121, 87] to construct accurate 3D bounding boxes for dynamic objects. LiDAR is a light-based sensor which measures the reflection of the light signal to perceive the geometry of the scene, creating dense 3D point clouds. However, LiDAR cannot penetrate through fog and dust, causing the sensor to fall prey to adverse weather conditions [7, 9, 10]. In contrast, radars are a robust sensing solution, which transmits millimeter waves (mmWaves) and remain less affected by adverse weather conditions [4]. The wavelength of mmWaves allows them to easily pass through fog, dust, and other microscopic particles.

Although radars are all-weather reliant sensors, they need to provide LiDAR-like high-quality perception performance to enable adverse weather perception. However, a challenge with the radar is that it cannot create dense point clouds like LiDAR. The primary reason is that an automotive radar emits mmWave signals, which reflect specularly off of surfaces unlike light signals which scatter in every direction, allowing only a fraction of incident waves to travel back to the radar receiver, as shown in Figure 1.1. Even a high angular-resolution automotive radar suffers from the curse of specularity and would only create a sparse point cloud with insufficient

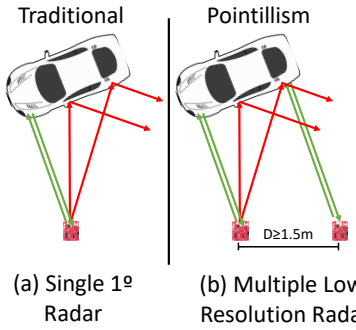


Figure 1.1. a) Single radar misses the reflective surface due to specular reflections (red). b) Multi-radar setup with optimal separation receives reflections back almost surely from the reflective surface (green).

information for precise bounding box estimation. To further exacerbate the issue, radar data contains structured noise due to sensor leakage, background clutter, and multi-path effects. The noise pollutes the point clouds by causing unwanted points to appear in the scene point cloud. This leads to inaccuracies in identifying the number of objects in the scene by increasing false detections.

In this work, we propose Pointillism, a system that enables radars to overcome the challenges posed by specular reflections, sparsity and noise in the radar point clouds, to provide high-fidelity perception of the scene with 3D bounding boxes. Pointillism consists of multiple low-resolution radars placed in a optimal fashion to maximize the spatial diversity and scene information. Pointillism combines this spatial diversity with novel multi-radar fusion algorithms to tackle the problem of specular reflections, sparsity and noise in radar point clouds. Building upon the hardware and algorithms, Pointillism also introduces a novel data-driven approach that enables the detection of multiple dynamic objects in the scene, with their accurate location, orientation and 3D dimensions. Furthermore, Pointillism enables such perception even in inclement weather, thereby paving a way for radar to be the main-stream sensor for autonomous perception.

A natural question is how does Pointillism overcomes the physics of wireless signals i.e., specularity. Pointillism's key idea to overcome specular reflections is to use multiple radars placed at spatially separated locations overlooking the same scene and illuminate an object in the scene from different viewpoints. This, in turn, increases the probability of receiving a

reflection back from multiple points/surfaces of the object, which single radar could have missed, as shown in Figure 1.1. Pointillism formulates the problem of optimal placement of multiple low-resolution radars as an optimization problem to achieve multiple reflection points from a vehicle at all orientation. The optimization reveals that the optimal radar placement of around 1.5 meter apart (typical width of a car) to achieve high-fidelity in the estimated pose of the surface.

A good placement ensures spatial diversity in the point clouds, but noise is still a major challenge. Pointillism presents a novel multi-radar fusion algorithm that reduces the noise points to enable accurate detection of multiple dynamic objects. The insight is to use spatial diversity generated by multiple radars to reduce the noise and enhance points corresponding to actual dynamic objects (eliminate noise). A naive approach to leverage spatial diversity would be to translate the point clouds collected by each radar of the multiple-radar to a common frame of reference and combine them to densify the radar point cloud. However, this approach adds up the noise points as well, doesn't reduce the noise and misses out on the crucial information encoded in the spatial locations of radars.

We make a key observation that across multiple viewpoints (radars), the noise points appears independent of each other in space and the points belonging to actual surface/object appear at nearby location consistently in most of the views (radars). To best leverage the observation, we create a space-time coherence based framework for combining of 3D point clouds from multiple radars. The output is a novel representation of *Cross Potential Point Clouds* that have the information regarding the confidence of each point coming from an actual object as soft probability value, along with all the properties of a point cloud.

With the knowledge of confidence estimates for the points, we can infer whether they belong to objects or noise. However, merely identifying all the relevant points out of noise is not sufficient for multi-object 3D bounding box estimation. Firstly, depending on the distance, orientation, and the exposed surface of an object, only a limited set of points could be captured by the radar. Secondly, in a scene with multiple objects, precise 3D bounding box estimation requires segmenting out the points belonging to each object. This results in massive uncertainty

in the exact orientation and the location of the bounding box.

A naive approach to solve for uncertainty could be to design hand-crafted features by taking into account the shape and size of the vehicles and all possible orientations. However, such an approach is not trivial because crafting the features that can incorporate all possible cases is very challenging. Our insight here is that a data-driven method could potentially solve this problem by building experience over time and learn the non-uniform distribution of radar point clouds.

We reap the advantages of data-driven learning for precise 3D bounding box estimation and propose a novel deep learning-based approach *RP-net*, that leverages the sparsity of Cross Potential Point Clouds. *RP-net* combines the two problems of point cloud segmentation and 3D bounding box location estimation in space, and performs a region of interest (RoI) based classification. However, picking RoIs uniformly throughout the 3D space is not computationally feasible. So, we define a unique set of anchor boxes that allow us to iterate over all the possible configurations of bounding boxes over sparse point clouds. Further, our experiments show that with this set of anchor boxes, we can exhaustively cover all the configurations while efficiently reducing the search space. Our solution is an end-to-end trainable network, which can easily be deployed for real-life testing.

It is well known that a data-driven approach needs a lot of data for training and validation. To the best of our knowledge, no publicly available dataset contains data from multiple low-resolution radars with overlapping field of views. Due to the lack of data, we built an automated data collection platform with multiple radars along with a LiDAR and a depth camera for ground truth labels of scene.

Pointillism achieves a median error of less than **37cm** in localizing the center of an object bounding box, and a median error of less than **25cm** in estimating the dimensions of the bounding boxes. Moreover, Pointillism achieves an overall mAP score [38] of **0.67** with an IoU threshold of 0.5 and **0.94** with an IoU threshold of 0.2 for estimating 3D bounding box, which is comparable with the state-of-the-art bounding box estimation techniques [87, 101, 100] using

LiDARs. Furthermore, with our approach, the mAP values increase to **0.67** compared to **0.45** for a single radar system. This means that Pointillism improves the performance by **48%** with its multi-radar fusion compared to a single radar. RP-net can make inference at a frame rate of 50Hz which is well beyond the real-time requirements.

In summary, our contributions are as follows:

- We propose Pointillism, a new framework for radar perception that leverages spatial diversity induced by multiple radars and optimizes their separation, to counter the fundamental challenge of specular reflections in mmWave radars.
- We conceptualize the notion of *Cross Potential Point Clouds* by utilizing space-time coherence on point clouds from multiple radars, to reduce the noise in radar point clouds, thereby increasing quality of signal.
- We provide the design of *RP-net*, a novel deep learning framework, designed to leverage the non-uniform distribution of radar point clouds, and estimate precise 3D bounding boxes on Cross Potential Point Clouds.
- We build the first real-world dataset of labeled radar point clouds from multiple radars with overlapping field of view, that contains 54k radar frames and corresponding LiDAR point clouds and RGB images in good/bad weather conditions. We believe that this dataset would enable innovation in a variety of sensor-fusion approaches.

1.2 Background and challenges

In this section, we briefly describe the point cloud generation process used in the current automotive radars and outline the challenges involved in working with radar point clouds.

1.2.1 Radar Processing Pipeline

FMCW (Frequency Modulated Continous Wave) technique has become a standard in the automotive radar market for point cloud generation and velocity estimation [2, 76].

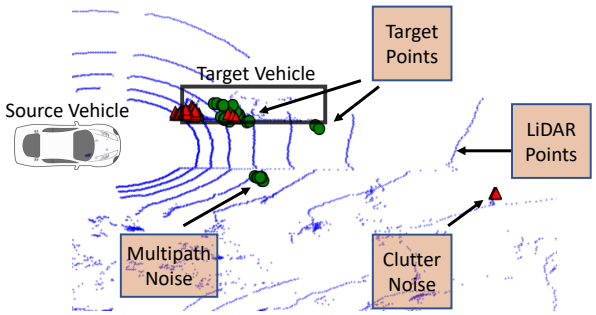


Figure 1.2. Point-cloud of a scene with one target car using multiple radars (red triangles and green dots are two different radars, blue points are from LiDAR). Multiple such target vehicles could be present in the scene. Figure shows noisy radar points generated due to clutter and multipath effects, and their independence across two radars.

Additionally, multiple transmit and receive antennas are used for the angle of arrival estimation using beamforming. We would refer the reader to [2] for a detailed description of the entire point cloud generation. The resolution in range and angle estimation in FMCW depends on the bandwidth of operation and the number of available antennas, respectively.

The current market trend is to achieve higher angular resolution by using several co-located antennas and improve bounding box estimation performance. However, radar point clouds are affected by more fundamental challenges that inhibit the bounding box estimation performance on radar data. In the next section, we describe the challenges with radar point clouds and discuss how they affect the bounding box estimation performance.

1.2.2 Challenges in radar point clouds

Millimeter wave sensing has enabled high resolution radars for automotive sensing but it has its own perils. There are three main challenges faced in mmWave sensing:

(a) Specular reflections: For an incident electromagnetic wave on a surface, the size of its wavelength compared to the roughness of the object's surface determines the degree of scattering of the wave. mmWaves undergo a negligible scattering effect, resulting in a specular reflection (angle of incidence = angle of departure) from the surfaces. Consequently, for a small aperture radar, a lot of reflected signal does not make its way back to the sensor, causing blindness of the

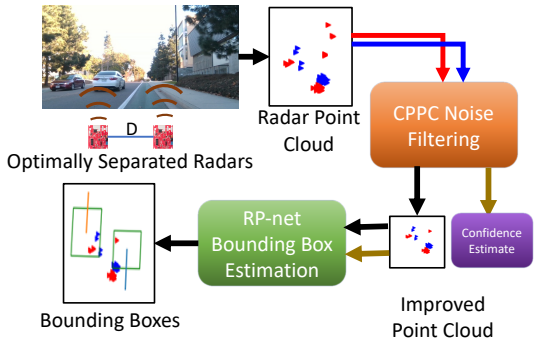


Figure 1.3. Overview of Pointillism. PC: Point Cloud

objects. This blindness is even independent of the resolution capabilities of the sensor.

(b) Radar clutter and noise: Radar detections are commonly known to be polluted by signals from clutter, noise, and multipath effects. Radar clutter is defined as the unwanted echos from the ground or other objects like insects that can be confused with the objects under consideration. In a congested environment like cities, a signal emitted by a radar sensor could suffer multiple reflections before coming back to the sensor. The result is the formation of ghost objects, which are reflections of actual objects in some reflector formed because of multipath(Figure 1.2).

(c) Sparsity: Outdoor scene point clouds are inherently sparse due to the empty volume between the objects, which are at a substantial distance from each other. Additionally, due to different interaction properties of mmWaves with different objects (non-uniform interactions), this effect is compounded in the case of mmWave radars. The result is a sparse and non-uniform point cloud (Figure 1.2).

Pointillism aims to tackle each of these challenges and provide accurate bounding boxes. Our approach towards achieving this goal is to couple multiple radar fusion with a novel noise filtering algorithm and estimate the bounding boxes. Figure 2.2 shows an overview of our entire processing pipeline.

1.3 Multi-Radar Perception

Specular reflections of millimeter waves can cause direct blindness of object surfaces, which could lead to fatal accidents. To better scrutinize the effect of specularity, we need to un-

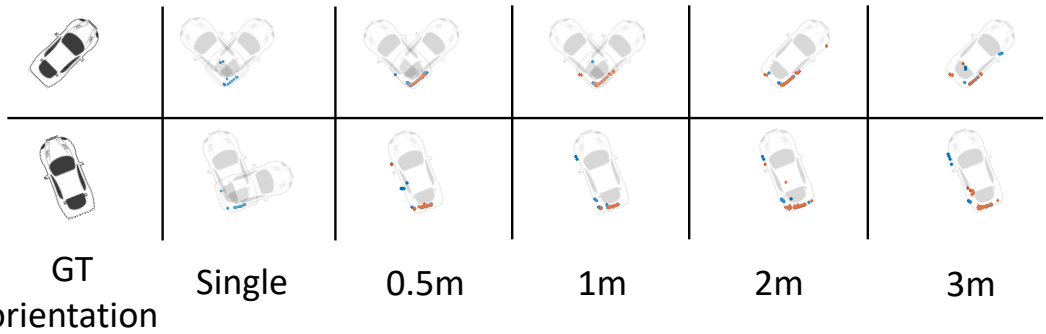


Figure 1.4. Wireless Insite simulations: Point clouds with single radar or smaller separation could lead to ambiguity in pose of the car which gets eliminated by increasing radar separation. Orange and blue points are from 2 different radars.

derstand the distribution of a radar point cloud. Point cloud generated by a radar depends mainly on two aspects: *geometry* of the scene and *resolution* of the radar (figure 1.6b). Importantly, the adverse effect of specular reflections is a *geometric* shortcoming and can not be tackled simply by increasing the resolution of the radar.

1.3.1 Studying Radar point clouds

To study the effect of scene geometry, we need to consider radars with a perfect resolution that can resolve any two reflections which are arbitrarily close to each other. We use a high-fidelity EM wave propagation tool Wireless InSite to model the EM wave interactions [74]. Our experiments using a 1-degree radar (1.8), and the results in past work [95], show that the radar reflection characteristics obtained from these simulations are similar to a real-world radar. We create simulations using a CAD model of a car along with its material properties. We create multiple simulations by placing the car at several locations in an area of 10m x 10m in front of the radar(s). At each location, simulations are created for multiple orientations(angles) of car, spanning 360 degrees in 36 discrete steps. For each placement and orientation, we consider two radar setups: *Single radar* and *Two radars with varying separation distance*. In simulations, the radars can resolve all the rays returning to the receiver (perfect resolution). With this, we can independently compare how the geometry of the scene affects performance.

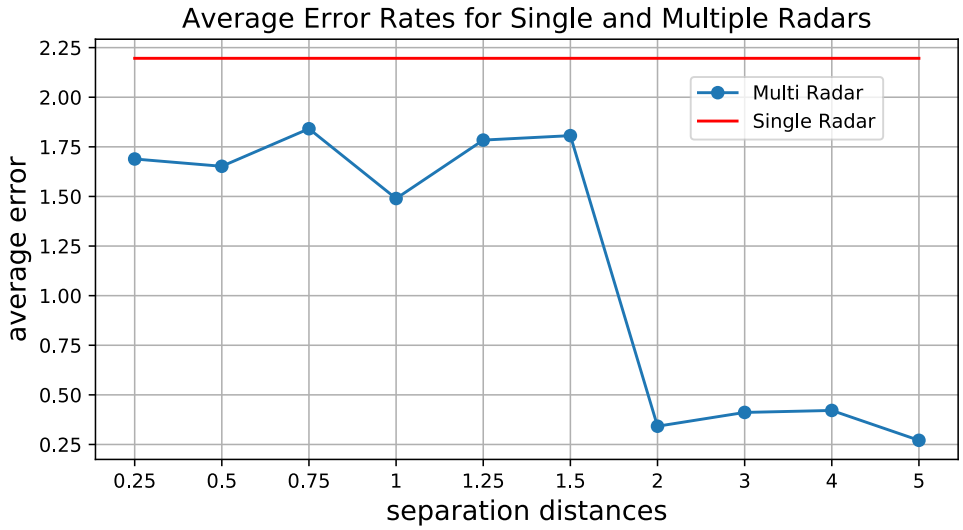


Figure 1.5. Comparison of average error in radian between single and multiple radars. For a separation greater than 1.5m, the performance improves.

1.3.2 Performance comparison of single and multiple radars

For a rectangular bounding box of a car, the system that captures more surface points (less affected by specularity) would perform better in estimating the bounding box's orientation. To estimate the orientation angle of the box from point clouds, we use an MLP (multi-layer perceptron) regressor from *scikit-learn* [84] trained on the data generated from the simulations. MLP regressor takes a simulated point cloud as input and outputs the orientation angle of the car. Figure 1.5 shows the comparison of performance in terms of the mean error made in angle estimation. We make two important observations from these results. Firstly, the results show that the two radar system outperforms the single radar system. Secondly, a more interesting observation is that there is a sharp increase in performance between the separations 1.5m and 2m. It remains relatively constant before and after. The width of the car we used in simulations is $\approx$ 1.7m, which is the width of a standard car. Hence, the results show that the optimal distance between two radars for estimating a bounding box on a vehicle should be comparable to the vehicle's width. Figure 1.4 further shows how increasing the separation between radars would improve the point cloud generated.

What do multiple radars bring to the table? Based on our analysis following are the key benefits of using multiple radar system:

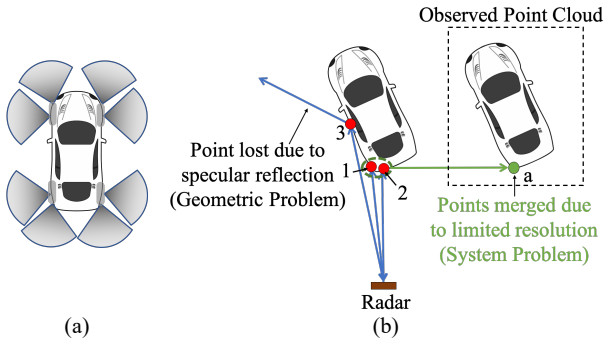


Figure 1.6. (a) Regions of reflection on a car. (b) Specularity vs Resolution: In the observed point cloud, point 3 is missing due to specular reflection while point *a* is formed from points 1 & 2 due to limited resolution.

- **Larger Virtual Aperture:**The reflections arriving from a car originate from specific scattering points around the car (e.g., wheelhouses, corners) [18, 46]. Each scattering point has a specific visibility region (Figure 1.6a). Multiple radars present at spatially separated locations create a larger virtual aperture, thereby capturing more of these reflecting centers. Occluded parts of an object from one viewpoint would appear in the other viewpoint, inherently increases the probability of capturing more meaningful points. This increase directly impacts the system's orientation estimation performance, as is evident from the results.

- **N times the number of points:** One obvious advantage of having multiple sensors is the increase in the number of points, which would densify the sparse point cloud.

- **Rich Spatial Diversity:** A larger virtual aperture with multiple radars provides rich spatial diversity. In the next section, we would show how this spatial diversity could be used to curb noise by building confidence for the points in the point cloud.

1.4 Cross Potential Point Clouds

Multiple radars work together to overcome the challenges posed by specular reflections of mmWaves by providing rich spatial diversity. However, noise (due to clutter, multi-path, and system noise) still presents a big challenge for object detection from radar point clouds. The noise

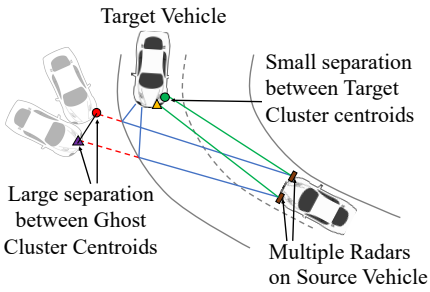


Figure 1.7. Formation of ghost clusters due to multipath. The centroids in the ghost clusters are more separated in comparison to the centroid of target cluster and given low potential values in Cross Potential Point Clouds.

points misguide object detection and introduce false positives. A single radar has fundamental limitations in removing noise. The features corresponding to the Cartesian coordinates, e.g. (x,y,z) and velocity, can provide a rich context for object detection. However, they do not provide information necessary to segregate the noise.

In this section, we would introduce a novel representation of *Cross Potential Point Clouds*, formed by the fusion of multiple radar point clouds. Our key observation is that points belonging to noise are independent across multiple radars placed at different spatial locations. In other words, points belonging to an actual object are more likely to be present in the point clouds of more than one radar. In contrast, the points belonging to random noise are specific to each radar. By leveraging this observation, we design a novel algorithm that filters noise from radar point clouds and creates low-noise Cross Potential Point Clouds. In the following subsection, we would describe how we can use the data from multiple radars to encode information regarding noise.

1.4.1 Space coherence with Radar point potential

Noise harms the bounding box estimation as it creates false positives. Our key insight on tackling noise takes inspiration from signal processing techniques. In signal processing, we collect multiple noisy data stream and take their average to reduce the noise variance and improve the overall SNR (signal to noise ratio). The idea is that the signal present in each data stream adds up coherently. At the same time, the noise is random and will not add constructively. While the idea is intuitive, it is non-trivial to extend it to the point clouds. We cannot simply add point

clouds from multiple radars in the hope to reduce noise because of the following reasons: (1) The 3D point cloud is sparse and incoherent in space, i.e., multiple radars may capture different points in 3D space for the same target object. (2) It is hard to build confidence for every point, whether it contributes to the object bounding box or corresponds to the noise point (generated by clutter or multi-path).

To apply the space coherence in the point cloud domain, we use the geometric information of point clouds. The above insight propagates to the fact that if a region of 3D space generates a response in multiple radars, it is likely to be generated from an object and not noise. To capture this effect, we need to measure the coherence between point clouds originating from multiple radars across 3D space. As shown in Figure 1.6, radar points from an object are clustered around some scattering regions on a vehicle. By identifying these clusters, we can define our confidence of a point being generated from an object by looking at the same scattering region in multiple radar point clouds (space coherence). Our approach runs in two steps: (a) Clustering the point clouds and (b) Enforcing space coherence by defining cross-potentials, detailed as follows:

(a) Clustering the point clouds: Radar point clouds are present in the form of clusters of points originating from a scattering region on the object (Section 1.3). We use the standard DBSCAN [33] algorithm to find clusters in our point cloud. DBSCAN works on the notion of defining a neighborhood of points based on distance ε given as an input parameter. If a specific number of points (another input parameter) is present in that neighborhood, the point and its neighborhood are identified as a cluster. For each cluster i, the centroid c_i of its points is used as the cluster's representative point. For the multi-radar case, we use c_i^j to represent the centroid of cluster i in the point cloud of jth-radar, which is represented by Γ_j. In this way, we create multiple clusters individually for each radar point cloud.

(b) Cross-potential: Next, we discuss how we find the correspondence between clusters across multiple radars. We define *cross-potential* between two clusters from two different radars as a confidence metric if the two clusters belong to the same object. Intuitively, the cross-potential between two clusters is inversely proportional to the distance between the two clusters' centroid.

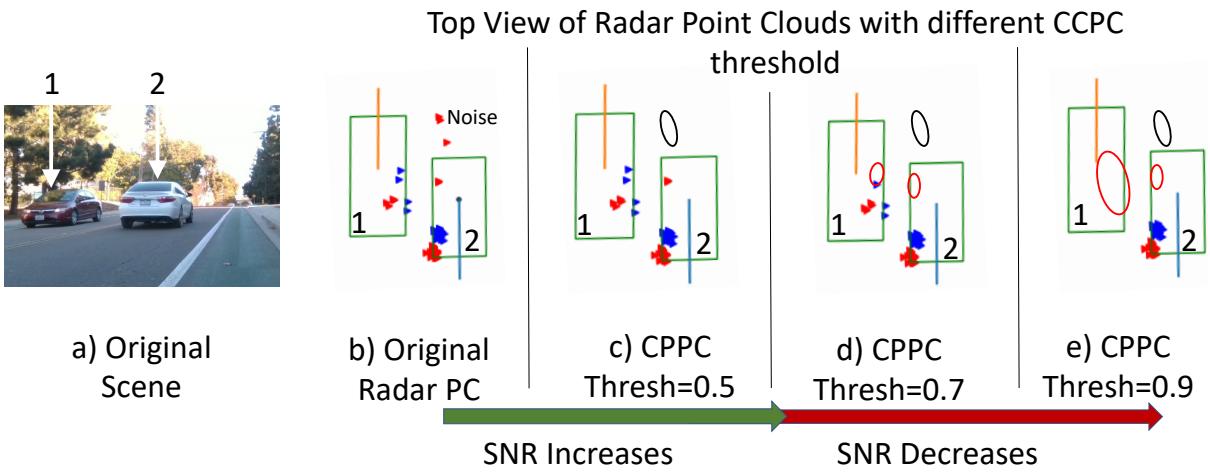


Figure 1.8. Scene with multiple cars and effect of changing potential thresholds. Black (or red) ellipses show regions from where noise (or actual) points get filtered. At a very high threshold (0.9), all points from one car get filtered out, causing miss-detection. The blue and orange line shows the direction of car.

We denote the cross-potential as $P(c_i^j|\Gamma_k)$ for the ith cluster in radar j (c_i^j denotes its centroid) with the kth radar for $k \neq j$, $k \in \{1,2,\ldots,N\}$ in an N-radar system. Mathematically, we define $P(c_i^j|\Gamma_k)$ as

$$P(c_i^j|\Gamma_k) = \frac{1}{1+\left[\frac{r_i^{jk}}{2}\right]^2} \tag{1.1}$$

where r_i^{jk} is the distance between centroid c_i^j from the respective nearest cluster centroid in kth-radar's point cloud Γ_k.

The reason behind this specific choice of potential is motivated by the dimensions of a car. A typical car has a width of around 2m. Therefore, we choose a function that gives a high potential to all the points within the 2m neighborhood of a point, i.e., P¿0.5 if r_i^{jk}¡2m. Note that our choice of potential function also preserves a point lying on a vehicle, which is present in only one radar, as long as it is in the 2m neighborhood of another high potential point (possibly on the farther corner along the width). With this, we preserve the extra points added due to the spatial diversity of the multiple radar system (Section 1.3). Using this potential function, we can quantify the *space coherence* of signals coming from multiple radars. We combine all the points from multiple radar point clouds and add confidence information to them. Each point gets the

same potential as its respective cluster centroid. Further, we filter all the points below a certain potential-threshold (discussed in detail later) to create *Cross Potential Point Clouds*. (We would refer to this as CPPC algorithm for brevity). Figure 1.7 shows how CPPC algorithm gives a low potential to noise or multipath points.

To study the effect of different threshold values on CPPC, we define SNR (signal-to-noise ratio) of a point cloud as the ratio of the total number of actual points against the noise points. Noise points are defined as the points lying outside the bounding box of the vehicle. Figure 1.9 shows the CDF plot of SNR improvement by using CPPC over the case where no CPPC transformation is applied (points from all the radars are combined without any filtering) for multiple potential thresholds. The plot shows that improvement in SNR increases with a higher potential threshold (green region) but with diminishing returns for large threshold values. Certain cases with low or no noise do not show improvement (blue region). However, a large threshold also decreases the SNR to a large extent in some cases, as it regards some actual points as noise and removes them, thereby reducing the SNR (Red region). The SNR decrease is severe for higher thresholds (¿0.5) and could lead to missed-detections (Figure 1.8). Hence, we choose 0.5 as our operating point to achieve a good trade-off between SNR improvement and missed-detections.

1.4.2 Time coherence with multiple frames

Space coherence can help to determine the noise out of the actual signal and boost detection performance. However, the task of estimating a bounding box is still non-trivial due to the sparsity of point clouds. Specifically, estimating the heading direction (also called pose or yaw angle of the bounding box) of a vehicle from a sparse point cloud within a single frame is challenging. In many cases, the points in a single frame might not contain enough information to reveal the heading direction.

To solve this problem, we make an important observation that along with the space coherence across the radars, the points also follow a *time coherence* across multiple frames from

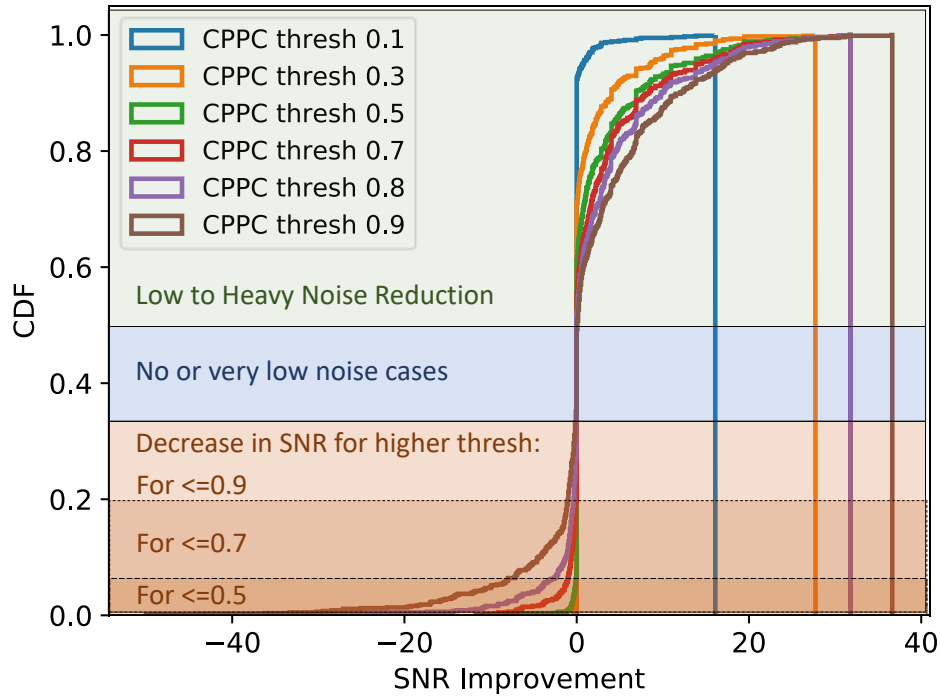


Figure 1.9. SNR Improvement: Comparison of SNR improvement over base case (without CPPC) for multiple potential-thresholds. Higher threshold rejects more noise (green) with diminishing returns. Blue region shows the cases with very low noise in base case, hence there is no improvement. Red region shows the cases where SNR decreases severely for higher thresholds.

radar. For a point originating from a rigid body, the linear motion would be the same as the car's motion across multiple time frames. We track the movement of points in consecutive frames and use it to estimate heading direction. Tracking is performed on self-motion compensated frames to remove the effect of the source vehicle movement. Kalman filter-based corrections are employed to tackle sensor uncertainties and noise. We track the points with the highest cross-potentials for each object. Using the time coherence between points, we can get a prior estimate of vehicles' heading directions in the scene.

1.5 Multi-Object 3D bounding boxes

Using multiple radar fusion to create *Cross Potential Point Clouds*, we can tackle noise and potentially eliminate in-accuracy in the detection of the number of dynamic objects. However, estimating the number of dynamic objects and their bounding boxes is still not a trivial task. To understand this, we need to look into the mapping function of radar. Given a scene geometry of

multiple vehicles and environment, a radar maps it to point cloud information as

$$x, y, z = \mathscr{F}(\text{scene geometry}) \tag{1.2}$$

Our aim is to inverse this mapping and estimate the scene geometry in terms of object bounding boxes.

$$(p_N, \psi_N) = \mathscr{D}(\Gamma_{CPPC}) \tag{1.3}$$

where N is the unknown number of objects present in the scene; Γ_{CPPC} is Cross Potential Point Clouds where each point is denoted by its cartesian coordinates, velocity, intensity and CPPC confidence; p_N represents the confidence of detection for Nth objects's bounding box in a scene and $\psi_N : \{c_x, c_y, c_z, w, h, l, \theta\}$ denote the tuple of bounding box parameters which are center coordinates, dimensions and yaw angle (angle with respect to z-axis) respectively. $\mathscr{D}$ is the multiple 3D bounding box estimation system. Estimating 3D bounding boxes for objects from a radar point cloud has the following challenges:

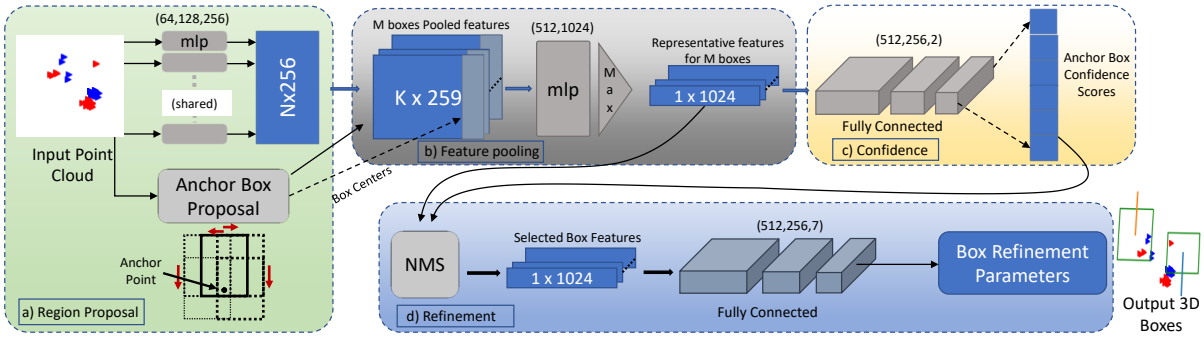


Figure 1.10. RP-net architecture: a) Our network takes *CPPC* as input with N points and 6 channels, extracts features using a pointnet layer, and generates anchor boxes for each point. b) A Feature pooling layer pools point features from K points lying inside each box, which are passed through pointnet and max pool layers to generate representative features for anchors. c) Confidence values are predicted for each anchor. d) High confidence anchors are taken to the box refinement stage. The output is a set of 3D bounding boxes for each object in the scene (top view box is shown).

- **Segmentation:** The radar point cloud of a scene is sparsely distributed where any subset of points could belong to a single object. Also, the number of objects and their locations are not known a-priori. Bounding box estimation requires proper segmentation of points

belonging to each object in the scene. This is a complex mapping problem where the number of targets is not known.

- **Uncertainty in bounding box parameters:** Radar can only see the part of an object which is exposed to the sensor. Therefore, the point cloud of an object may not contain crucial information regarding all dimensions, orientation, and center-location of the bounding box. As a result, there is uncertainty in bounding box parameters.

Estimating $\mathscr{D}$ with an exponentially large search space is a complex problem. A naive solution like fitting a fixed-sized bounding box based on point clusters will not work due to a lack of segmentation information. Proper segmentation of points into different objects and accurate 3D bounding box parameter estimation would require complex hand-crafted feature representations of the point cloud. However, recent advances in deep neural networks [89, 101] have shown that we could overcome these challenges by learning complex feature representations from data. In other words, a deep network can build experience regarding the relation between point clouds and exact object placements and use that experience to estimate accurate 3D bounding boxes.

1.5.1 RP-net Overview

To overcome the challenges mentioned above, we present a novel deep learning architecture called *RP-net*. Our network is specifically designed to handle the sparsity in radar point clouds and output accurate 3D bounding boxes. Directly giving the raw point cloud as an input can not solve the problem due to the unknown number and locations of the objects in the scene. Usually, an object detector uses a fixed set of initial region proposals to solve this problem. Our network proposes a novel way of generating these region proposals (anchor boxes), based on the radar response due to vehicle geometry and space-time coherence. Unlike LiDAR, where many points originate from ground and other static objects like buildings, radar data is sparse and contains mostly the points from dynamic and metallic objects like cars after CPPC noise suppression due to strong EM reflective properties of metals. Specifically, the sparsity in radar data and the fact that all the points originate from the vehicles' surface allow us to define

point-based region proposals. These fixed size anchor boxes are used as initial estimates of 3D bounding boxes. The size of these anchor boxes is determined by the average size of vehicles in the training dataset. Now, instead of the entire scene point cloud, the network works separately on each of these anchor boxes. For each of these anchor boxes, the task is reduced to generate a confidence number p of whether the points inside that anchor box belongs to an object.

Confidence scores are generated for all the anchor boxes. A set of high confidence boxes is chosen. These anchor boxes are passed through a refinement stage that solves the uncertainty issue in bounding box parameters. In this stage, the anchor boxes are refined to provide accurate 3D bounding boxes of the objects present in the scene (parameter ψ).

1.5.2 Network Architecture Design

The job of RP-net is that given a scene point cloud, output the set of 3D bounding boxes of the objects present in the scene. It takes CPPC as input, which has 6 channels corresponding to x,y,z coordinates, velocity, peak intensities, and cross-potential values for each point. RP-net comprises of following blocks:

(a) Handling Multiple Objects. For a given scene with multiple vehicles, region proposals are defined by placing multiple anchor boxes in the scene. As mentioned above, we propose a novel point-based region proposal (anchor boxes) generation scheme based on the radar response due to vehicle geometry and space-time coherence. Given a point, we use five different placements of anchor box around the point (we call this point as *anchor point* for those anchor boxes. Refer Figure 1.10). We use the pose values derived for each anchor point using space-time coherence for the orientation angle, as explained in Section 4.

(b) Segmentation by Feature Extraction and pooling. Our objective is to perform classification and 3D bounding box parameter regression by learning meaningful feature representations from the point cloud data. RP-net extracts these features in two stages, i.e., before and after generating *anchor boxes*. We first use a pointnet [89] encoder of shared MLP to extract features from the entire point cloud. We would refer the reader to [89] for an in-depth understanding of pointnets.

In the second stage, the anchor boxes are determined for each point. An RoI (Region of Interest) feature pooling block of the network pools the features from all the points lying inside an anchor box. These features are passed through another pointnet layer and then max pooled into a single representative feature for every anchor box defined per scene.

(c) Bounding Box confidence prediction. The entire set of representative features of anchor boxes, obtained after the previous block, is passed through a classification network consisting of fully connected layers. The fully connected layers learn a mapping from anchor boxes' representative features to the confidence value for each box.

Performing classification on RoI based max-pooled features always ensures that the contextual information from all neighborhood points of the anchor point, lying inside the anchor box is accounted, leading to better classification results. The problem of segmentation is solved by performing classification directly on the anchor boxes. The network will learn to choose the corresponding anchor box with high confidence, which contains all the points belonging to an object.

(d) Refinement of Box Parameters. In the previous step, the anchor boxes were rough estimates of the dimensions, center, and orientation of final 3D bounding boxes as we used fixed-size anchor boxes. We still need further refinement of these parameters to get accurate bounding boxes. Note that this step is quite essential to estimate the accurate dimensions and location of the boxes. After the classification step, we obtain the confidence scores for all the anchor boxes. Since we obtained anchor boxes for each point, there could be many overlapping high confidence boxes belonging to the same object. Non-maximal suppression(NMS) sampling is performed on this set using the confidence values. NMS sampling removes boxes which have a high overlap with another high confidence box of the same object. The representative features from the remaining anchor boxes are passed through three fully connected layers to output a tuple $[h', w', l', x', y', z', \theta']$ corresponding to refinements of length, breadth, height, center coordinates, and orientation angle respectively. These refinements are added to the anchor box parameters to generate the final 3D bounding box prediction.

(e) Loss functions. The anchor box classification in the first stage of the network is a binary classification problem that uses a cross-entropy loss, given by

$$\mathscr{L}_{RPN} = \sum_{i=1}^{N} -(y_i \log(p_i) + (1 - y_i) \log(1 - p_i)) \tag{1.4}$$

where $y_i = [0, 1]$ is the ground truth and p_i is the predicted confidence value. Refinement of the bounding boxes is a regression problem and we use Smooth-L_1 loss for this purpose. The loss is given by:

$$\mathscr{L}_{refinement}(r, r') = \begin{cases} \frac{1}{2}(r - r')^2, & \text{for } |r - r'| < 1. \\ \delta|r - r'| - \frac{1}{2}, & \text{otherwise.} \end{cases} \tag{1.5}$$

where r and r' are ground truth and regressed refinement values respectively for each parameter $[h', w', l', x', y', z', \theta']$.

1.6 Experimental Setup and Dataset

Table 1.1. Values of the radar parameters used. (Res.=resolution)

Parameter	Value	Parameter	Value
Start Frequency	77 GHz	Frame rate	30 fps
Bandwidth	2240 MHz	Range Res.	0.067m
ADC rate	7500 ksps	Velocity Res.	2.59 m/s
Chirp Duration	40 μs	Max velocity	20.74 m/s

The use of radars for perception in autonomous systems is quite recent. No publicly available dataset collects data from multiple radar sensors with overlapping fields of view in 3D. We collect our own dataset for training and testing Pointillism. We collected data from three sensors: a 16 channel Ouster LiDAR [5] placed and an Intel RealSense D415 [3] Camera at the center of a rail , and 2 TI IWR1443BOOST [2] radars placed at the end-points of the rail at a distance of 1.5m, all placed on a car. Figure 1.12 shows our data collection platform. We use LiDAR for ground truth annotations and camera for visualization only. Data is labeled using an online tool called *scalabel* [8]. All the sensors are controlled using a central controller

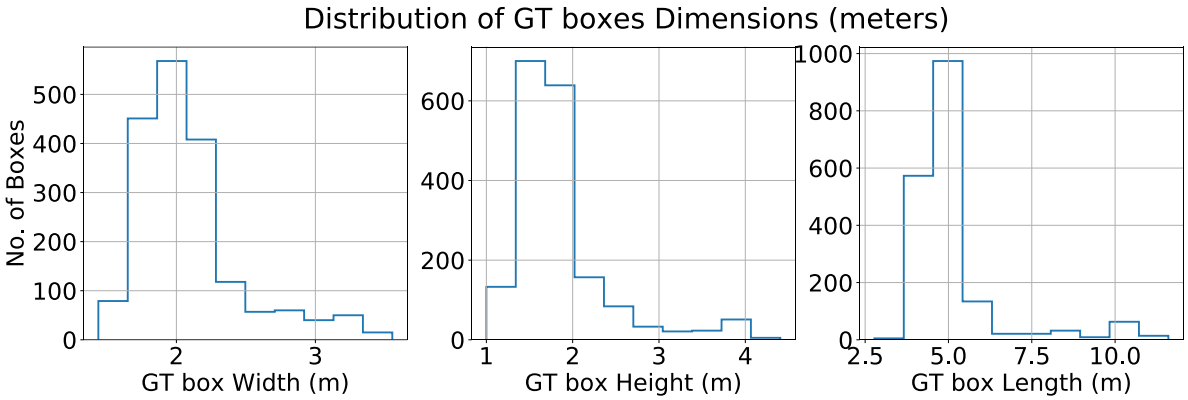


Figure 1.11. Dataset Distribution: Histogram plots of the dimensions of the vehicles present in the dataset. It includes small golf carts to large buses.

running ROS (Robotic Operating System), running on a laptop placed in the car. Each data frame is timestamped for synchronizing the sensors. We perform controlled experimentation based sensor calibration. The data from all the sensors are brought to the same coordinate system by deriving extrinsic matrices for each sensor. Our hardware design is modular so that sensors can be swapped, and only data extraction for that sensor needs to be updated.

Each radar has a 3Tx and 4Rx MIMO antenna array. We use the FMCW processing chain provided by TI for generating point clouds characterized by x,y,z coordinates, doppler, and the intensity (peak values) of points [2]. Table 1.1 summarizes the operating parameters of our radars. Figure 1.11 shows the distribution of the different sizes of objects/cars present in our dataset. The dataset includes small golf carts to large buses with a maximum number of 4 objects in the scene. In total, we collect 54k radar frames for five real-life traffic and driving scenes. Data was collected in different operating conditions, including day, night, and adverse weather (fog) conditions (smoke machine experiments). It is the first of its kind radar dataset with data from multiple radars, LiDARs, and RGB cameras that can enable various sensor-fusion-based approaches. We will release the dataset for public use. We plan to keep expanding this dataset with more samples over time.

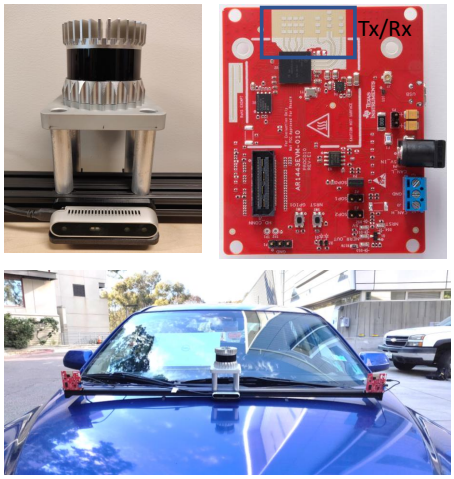


Figure 1.12. Pointillism data collection platform. (Top Left) 16-channel Ouster LiDAR, RealSense Depth Camera for ground truth labels. (Top Right) our radars. (Bottom) Our System: Two radars, LiDAR, Depth camera combined synchronously to common clock.

1.7 Implementation

In this section we would describe the implementation details of Pointillism.

Deep learning parameters: For implementing our network, we use the PyTorch framework. The number of output channels is mentioned above each layer in Figure 1.10. We use Adam optimizer with a learning rate of 0.0002 and momentum 0.9. A batch-normalization layer follows each layer of the network. We use (2m, 2m, 5m) as (width, height, length) to initialize the anchor boxes' dimensions. 70 random points are sampled from each Cross Potential Point Cloud for giving input to the network. The network takes 6 channel inputs for xyz-coordinates, velocity, peak intensity values, and confidence estimates. We sample 32 points at random from all the points lying inside an anchor box during the feature pooling stage. If the number of points is less than 32, then the same points are repeated to maintain consistency throughout network operation.

End-to-end training and deployment: For training the classification network, ground truth is generated based on the IoU (Intersection over Union) values of the anchor boxes against the ground truth bounding boxes. We use the top view of the anchor boxes to calculate 2D IoUs. We choose choose top 100 anchor boxes based on IoU match for classification. Boxes with 2D IoU overlap of greater than 0.2 are considered as positive examples for classification. For the

Figure 1.13. Predicted 3D bounding boxes using only radar (red) along with ground truth boxes (blue). The 3D IoUs of predictions are 0.52, 0.43 & 0.83 respectively. Radars can be used as a standalone sensor for object detection despite sparsity and low resolution.

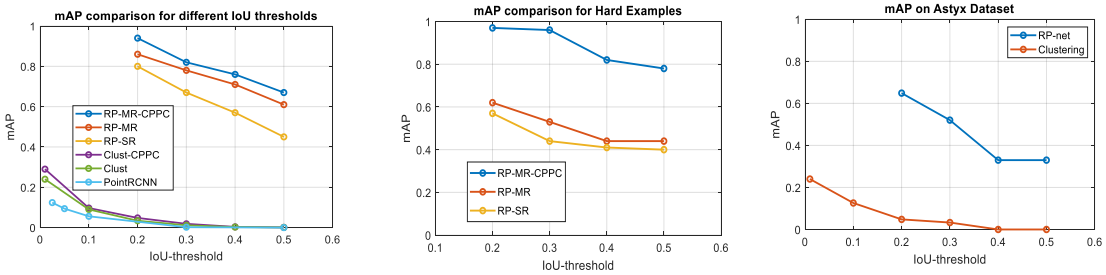


Figure 1.14. a) RP-net outperforms PointRCNN and clustering baselines for bounding box estimation. RP-MR performs better than the RP-SR system. CPPC based combination further improves performance (RP-MR-CPPC). b) CPPC based fusion performs significantly better in hard examples due to noise estimation and prior heading direction using space-time coherence. c) mAP performance of RP-net on Astyx dataset compared to the clustering baseline.

refinement network, NMS is performed on the boxes to remove boxes with more than 0.5 2D IoU overlap using predicted confidence scores. The training happens simultaneously for both region proposal and the refinement network. However, at the start of the training, the classification head can not provide correct confidence scores; hence, for the initial 30 epochs, we use 2D IoU scores with ground-truth boxes as confidence values for NMS. Note that the network is already trained during the testing phase, and we do not need these IoU values.

1.8 Evaluation

In this section, we would comprehensively evaluate each part of Pointillism. We would compare our results against the commonly used deterministic clustering [128] and deep learning based approaches [101] for 3D bounding box estimation on point clouds. Figure 4.7 shows examples of predicted 3D bounding boxes by Pointillism.

Testing and Training data-set: We test our system on the dataset collected by us (section 6) that contains 54000 radar frames. We use a train test split of 9:1. The data used for testing comes from separate data collection runs than training data to ensure the generalization of our approach. We also use the publicly available astyx dataset [76] for automotive radars to demonstrate the generalization of our approach. We also compare the performance of Pointillism against LiDAR in bad weather conditions.

Following are the metrics we use to evaluate our system:

- **IoU**: IoU (Jaccard Index) is a measure of the overlap between the predicted bounding box and the ground truth box. 3D IoU is given by $\frac{\text{Intersection Volume}}{\text{Union Volume}}$. 2D IoU is defined for the top view (also called Bird-eye-view (BEV)) rectangles of 3D bounding boxes, as $\frac{\text{Intersection Area}}{\text{Union Area}}$. Two equal-sized boxes with half overlap would have an IoU of 0.33. Hence, even an IoU of around 0.5 is generally regarded as a good overlap.

- **mAP**: (mean Average Precision) mAP is the area under the precision-recall (PR) curve, which is a measure of the number of actual boxes detected (recall) along with the accuracy of detections (precision). Specifically, precision is obtained for incremental recall values to get PR curve.

$$Precision = TP/(TP+FP)$$

$$Recall = TP/(TP+FN)$$

$$mAP = Area(\text{precision-recall curve})$$

An estimation is regarded as a true positive (TP) if it is above a particular IoU (Intersection Over Union) threshold. Note that a higher recall rate can easily be obtained by predicting a large number of boxes, but at the cost of sacrificing precision (more False Positives (FP)) and vice-versa. A higher mAP means better performance on both accuracy (precision) and exhaustiveness (recall) of estimation. An FP could also be obtained because of noise. An FP generated due to noise will have a very small (almost 0) IoU with any ground truth box.

Hence, in a lower IoU threshold regime, the mAP is more sensitive to the amount of noise and allows us to better compare our noise suppression performance.

In summary, Pointillism achieves a median error of less than **37cm** in localizing the center of an object bounding box and a median error of less than **25cm** in estimating the dimensions of the bounding boxes. We use 2D IoU (BEV IoU) as the thresholds for our mAP metric [38]. Pointillism achieves an mAP score of **0.67** for an IoU threshold of 0.5 and a score of **0.94** for a lower IoU threshold of 0.2, which is a **45%** improvement over a single radar system. We further show that RP-net is easily generalizable to other datasets by evaluating its performance on *Astyx* dataset [76] and achieve an mAP of **0.65** with IoU threshold of 0.2.

1.8.1 Performance on Bounding Box estimation

Our entire system (dubbed RP-MR-CPPC) consists of multi-radar (MR) fusion to create Cross Potential Point Clouds (CPPC) and RP-net to estimate 3D bounding box. To individually compare the mAP performance of CPPC and RP-net we define the following baselines:

- **RP-SR:** RP-net on single radar data.
- **RP-MR**: RP-net on multiple radar data without Cross Potential Point Clouds fusion. The point clouds from multiple radars are simply added in the global coordinate system.
- **Clust**: A clustering based bounding box estimation baseline. A predefined size bounding box is estimated for each cluster found using DBSCAN, coupled with angle estimation using *Principle Component Analysis* [123]
- **Clust-CPPC**: The clustering based approach used on Cross Potential Point Clouds.
- **PointRCNN**: Official implementation of well-known LiDAR based 3D bounding box estimation network PointRCNN [101] used on our collected dataset.

Figure 1.14a shows the overall performance of these systems. The X-axis is the IoU thresholds used for the mAP values. Further, to best examine the performance improvement

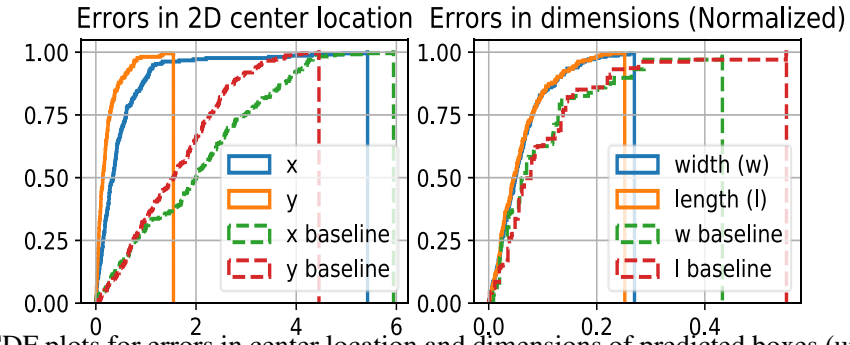


Figure 1.15. CDF plots for errors in center location and dimensions of predicted boxes (upwards is better). The x,y-coordinate is the depth and lateral dimensions, respectively. Compared to the clustering-based baseline, RP-net performs much better as it learns refinement from the data.

brought in by CPPC, we also evaluate the performance on a subset of validation set that contains hard examples following KiTTi evaluation framework [38]. Hard examples are characterized by point clouds containing more than one-fourth of the points coming from noise, and the cars are undertaking complex maneuvers (sharp turns). The results of hard examples are shown in Figure 1.14b.

CDF comparison. In this experiment, we evaluate the performance of Pointillism on estimating the center coordinates of bounding boxes and its dimensions in 2D Bird-Eye-View (BEV). Proper estimation of bounding box parameters in BEV is critical for autonomous driving[38]. Figure 1.15 shows the CDF plots of L1 absolute errors in center coordinates and dimensions of bounding boxes predicted by RP-net compared against the clustering-based baseline (clust-CPPC[123]). Errors in dimensions are normalized to actual ground truth box dimensions. Note that the clustering baseline does not have the refinement stage as in RP-net, and hence Pointillism clearly outperforms the baseline. RP-net can learn the complex point cloud to bounding box mapping function (equation 1.3) using data and hence provides accurate bounding box estimation results.

Performance of RP-net. Figure 1.14 shows that RP-net performs significantly better compared to LiDAR-based network (PointRCNN) and the clustering baseline. This result is expected as a LiDAR-based network can not be directly applied to radar point clouds due to inherent differences in point cloud distribution. LiDAR-based networks are designed to work in cases where background points are much larger compared to the foreground(points on objects of

interest). Also, the poor performance of the clustering baseline conveys the complexity of the task and the requirement of data based learning. The performance for the clustering baseline increases with lower threshold values because it lacks any refinement of the bounding box to get the correct dimensions of the box. RP-net architecture is specifically designed to handle the sparsity and non-uniformity of radar point clouds by learning to model their intricacies and using a special set of anchor boxes.

Effect of using multiple radars. Multiple radar fusion based approach (RP-MR) outperforms the single radar baseline (RP-SR) in all the cases (Figure 1.14). The mAP value increases from 0.45 to 0.61 for IoU threshold 0.5 as we move to multiple radars from a single radar. The reason behind this is the specularity of mmWaves that prevent all the features from being captured and hence makes it hard to get the dimensions of the box right, inherently affecting the mAP values. Using multiple radars decreases the adverse effect of specularity, as evident from the increase in mAP values from single to multiple radar approaches.

Performance with multiple cars We evaluate the performance of different networks in detecting the vehicles present in the scene. Table 1.2 shows the recall rate achieved with the different number of cars present in the scene. The performance drops when the number of vehicles in the scene increases due to occlusions. Still, RP-MR-CPPC outperforms the single radar non-CPPC based models in all the cases, which proves the superiority of our algorithm over baselines.

Table 1.2. Recall rate for different number of cars in the scene.

Model vs # Cars	**1**	**2**	**3**	**4**	**Entire Dataset**
RP-SR	0.4	0.31	0.2	0.1	0.32
RP-MR	0.61	0.45	0.3	0.28	0.47
RP-MR-CPPC	**0.75**	**0.52**	**0.41**	**0.38**	**0.6**

Effect of using CPPC. Figure 1.14a shows that using a CPPC based fusion approach is better than just a naive combination of multiple radar point clouds (RP-MR). In the hard examples (figure 1.14b) the effect is even more significant. The performance increases from **0.44** to **0.78** by using CPPC on multiple radars for IoU threshold of 0.5. Note that even for the

Table 1.3. Ablation studies: change in performance with different input channels to the network.

Model	Velocity	Intensity	CPPC	mAP
Ablated Model	✗	✗	✗	0.56
	✓	✗	✗	0.60
	✓	✓	✗	0.61
Final Model	✓	✓	✓	**0.67**

clustering based baseline, using CPPC fusion improves the results for bounding box estimation. The space-time coherence in CPPC reduces the noise, thereby minimizing false positives and improving results.

Comparison to other approaches (LiDAR/sensor fusion). In order to put these results into perspective, we provide some results from LiDAR-based approaches. PointRCNN [101] achieves 0.78 mAP on KiTTi LiDAR dataset [38]. A camera and radar fusion approach, on a dataset provided by Astyx GMBH [76], achieves an mAP value of 0.45 on the complete dataset. Although it is not fair to directly compare with the LiDAR results, as the underlying datasets and IoU thresholds used are different (LiDAR-based approaches use a higher IoU threshold of 0.7), but we want to emphasise on the fact that it is possible to achieve LiDAR or camera like perception using just radars. Moreover, a higher resolution radar can produce a denser point cloud, that would further increase the IoU performance using our method.

Table 1.4. Effect of model size on time and performance

Model	mAP (IoU>0.5)	Time
256 channels	0.50	0.018 seconds (55 fps)
1024 channels	0.68	0.0208 seconds (48 fps)

1.8.2 Model Generalization on Astyx dataset

In order to evaluate the generalization of our network, we test our network on the recently released Astyx dataset [76]. The released data set contains only 545 scene point clouds, which are quite less in order to allow efficient training. Also, the dataset contains only a single radar, which prevents us from applying our entire processing chain to the dataset. We divide the dataset into 5% test (and rest training) set and evaluate the object detection performance. RP-net is able to achieve a performance of **0.65** mAP on this dataset with an IoU threshold of 0.2 while the

clustering baseline achieves a significantly lower mAP of 0.05, as shown in Figure 1.14c.

1.8.3 Effect of input channels

In this experiment, we would perform the ablation study for adding different channels to our input. We report the mAP score changes with the incorporation of a different number of input channels. Table 1.3 summarizes the performance of this experiment. As expected, adding the velocity as an input channel brings the performance improvement as it provides the context about the direction of the vehicle's movement. Adding the intensity values do not have much effect on the performance. We hypothesize that as the target objects are only cars, there isn't much difference between the intensity of reflections. Lastly, using Cross Potential Point Clouds based fusion further improves the performance to mAP score of 0.67 because of space-time coherence.

1.8.4 Model Architecture validation

In this section, we evaluate the effect of model size on the performance. We change the number of channels for the representative feature vector of anchor boxes and compare the performance. Table 1.4 shows the result of our analysis. This experiment sheds light on the trade-off of time and accuracy. Clearly, with the increase in the number of layers from 256 to 1024, the model performs better (0.18 increase in mAP) but lags on time (0.01 sec slower).

One of the major requirements of an autonomous perception system is real-time performance. Our best model makes inference at 50fps while running on an NVIDIA GTX 1080 Ti GPU(Table 1.4). Normal human reaction time is 100ms, which means that our model can provide 5 different scene perception outcomes until a normal human reacts to an event. Thus, Pointillism has huge potential to be deployed on a real-time autonomous vehicular system.

1.8.5 Performance in bad weather

The main objective of using radar as a primary sensory modality is to enable all-weather perception. To compare the performance of different sensing modalities in adverse weather

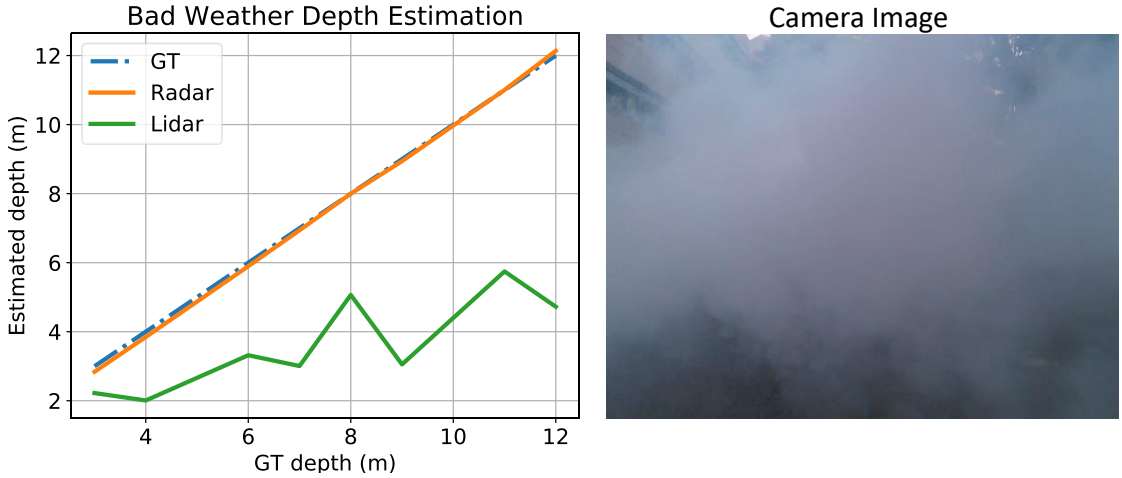


Figure 1.16. Performance comparison of different sensors in the presence of adverse conditions. The left plot shows the depth estimation performance of Radar and LiDAR for an object directly in front of the sensor in the presence of fog. The right figure shows the camera image for the experiment.

scenarios, we perform experiments in foggy conditions. We use an artificial fog generator to simulate adverse weather conditions and estimate the performance of the depth estimation of a single-vehicle present in front of the sensor. The first point of return is regarded as the predicted depth. Figure 3.9 shows the performance comparison between LiDAR and Radar (the results are an averaged over 100 frames). LiDAR and camera get severely affected in the presence of fog, which would pose severe problems for tasks like depth estimation. Radar, on the other hand, remains unaffected. Moreover, past work [24, 60] conduct large scale experiments using fog chambers and artificial rain generators to show the prominence of radar sensors compared to LiDAR sensors in adverse weather conditions.

1.9 Related work

The techniques designed in Pointillism are fundamental and could be used to improve LiDAR, ultra-sonic, and infra-red imaging other radar applications. Spatial separation analysis could be used for other applications to identify optimal separation. Similarly, Cross Potential Point Clouds and RP-net can potentially be extended to benefit all the radar sensing applications relying on radar point clouds. Our work is closely related to the work on the following:

MIMO radar and Multi-radar fusion. Radars have been long-standing sensors in the field of surveillance and detection. Past works have considered the fusion of multiple radar sensors for aerial surveillance [35, 124, 55, 86] but do not have much relevance to self-driving scenarios involving high-resolution MIMO radars. In [86], a nearest neighbor based association scheme is given, which is again limited to aerial surveillance. Also, the presence of ghost objects is not taken into account, which is a major part of real-world automotive radar data. We perform all our analysis on real data of dynamic traffic environments. A multi radar setup is considered in [94] with a common field of view. They use *Labelled Multi Burnoulli* filter to track and associate the detections in multiple radars but fail to leverage the multi-sensor input or optimize spatial separation. They instead treat it as a single sensor. A multi radar system deployed in [75], compares performance with a single radar system for the task of blind-spot detection using Kalman filter based tracking. They use multiple radars to achieve a wider FoV by simply combining the detections generated from two radars and could not report any significant performance gains. Pointillism takes advantage of multiple radars for increasing the point cloud density while minimizing noise recorded in the scene, as opposed to getting a wider FoV.

Synthetic Aperture Radar Synthetic Aperture Radar(SAR) [36, 39, 40] based techniques are used to increase the resolution of radar, but they have limitations. SAR approaches can not resolve the specularity problem of mmWaves, which requires a different viewpoint as provided by multiple radars in Pointillism. For a vehicle moving on the road, a SAR based approach needs to be implemented using the motion of the car as done in [36, 39], instead of physically moving the radar. This approach to SAR only increases the resolution on the sides of the ego vehicle (side-looking SAR). Applications of this method include parking spot detection, but object detection in front of vehicles is not improved. A forward-looking SAR is implemented in [40] by using a beam scanning 300 GHz radar, which is different from the 77 GHz MIMO radar we used. This approach tries to improve angular resolution at small boresight angles. It does not take care of noise generated by the multipath and random clutter, which is handled by Pointillism's CPPC formulation. Pointillism's approach for multi-radar fusion generalizes to the

fundamental distribution of radar point clouds and could be applied to all the above work and reap significant benefits for tackling specularity, noise cancellation, and sparsity.

Deep Learning based 3D bounding box. With the abundance of open-source LiDAR datasets [38, 19, 11] for autonomous driving, object detection on LiDAR point clouds has seen numerous advancements. Point clouds are converted into voxel grids for object detection with bounding box prediction in [72, 88]. Direct object detection on point cloud data is performed in [87, 121]. However, they use calibrated 2D RGB images to segment instance-based frustums from a point cloud. A camera-radar fusion approach for bounding boxes is provided in [78] but would be infeasible for all-weather perception tasks. PointRCNN [101] performs 3D bounding box estimation only using LiDAR point clouds. However, due to the sparsity of radar point clouds, the approach fails to capture enough context for segmentation. RP-net (Our approach) combines the region proposal and segmentation to overcome the sparsity problem. Andreas et al. An approach based on frustum pointnet is followed in [25], where patches are considered throughout the 2D image followed by a segmentation head that identifies the radar targets lying on the object of interest. However, the patch proposal can not be readily extended to 3D detection.

Radar sensing for health-care and indoors. FMCW is a popular technique for RF sensing. FMCW based radars (WiFI frequency) have also been used for indoor health monitoring and localization applications [129, 13, 12, 52, 127, 109, 51, 53]. Low-frequency radars (WiFI frequency) have been used [130, 131, 132, 109] for tasks like human pose estimation, fall detection, and sleep stage detection using deep learning. All these works use low-frequency FMCW based and can not be readily applied in outdoor settings. mmWave radars have been used indoors for ego-motion estimation and corridor mapping [68, 14]. The techniques cross-potential point cloud and RP-net can be used by the indoor radars as well, as they are complementary to all existing work and provides a neat framework that is inspired by the distribution of the radar point clouds. For example, reflection tracking is a key technique for all the above work. One could apply Pointillism's cross-potential based ideas to the reflections and their tracking to make them robust against noise and dynamic multi-path [12].

Automotive Radar data processing using deep learning. Deep learning has been applied to various stages of radar data processing. Past works [112, 67] perform object classification on radar data using CNNs. Pointnet++ [89] is used directly on radar point clouds to perform semantic segmentation in [96]. These approaches do not consider the major challenges of specularity and sparsity with radar point clouds. A GAN based approach[44] has been proposed to image cars using synthetic aperture based sensing but is limited to static scenes. 2D box prediction using radar tensors for the highway has been done in [71]. Their approach is only limited to traffic scene where vehicles are moving straight and can not retrieve the 3D pose of vehicles. The special distribution of radar point clouds requires an approach specifically designed for radar data, such as CPPC. We believe Cross Potential Point Clouds can help existing work to improve the performance significantly.

1.10 Limitations and Future Work

Pointillism presents a system that enables object detection using only automotive radars. It tackles the fundamental problems of noise, specularity, and sparsity present in radar point clouds by using multiple radars at once. The radars used during the evaluation of our system offer an angular resolution of 15 degrees. Current high-resolution automotive radars provide an angular resolution of 1 degree that would potentially lead to better bounding box refinement and hence better IoU values. This means that the current mAP performance of pointillism at higher IoU thresholds (section 1.8) could further be improved by using higher-resolution radars. A natural extension of this work is the time-based tracking of dynamic objects present in the scene. We describe how we use space coherence to eliminate noise and time coherence to get a pose estimate. A tracking based formulation could possibly combine the noise filtering with pose estimation by looking at time-based data from multiple radars. Secondly, the radars also provide doppler estimates of the objects. We use doppler as an input channel to RP-net. However, it could also be used to classify different types of road users like pedestrians, bikes, and cars. The concepts introduced in our paper relate to the fundamental nature of radar point clouds and can

be extended to enable more complicated tasks than object detection.

Chapter 1, in full, is a reprint of the material as it appears in "Pointillism: Accurate 3D Bounding Box Estimation with Multi-Radars" by Kshitiz Bansal, Keshav Rungta and Dinesh Bharadia, which appears in the 18th ACM Conference on Embedded Networked Sensor Systems (SenSys '20), November 16–19, 2020, Virtual Event, Japan. ACM, New York, NY, USA. The dissertation author was the primary investigator and author of this paper.

Chapter 2

RadSegNet

2.1 Introduction

Rapid research in self-driving sensing systems has significantly improved the quality of perception tasks like object detection in the past few years. Despite these advancements, level 4 or 5 self-driving ability in commercial vehicles is not commonplace. A major reason behind autonomous cars not being more commonplace is their dependence on lidars, cameras, or their fusion which cannot perform reliably in cases of occlusions and adverse weather conditions [37]. This shortcoming in cameras and lidars has sparked a major interest in automotive radar-based sensing, particularly in camera radar fusion systems [80, 57, 81].

A camera radar fusion system ideally can combine the benefits of both cameras and radars while also addressing the shortcomings of each sensor. While cameras provide rich texture and semantic information, they start failing in cases of long range, occluded objects and adverse weather conditions [37, 20]; and while radars are capable of providing all-weather reliance, long range and occlusion-free detection [37, 83, 50], they struggle in clearly identifying objects due to lack of rich texture and semantic features [16]. In this work, the primary question we try to answer is, how do we fully realise the advantages of both modalities for achieving accurate and also reliable object detection.

An effective sensor fusion involves combining the strengths of multiple sensors while mitigating their limitations. Previous approaches that predominantly rely on the camera's

perspective view, where the radar data is projected onto the camera image, use radar only as a secondary aid. Such methods have significant limitations, particularly in scenarios where objects are occluded in the camera image and hence fail to leverage the full potential of radar data. Advanced state-of-the-art techniques, such as AVOD-fusion [57], utilize multi-view fusion, combining both camera perspective and radar bird's eye view (BEV) as separate input branches. This fusion method heavily relies on the quality of camera images, and does not consider scenarios where the camera's reliability is compromised, such as when objects are occluded or in adverse conditions. Consequently, this can result in a considerable decrease in the overall performance of the system. Fig. 2.1, shows how the quality of detections from AVOD-fusion degrade, when the camera is subjected to artificially generated adverse weather.

Undoubtedly, there is a need to improve the reliability of camera-radar systems to achieve good performance even when camera input is degraded. In order to achieve this goal, we argue for a fundamentally different approach for fusing information from radars and cameras. We propose that if we can extract the useful information from both the sensors independently, then we would get the advantages of both modalities and break the dependence on camera in case of unreliability. This new philosophy of fusion uses the fact that the sensors provide different pieces of information, hence it is possible to extract information independently, thereby improving the reliability of the entire system.

In this work, we propose RadSegNet, which realises a system that breaks the dependency between the two sensors. RadSegNet, that achieves the required functionality by using mainly two design principles. First, we note that rich textural information in cameras is mainly used to clearly identify the instances of foreground objects like vehicles out of the background clutter. Hence, we bank on the significant advancements made in the camera semantic and instance segmentation literature to first independently extract the semantic information from camera RGB images. Second, we note that for radars, BEV representation offers several advantages over perspective view [120] especially in case of occlusion. Hence in RadSegNet, we exhaustively encode all the information present in radar point clouds in a BEV representation. Overall, in

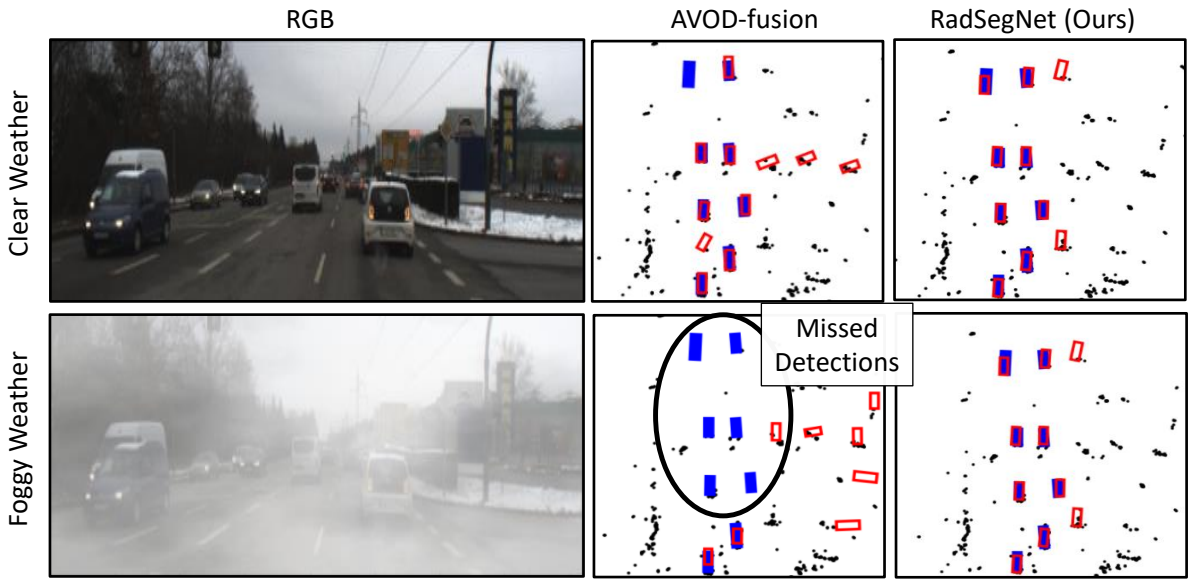


Figure 2.1. Performance of radar-camera fusion architectures when camera input is augmented with artificial fog. AVOD-fusion [57] gets significantly deteriorated while our method continues to provide robust results even in fog. Filled blue boxes ■ are ground truth and red empty boxes □ are prediction results.

RadSegNet, we first independently extract the semantic information from camera RGB images and then use it to augment radar data and create a BEV representation. Finally, the detection is performed only on this augmented BEV radar representation, thereby breaking the dependence on the camera. Even when camera input is unreliable, RadSegNet continues to work reliably using the radar data (Fig. 2.1). Note that these unreliable conditions include both adverse weather conditions as well as occlusions and long-range in clear weather, where camera data becomes poor.

A major challenge in propagating the semantic information extracted from the camera to radar data is the lack of depth information in camera. This prevents us from directly projecting the camera pixels on the radar point cloud. To address this issue, we developed a new method called semantic-point-grid (SPG) representation, which encodes semantic information from camera images into radar point clouds. Rather than projecting the camera image onto the radar point cloud, SPG finds the most probable camera pixel correspondence for each radar point. Although projecting radar points onto a camera image using intrinsic parameters is possible, the

noisy nature [16] of radar point clouds and errors in sensor calibration leads to inconsistencies in associating the correct pixel from the camera image with the corresponding radar point. SPG overcomes this issue by using a novel Instance-Informed Weight (IIW) module that inhibits inaccuracies resulting from poor projection. The IIW module in SPG encoding utilizes the spatial separation between the correct and incorrect projection for each object instance to reduce the focus on the points associated with incorrect semantic information.

We evaluate our approach on two publicly available radar datasets with different types of radars. For comprehensive benchmarking we use Astyx [77] dataset that has radar point clouds and augment it for bad weather. For adverse weather testing on real world data, we use the RADIATE [99] dataset that has dense radar data. For the task of object detection, RadSegNet sees an average precision (AP) improvement of 27% in Astyx dataset and 41.46% in RADIATE dataset, when compared to the state-of-the-art (SOTA) approach of camera-radar fusion, AVOD-fusion [57]. More importantly, in cases of adverse weather conditions, where camera images have deteriorated, SOTA could see more than 50% AP degradation, while RadSegNet degrades less than 6%. Figure 2.1 shows the robust performance of our approach compared to AVOD-fusion [57] when fog is inserted into the images from Astyx dataset. To summarize, our approach to radar-camera fusion is universal (works with any type of radar), reliable (maintains reliability in all-weather conditions), and complete (completely use all advantages of radar sensing). It can also be readily integrated with other sensor-fusion approaches as it does not have any dependency in the feature extraction stage.

2.2 Related Work

Projection based

Felix et al. [81, 80] project the radar point on the camera plane and augment them into vertical lines and pillars respectively to encode height. Chang et al. [21] use a spatial attention network to process projected radar image. Grimm et al. [43] use a differentiable warping function to warp radar tensor to camera image for using camera labels for training. Operating in

perspective view makes it harder to distinguish between small objects close to the sensor and large objects at a longer range, hence achieving sub-optimal performance. Millieye [102] and BIRAnet [125] are restricted to detecting objects in 2D image plane, which is not sufficient for downstream tasks like path planning. Lim et al. [65] use planar homography transformation to project camera image to radar BEV but an inverse mapping of camera to BEV plane is ill-defined due to the lack of depth information in camera images causing ambiguities in detection.

Region proposal

This class of methods tend to use radars for generating region proposals for performing object detection. Nabati et al. [79] use radars for 2D proposals to improve detection in autonomous driving, while Kowol et al. use radar detections to prune predictions and improve performance. However, both methods only partially utilize the advantages of radar by using it to assist cameras.

Multi-view feature aggregation

In these methods, a set of object proposals are projected separately to camera and radar plane to extract features. Kim et al. [57] and GRIFnet [58] fuses these features to obtain final predictions, while CRAMnet[54] uses these features to create a 3D point cloud and refine the final detection for each object. These approaches yield SOTA results for camera-radar fusion, but the performance is unreliable when camera input deteriorates in adverse conditions.

Lidar-radar fusion

Wang et al. [118] uses a lidar to identify ghost points in the radar data and [126] uses lidar as the primary sensor while being assisted by low-resolution radar for object detection. As the main focus of our work is on radar and camera fusion, fusion with lidar remains out of scope. However, such refinements can still be used in tandem with RadSegNet's radar camera fusion to further improve performance.

2.3 Radar Primer

At their core, radars use the same time of flight (ToF) analysis of reflections to generate points, as in LiDARs, but they differ in the wavelength of operation. While, lidars use nanometer wavelength signals, which provide them with very high resolution due to surface scattering, radars use millimeter wavelengths where the reflected power is divided between specular reflections and diffused scattering [16]. The raw radar data is dense, but contains background thermal or multipath noise [17]. Radar data is also commonly subjected to Constant False Alarm Rate (CFAR) filtering, which generates a light-weight and sparse point cloud output [77]. Consequently, object edges are not as clearly defined in radar point clouds as they are in LiDAR point clouds. For example, in radar point clouds, a point cluster originating from a wall could have a similar spatial spread as that of a cluster originating from a car. This effect makes it challenging to learn any shape based features directly from radar point clouds to distinguish objects of interest (Cars, pedestrians, etc) from background objects. Figure 2.1 shows the non-uniformity present in radar point clouds.

However, at the same time, radars also offer the following unique advantages due to the millimeter band transmission: a) They provide a *longer range* than lidars, because larger wavelength signal has a lower free space power decay rate allowing radar waves to travel over longer distances [37]. b) They can *see-through occluding vehicles* because their signals bounce off the ground allowing them to also sense vehicles which are completely occluded [83, 50]. c) they are *all weather sensors* because the larger wavelength of millimeter waves allows them to pass unaffected through adverse conditions like fog, snow and rain [16].

2.4 Methodology

2.4.1 Input representation for Radar data

The input representation of the sensor data has a significant impact on deep learning architecture's performance for object detection tasks. Specifically for radar data, high sparsity and

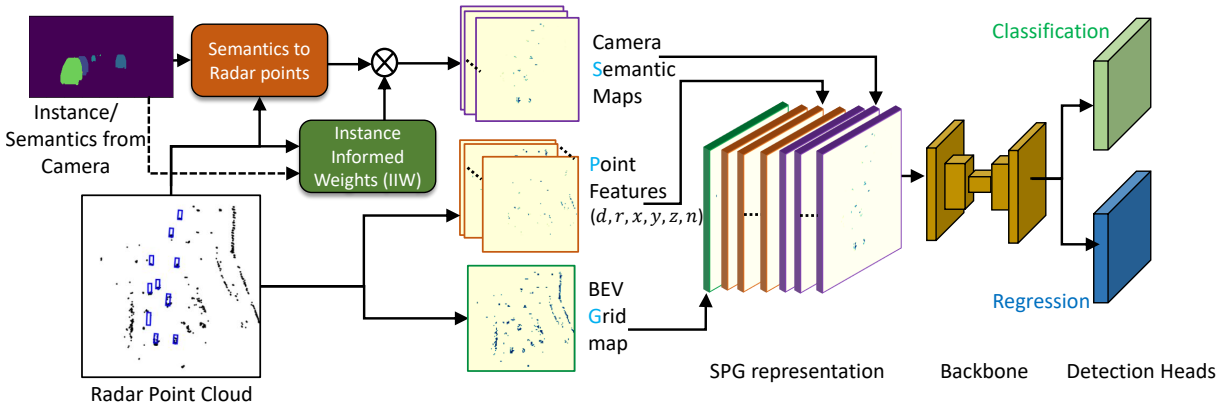


Figure 2.2. Overview RadSegNet: Our approach utilizes encodings from our SPG module to detect objects. The encodings are generated from the semantic features from a semantic segmentation network along with radar point-based features and an occupancy grid. These encoded maps are concatenated and passed through the bounding box detection network.

non-uniformity make it extremely crucial to choose the correct view and feature representation. Past work[120] has shown that performance gains can be obtained by just transforming data from a perspective camera view to a 3D/BEV view. The reason behind this is that in the perspective view, there is scale ambiguity with depth as well as object overlap due to occlusions. BEV representation, on the other hand, is able to clearly separate objects at different depths, offering a clear advantage in cases of partially and completely occluded objects [120].

For radars, BEV becomes an absolute necessity, as they even get signals from occluded objects due to the radio waves bouncing off of the ground (section 2.3). Representing radars in perspective view to extract features is not only sub-optimal but may also cause confusion in the case of occluded objects. Hence, for getting a good and reliable performance, RadSegNet uses BEV representation as input but at the same time, also retains the necessary information from the radar point cloud, required for optimal object detection.

Occupancy grid

In order to generate a BEV representation, we project the radar points onto a 2D plane by collapsing the height dimension. The plane is then discretized into an occupancy grid. Each grid element is an indicator variable that gets a value of 1 if it contains a radar point otherwise it is represented as 0. This BEV occupancy grid also preserves the spatial relationships between the

different points of an unordered point cloud and stores radar data in a more structured format [61].

Radar point features

The BEV occupancy grid provides an optimal representation for radars and provides order to the un-ordered radar point cloud. However, naively creating a BEV grid also discretizes the sensing space into grids which dissolves the useful information required for the refinement of bounding boxes. In RadsegNet, we retain that information by adding the point based features to our BEV grid as additional channels. Specifically, we add the cartesian coordinates, doppler and intensity information as additional features. The BEV grid input to the network is then defined as follows:

$$s := (\mathscr{I}, d, r, x, z, y, n)(u, v) \tag{2.1}$$

Here, $\mathscr{I}$ represents the 2D occupancy grid where each grid element is parameterized as (u, v). All the positions in $\mathscr{I}$ where radar points are present store 1 or else 0. d and r represents the doppler and intensity value of radar points. They help identify objects based on their speeds and reflection characteristics. (x, z) is the average depth and horizontal coordinate in the radar's coordinate system. In order to encode height information, we generate height histograms by binning the height dimension (y) at 7 different height levels and creating 7 channels, one for each height bin. The cartesian coordinates (x, y, z) help in refining the predicted bounding box. The n channel contains the number of points present in that grid element. The value of n is usually proportional to the surface area and reflected power which helps in refining bounding boxes. An overview of all the point features is shown in figure 2.2.

2.4.2 Fusion with Camera Data

The BEV occupancy grid, along with the radar point features, represents all the information in radar point clouds in a well structured format. Now, we need to add camera information to

this representation to complete our fusion system. Note that a direct projection of camera data to the BEV is non-trivial and challenging as camera lacks depth information. To solve this problem, current state-of-the-art radar-camera fusion systems [57] simultaneously extract features from both modalities. The features are then fused on a per-object proposal basis. However, in this approach, the performance drops significantly when the camera data is unreliable for any object, in case of occlusion or adverse weather (drop of more than 50% in some cases, refer section 2.6).

In RadSegNet, we define a novel SPG (Semantic-point-grid) encoding that solves the above challenge by independently extracting information from cameras, in a reliable way. In the next section, we provide details of our SPG encoding and how it makes use of all the advantages present in both modalities while being reliable in cases of camera uncertainty.

Capturing semantics with SPG

The rich texture and semantic information in camera images is very useful for understanding a scene and identifying the objects in it. This information complements well with the radar, where non-uniformity in point clouds makes it harder to learn features that can identify the objects well (section 2.3). Our key insight for using this complementary nature while maintaining reliability in adverse conditions, is to first extract the useful information from camera images in the form of scene semantics and then use it to augment the BEV representation obtained from radar. In contrast to fusing the features on a per-object basis, our approach keeps a clear separation between information extraction from two modalities, hence performing reliably even when one input is degraded. We use a robust pre-trained instance segmentation network to obtain semantic masks from camera images of each object instance present in the scene. However, we still need to add this information to the radar BEV without the presence of depth information for the camera image.

To associate camera based semantics to radar points, we create separate maps for each output object class of the semantic segmentation network. These maps are of the same size as the BEV occupancy grid and get appended as semantic feature channels. To obtain the values of the

semantic feature channels for each grid element, we first transform the radar points to the camera coordinates using camera intrinsic parameters. Next, we find the nearest pixel in camera image to the transformed point and use the semantic segmentation output of that pixel as the values of semantic feature channels in SPG. For multiple radar points belonging to same grid element, an average is taken over semantic values. These feature channels contain the semantic information extracted from the camera, helping in the object detection from radar BEV occupancy grid. They effectively reduce the possible false positive predictions generated by radars, as the radars may get confused in identifying objects due to inherent non-uniformity (section 2.3) in radar data. Figure 2.3 (bottom middle) shows an example of how the semantic features are encoded with the radar BEV grid, for the class car.

Instance Informed Weights (IIW) A crucial aspect to take into account while associating radar points with camera pixels is the noise present in radar point clouds and errors in sensor extrinsic calibration. These noise and errors makes it challenging to correctly associate the radar point with the corresponding pixel in the camera image. Specifically, there would be cases where because of these errors, a point belonging to a background object like a building, gets projected on a foreground object like a car, which is in the same line of sight. In this case the point gets incorrectly associated with the semantics of the foreground object. To correct such mis-associations due to incorrect projections, we use a weighting scheme called IIW (instance informed weighing, section 4.2.1) in our SPG encoding.

As the first step in SPG, we project radar points onto an instance segmentation map, where each object in the scene has its own instance mask. In doing so, for any given object, we have a list of points that gets projected on its instance mask, consisting of both correct and incorrect projections caused by noise. These points are all assigned the same semantic information corresponding to that object's instance class and also an instance ID unique to that object instance. Now, the problem of identifying the mis-projections corresponding to an object instance is reduced to finding the outliers in the subset of points having the same object ID and assigning them a lower importance weight. The idea behind IIW is based on the insight that the

number of mis-projections would likely be lesser than the number of correct projections as the mis-projections tend to happen mostly near the object edges.

To get the weight for a point n, a voting mechanism is used. We define the neighbourhood of the point n with α. For a neighbouring point within the radius of α we add 1 and for a point outside radius α we subtract 1. Mathematically, we can achieve this using tanh function:

$$S = \sum_i -tanh(k_2(\|d_n - d_i\| - \alpha))$$

We use k_2 as a hyperparameter to tune the sharpness of $tanh$. Here, d_i is the cartesian coordinate of point i and $\|*\|$ denotes the l_2 norm. When $(\|d_n - d_i\| - \alpha)$ is positive, the $-tanh(*)$ outputs a value closer to -1; and when it is negative, its value is closer to 1. Since we are trying to correct the incorrect projections per object, we only need to sum over the points selected using that object's instance mask (or the points with the same instance ID). So, each term in the sum is multiplied by $\mathbb{1}(p_n = p_i)$, an indicator function to identify points belonging to same instance ID p_n as that of point n. Finally, to convert the value to the interval (0,1) we use the $sigmoid(S[*])$ function. Hence, the final weight value w_n value becomes:

$$w_n = S\left[k_1 \sum_i \left(tanh(k_2(\alpha - \|d_n - d_i\|))\mathbb{1}(p_i = p_n)\right)\right]$$

k_1 is another hyperparameter to keep the value of weights close to 0 or 1. We implement the IIW using pytorch functions and use *torch.cdist* function to calculate the distances between all pairs of points. Figure 2.3 shows how the IIW weighing scheme corrects for the mis-projections in SPG encoding.

Note, the reason that RadSegNet truly achieves independent feature extraction is because we don't filter out any radar points, while making better use of the advantages that both modalities

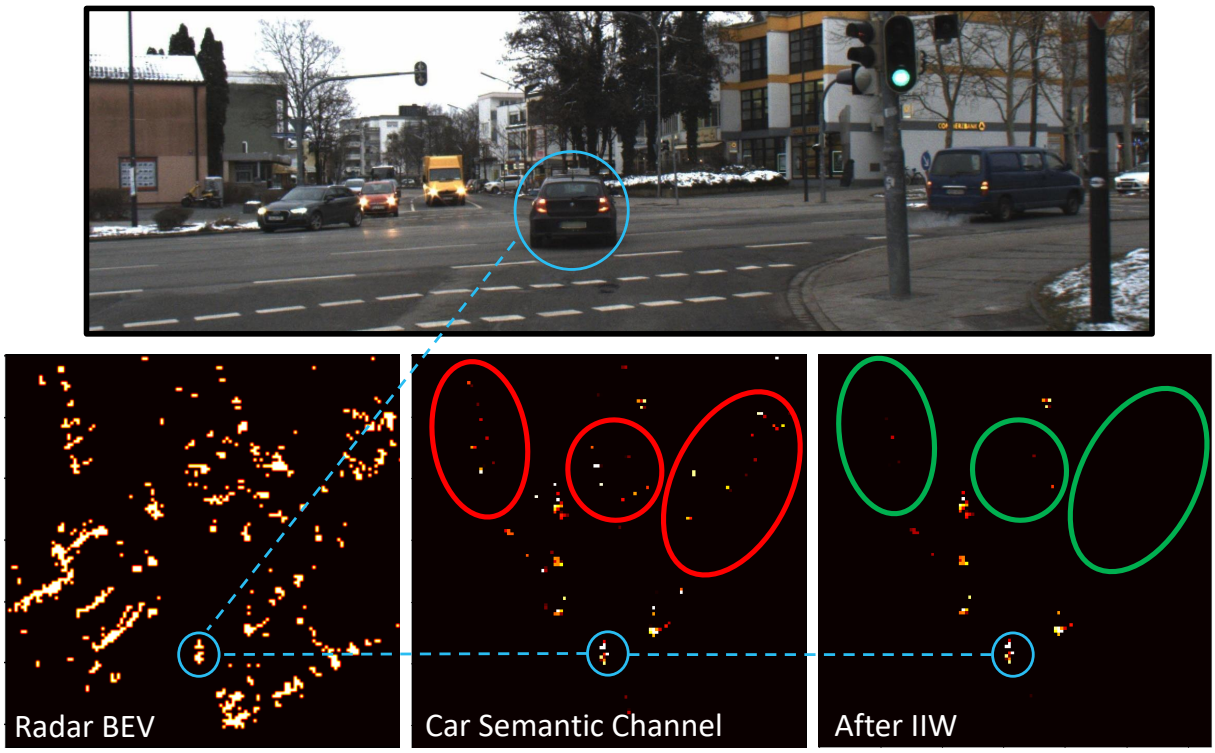


Figure 2.3. Adding the semantic channel in SPG encoding helps identify points belonging to objects of interest. The figure shows an example car represented in the semantic map and the corresponding BEV occupancy grid (blue circles). The bottom right plot shows how our IIW module cleans the semantic map for incorrect projections (Red and green ellipses).

bring to the table. This means that in cases where the camera based features become less informative, all objects in the scene are still visible to radars, which prevents any drastic drop in performance. The textural and high resolution information from cameras is condensed into semantic features which assists the all-weather, long range and occlusion robust sensing of radars.

2.4.3 Bounding Box prediction on SPG features

We use SPG encoding to generate BEV maps, which are fed into a neural network for feature extraction and bounding box prediction. As our backbone, we use and encoder-decoder network with skip connections that has 4 stages of down-sampling layers and 3 convolutional layers at each stage. This allows us to extract features of different scales and combine them using skip connections during the up-sampling stage. We use an anchor box-based detection

architecture to generate predictions using a classification and a regression head. The classification head uses focal loss [66] for to deal with sparse radar point clouds, and the regression head uses Smooth L1 loss.

2.5 Implementation

Image segmentation network

We utilized a pre-trained maskRCNN [47] model from pytorch's model zoo for our image segmentation network due to its accuracy and generalizability. However, depending on the use case, a faster alternative model can also be selected. Our approach remains agnostic to the chosen network type.

Metric

We use BEV average precision (AP) as our main metric in our evaluation with an IoU threshold of 0.5 to determine True Positives.

Perspective view Baseline

We compared our approach against two baselines: CenterFusion[80] and CenterNet[133]. CenterFusion is a state-of-the-art camera-radar fusion approach that creates a feature map of radar point clouds and processes it along with an image-based feature map to perform detections. CenterNet, on the other hand, is a camera-only approach. We fine-tuned the pre-trained CenterFusion network on the Astyx dataset and used the official GitHub implementation of both networks.

Multi-view Baseline (SOTA)

We adopted [57] as our multi-view fusion baseline. The approach utilizes an AVOD architecture from [59]. We trained it on the Astyx dataset using the official implementation. We refer to this approach as AVOD-fusion, which is currently the state-of-the-art approach for camera-radar fusion.

Table 2.1. BEV Average Precision scores for different IoU thresholds in brackets. Scores for the best baseline are underlined. R: Radar; C: Camera

Method	Modality	AP (0.5)		
		Easy	Medium	Hard
CenterNet [133]	C	12.40	13.36	13.11
Center Fusion [80]	R + C	9.78	11.22	10.87
Painted-pointpillars [116]	R + C	–	–	36.00
AVOD-Fusion [57]	R + C	40.38	37.46	36.11
RadSegNet (Ours)	R + C	**48.14**	**46.82**	**45.88**
% increase over underlined		+19.21%	+25.00%	+27.05%

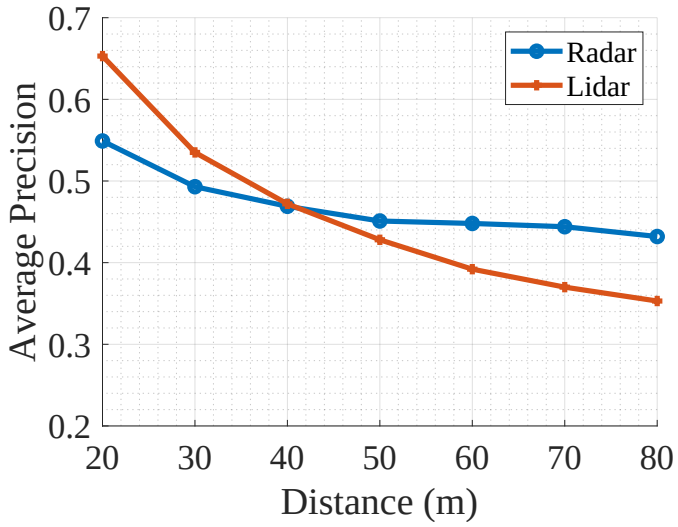


Figure 2.4. Comparison of bounding box detection performance over distance, between lidar and radar input

Testing datasets

We evaluate on two datasets. First, we show results on Astyx Hi-res radar dataset [77] and comprehensively benchmark our approach. In this dataset, we also create augmented weather to evaluate the reliability of different camera-radar fusion approaches in adverse conditions. Next, we evaluate on RADIATE dataset [99], which contains images from real-world bad weather environments.

2.6 Evaluation on Astyx Dataset

2.6.1 Dataset Details

The Astyx Hi-res radar dataset [77] provides high-resolution MIMO radar point cloud data collected on German roads while driving at different speeds. The dataset provides 3D

bounding box labels of vehicles and pedestrians with their occlusion level, generated via human annotation using an onboard lidar point cloud and camera images. We evaluate the dataset for vehicle detection task in 3 categories based on object occlusion level: ”No occlusion (Easy)”, ”Not fully occluded (Medium)” and ”Full Dataset (Hard)”. We split the dataset into 4:1 ratio for training and testing and limit evaluations to 80m, the range at which lidars cannot maintain enough point density causing label certainties to degrade drastically beyond this distance. We also perform k-fold cross validation on the dataset with different train-test splits.

2.6.2 BEV bounding box prediction

In Table 2.1, we present the AP score of our network and compare it with other radar-camera fusion methods. Our approach, RadSegNet, outperforms perspective view-based methods [133, 80] in all three occlusion categories, thanks to its BEV representation. The current state-of-the-art approach, AVOD-fusion [57], also uses BEV representation from radar and serves as the best-performing baseline. However, RadSegNet outperforms AVOD-fusion across all occlusion categories. In the easy category with no occlusion, the effective usage of radar point features in RadSegNet results in a performance improvement. Nonetheless, the percentage increase is higher in medium and hard categories that include occlusion. This demonstrates that our SPG representation in RadSegNet provides a significant advantage over AVOD-fusion’s multi-view fusion, particularly in cases of occlusion where camera features are unreliable, even in clear weather conditions. We also evaluate against painted-pointpillars [116] on camera radar data of Astyx. It is originally a lidar camera fusion approach that uses camera semantic segmentation. RadSegNet still outperforms [116] by a large margin which shows the importance of using radar-specific point features and IIW module based semantic map correction on the overall performance. Please refer to supplementary video for qualitative outputs of our network.

Table 2.2. Effect of different weather conditions. Percentage drop from clear weather performance in parenthesis.

	Clear	Fog	Snow	Rain
AVOD-Fusion [57]	36.11 –	2.38 (-93.41%)	33.56 (-7.06%)	17.41 (-51.79%)
RadSegNet	45.88 –	43.24 (-5.75%)	43.72 (-4.71%)	32.89 (-28.31%)

2.6.3 Performance on Lidar compared to Radar

We conducted an experiment on the Astyx dataset using 16-channel lidar data to analyze the advantages of using radars. We compared the performance of a network trained on lidar and camera data with a network trained on radar and camera data, using our proposed approach (RadSegNet). The results, as shown in Figure 2.4, indicate that lidar outperforms radar at shorter distances but suffers a significant drop in performance as the distance increases due to occlusions and decreased point density. In contrast, radars can reliably operate at longer ranges and perform well despite occlusions. We provide supplementary video with qualitative outputs. Note that even with a lidar with more channels, occlusion issues would still persist.

2.6.4 Performance in adversarial scenarios

In this experiment, we now evaluate the performance of the camera-radar fusion systems when camera images are subjected to adversarial conditions degrading the quality of the images, which is also the main problem we want to address with this work. To compare the performance drop against the normal conditions, we need to augment the camera images in the Astyx dataset with artificial bad weather conditions. Due to the unavailability of dense depth maps and stereo cameras in the Astyx dataset, it is not possible to use physical models of augmentation [45, 92]. However, to showcase the proof of concept we use imgaug [1] library that uses image filters to add bad weather in the images. We also compare results in real adverse weather data in the next section on the RADIATE dataset.

Table 2.2 shows the performance comparison of our work against the AVOD-fusion

[1] https://imgaug.readthedocs.io/en/latest/source/overview/weather.html

Table 2.3. Ablation study experiment. We present the effect of each channel on the overall BEV AP performance of RadSegNet at different IoU thresholds

SPG Channel				
Radar	Position	Num Pts	Camera	AP (0.5)
✓				32.52
✓	✓			33.09
✓	✓	✓		36.09
✓	✓	✓	✓	45.88

baseline. For each augmented weather condition, we also show the performance drop, as a percentage, from the clear weather performance. We use the network trained on clear weather and evaluate the test set with augmented weather augmentations. As shown by the results, AVOD-fusion's performance heavily degrades in cases of fog and rain. This is because AVOD-fusion learns feature representations that are heavily dependent on cameras for each object proposal, which becomes unreliable in adverse conditions. RadSegNet shows that its performance is much less affected in all conditions compared to the AVOD-fusion. These results show the shortcomings of the current radar-camera fusion approaches and the ability of RadSegNet to learn independent features from radar and cameras, that can perform reliably in adverse scenarios. Please refer to supplementary video for qualitative comparisons.

We also compare the IoU drop of semantic segmentation output for different augmentations, treating the segmentation output of the original image as ground truth (qualitative outputs in supplementary video). We get an IoU of 0.61 (Fog), 0.40 (Rain) and 0.57 (Snow). The quality degradation follows the same trend as the AP performance. Past work has shown that the performance of semantic segmentation output can be independently improved by fine tuning on adverse weather data [92, 93], which would further boost the performance of RadSegNet.

2.6.5 Ablation Studies

Table 2.3 shows the results of an ablation study, conducted to evaluate the contribution of each channel of our SPG encoding scheme to the performance of RadSegNet. The incorporation of (x, y, z) position maps and n Num Pts channel resulted in 1.76% and 9.05% improvements,

respectively. The addition of semantic features from the camera produced a significant 27.12% increase in performance, supporting our claim of effective utilisation of the strengths both modalities by RadSegNet. Interestingly, RadSegNet outperforms Center Fusion and performs similarly to AVOD-fusion, even without camera input. This is because RadSegNet's design prioritizes radar data, which experiences less degradation in long-range and occlusion cases and RadSegNet also uses other radar point features alongside BEV grid in SPG, to provide additional information like positional context (x, y, z) and the strength of reflections (n), to aid in object detection and identification.

2.7 Evaluation RADIATE dataset

In this experiment, we evaluate RadSegNet on a large scale radar dataset RADIATE [99]. This dataset uses a mechanical radar which provides dense radar data as output. The dataset also contains scenes from adverse weather conditions such as rain and snow, and bad lighting conditions such as night, making it ideal to test the performance on real world adverse conditions.

2.7.1 Implementation details

RADIATE uses a mechanical Navtech CTS 350-X radar and 2 ZED cameras. The radar data is stored as 2D intensity maps, without any height information. We use the left ZED camera in our evaluation. The ZED camera is only facing forward, so we crop out the intensity maps to only keep the forward direction. The labels are filtered accordingly. RadSegNet uses point clouds as input in order to perform SPG encoding. As the radar input is present in form of intensity maps, we use CFAR [30] filtering technique to convert the intensity maps to 2D point clouds. We use the height of the sensor as the height coordinate for data, in order to get a 3D point cloud. Once we have the camera image and radar point cloud we use to evaluate RadSegNet on the RADIATE dataset. The official training set contains 8890 clear samples and 4151 adverse samples and the test set has 4387 clear and 1222 bad samples.

Table 2.4. Results on the RADIATE dataset. The percentage scores in the bracket show improvement over the clear-only training for respective approaches.

	Trained on	Clear	Bad	Clear + Bad
AVOD-fusion	Clear	43.50	18.03	36.68
AVOD-fusion	Clear + Bad	44.98	22.60 (+25.34%)	38.17
RadSegNet	Clear	58.43	17.12	47.71
RadSegNet	Clear + Bad	**59.63**	**38.70 (+126.05%)**	**53.92**

2.7.2 Performance in Clear/Bad Weather

Table 2.4 shows the performance of object detection on this dataset. We perform two types of experiments 1) training only on clear samples and 2) training on both clear and adverse samples. The same test set is used for both experiments, which contains frames from both clear and adverse data. RadSegNet achieves better performance compared to the AVOD-fusion baseline in both clear and adverse scenarios (41.46% improvement when trained on clear+adverse and tested on clear+adverse). More interestingly, we compare the AP score increase, from first experiment to the second, seen on adverse weather samples in the test set. This increase for RadSegNet is much more significant compared to the same for AVOD-fusion (126% vs 25%). We make two observations: 1) for the dense radar type, there is a slight domain gap between clear and adverse radar data, as a network trained on clear-only does not generalize well over adverse data and 2) RadSegNet's approach of independent information extraction provides much more reliable performance than SOTA when some supervision is provided for adverse data. Overall, this experiment further proves that RadSegNet's fusion provides better performance in good weather conditions and much more reliability in adverse conditions, regardless of the radar type. Refer to supplementary video for qualitative analysis.

2.8 Discussion

RadSegNet performs instance segmentation on camera images and detections on SPG encoded representation. To minimize the overhead of obtaining segmentation in a practical scenario, the two steps can run in parallel by keeping one frame latency between the detection

and the instance segmentation network. Past works have explored the possibility of building such systems [116] and similar techniques can also be employed in our approach.

We demonstrated that independent feature extraction reduces the co-dependency of camera and radar inputs, allowing for more robust performance even in scenarios where camera segmentations are degraded. Future work can involve developing an uncertainty metric to switch off the camera input after reaching a certain level of degradation. It's worth noting that this is only possible due to the independence in camera and radar feature extraction provided by RadSegNet.

Chapter 2, in full, is currently being prepared for submission for publication of the material by Kshitiz Bansal, Keshav Rungta and Dinesh Bharadia. The dissertation author was the primary investigator and author of this paper.

Chapter 3

R-Fiducial

3.1 Introduction

A critical aspect of vehicle autonomy is to be aware of the traffic situation and infrastructure at all times. A significant part of this awareness roots from the proper detection of roadside indicators like stop signs. Apart from the dynamic objects in a scene, a fully autonomous car needs to sense the traffic-related signage, irrespective of the environmental conditions. Conventionally used visual sensors (cameras) entirely rely on color (RGB) signatures to detect things like stop signs. This dependence limits the functionality of cameras only to well-lit environments. Poor lighting situations and glare due to retro-reflective coatings are everyday occurrences that pose difficulties for a camera to perform detection reliably. [34] shows how a small modification could easily fool object detection algorithms, proving that such a fragile technology will not be enough for advanced applications like autonomous driving. Similarly, self-driving vehicles also rely on high-definition maps to obtain information about the static environment. However, due to the inefficiency of self-localization in the map, critical information regarding roadside signs can not be reliably obtained from maps.

Recently, radars have started getting much attention in the autonomous industry, given their robust all-weather operation. Radars can actively sense the scene by illuminating it with EM waves, specifically millimeter waves. With specialized Frequency Modulated Continuous Wave (FMCW) signals, radars can accurately localize objects with high reliability. However, the

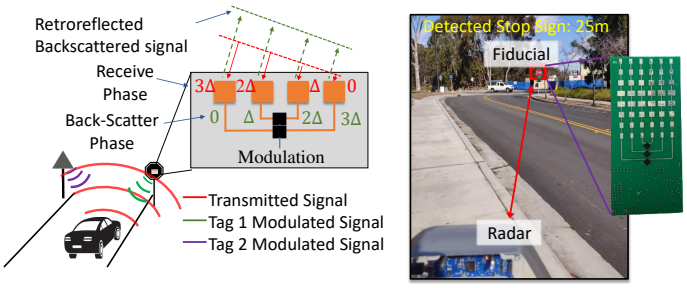


Figure 3.1. a) Retro-reflective tags are attached to traffic signage, and each R-Fiducial uniquely modulates the incident radar signal to distinguish itself from other objects. b) Radar detecting the deployed R-Fiducial on a STOP sign and localizing it w.r.t. to radar's location

Table 3.1. Comparison of R-Fiducial with the past approaches based on different design requirements, BF: Beamforming

	Radar Type	**Range of operation (m)**	**Sensing latency (ms)**	**Doppler performance**
RoS [82]	Custom	<25	Depends on car speed	Precise self-tracking required
Millimetro [105]	Digital BF only	>25	64	Doppler correction required
Omniscatter[15]	Digital BF only	14	≈84	Not Robust to high doppler
R-Fiducial	Analog/Digital BF	**6 - 25+**	**≈0.5 - 38**	**Robust against doppler**

radar reflections lack useful information regarding color and texture, limiting their applicability to certain tasks. The natural question is, if radars are necessary for all-weather perception, how can we enhance their sensing capabilities without relying on texture or color information? In this work, we answer this question by presenting a design of low-cost radar backscatter tags (*R-Fiducial*) that can be ubiquitously deployed to extend the current traffic infrastructure (Fig 3.1) and enable myriad perception tasks like stop sign detection to self-localization just by using radar. Next, we present the design requirements of such a radar fiducial, followed by a summary of the related works.

3.1.1 Fiducial Requirements

Large field of view

A radar fiducial needs to be easily detectable from various angles and distances. When using mmwave radar on a vehicle, the radar signal should be strong enough when reflected back to the same direction. To achieve this, the fiducial should have a wide field of view

Reliable identification

A radar fiducial must also be uniquely identifiable by radar systems. This is important because in a real-world deployment, a radar mounted on a vehicle would receive reflections from many different objects in the scene. All of these signals would act as clutter for the signal reflecting back from the fiducial. Therefore, the fiducial must be designed in such a way that it can be distinguished from other objects in the scene. They also should have an ID to identify them uniquely in the presence of other tags.

Low Latency and Long Range

In addition to accurate detection and localization, radar fiducials must have low latency to be effective in life-critical applications. For instance, if a driver needs to stop a car moving at 30 mph, it requires at least 25 meters of stopping distance after being alerted [1]. Therefore, the detection latency of a stop sign by the radar fiducials needs to be in the order of milliseconds to ensure a safe stopping distance..

Compatibility with existing radars

Radar fiducials must be compatible with existing radar systems to ensure wide applicability. This requires them to be detectable without any hardware changes to current widely-used mmwave radars. Additionally, they should support both digital and analog beamforming modes of operation.

Accurate Localization

Accurately identifying and locating radar fiducials is crucial for detecting traffic signage. The fiducials must have a unique identity, recognizable by radar, and should be detectable using traditional radar processing techniques for localization. This enables the radar to identify and locate the fiducials, which in turn enables accurate detection of traffic signage.

3.1.2 Related Work

Backscatter tags[115, 114, 105] based on van Atta array architecture [110, 98] have been proposed for transportation systems, but existing designs have the following limitations. Vitaz et al. [114, 115] developed a millimeter wave tag for object tracking, but it uses power-hungry PIN diode switches and is read by a pulsed radar, which is not commonly used in automotive applications. Mazaheri et al. [73] and RoS [82], attempt to isolate the tag's scattering by changing the polarization of the backscattered signal, but this is not sufficient to separate the tag's scattering from surrounding reflections because the scattering of a vertically polarized wave from any object generally contains both vertical and horizontal polarizations. Moreover, these are incompatible with automotive radars as automotive radars utilize antennas with a single polarization.

Some other works [105, 64, 15] use a low-frequency modulation signal to isolate the tag's scattering from the surroundings. However, REITS [64] only conducts simulations using millimeter wave tags for transportation systems and lacks real-world testing. Millimetro [105] uses sinc template matching in the Doppler domain to detect tags but becomes unreliable in cases of Doppler shift due to tag or vehicle movement. Similarly, OmniScatter [15] uses a low switching frequency modulation and takes a large point FFT to separate the tags from background objects. Still, this approach is based on the assumption that there is no significant movement between the tag and radar during the measurement duration, which is not always the case. In contrast, our proposed approach uses pseudo-random code sequences to modulate the backscatter signal, which spreads the signal across the entire range-doppler spectrum and provides resilience against high relative movement between radar and the tag. Furthermore, using pseudo-random sequences allows for detection and localization within a single chirp duration, making it suitable for beam-scanning radars. A comparison of the proposed approach with the prior works is shown in Table 3.1.

3.2 Proposed Method

We first briefly explain the functioning of an FMCW radar and how it measures an object's distance to the radar. It transmits a chirp signal of bandwidth B and duration T_c, receives the scattered signals from the objects and dechirps them by mixing with the transmitted chirp. The reflections from an object at a distance d reach the radar after being delayed by time $\tau = \frac{2d}{c}$ where c is the speed of light. The time delay between the received chirp and the transmitted chirp at the radar results in a dechirped signal of frequency $\Delta = \frac{B}{T_c} \times \frac{2d}{c}$, which is a function of radar-to-object distance. The dechirped signal is sampled by an Analog to Digital Converter (ADC) to find its frequency and compute the radar-to-object distance. Next we describe our tag design and how it works as a radar fiducial.

3.2.1 Tag architecture

To ensure reliable detection of objects using radar, it is crucial to receive a sufficient amount of reflected power from the object. This reflective strength is quantified by the Radar Cross Section (RCS), which is the ratio of the reflected power to the incident power. However, RCS is direction-dependent. When an electromagnetic wave is incident on a metal plate, most of the power is reflected away in specular reflection direction, resulting in little reflection back in the same direction. To overcome this, a Van Atta array can be employed, which consists of a linear array of N antennas. The first and Nth antennas, second and $(N-1)^{th}$ antennas, and so on, are interconnected with transmission lines that provide the same phase shift. This interconnection results in the interchanging of phases on the connected antenna pairs and ensures that the reflected wavefront leaves in the same direction as the incident wavefront, as shown in Figure 3.1.

Modulation of Van atta tags: To impart an identity to retro-reflective Van Atta array tags, a modulation signal $m(t)$ is applied to the tag. The modulation is achieved by switching the tag between the ON and OFF states using an RF switch. The antenna pairs are connected in the

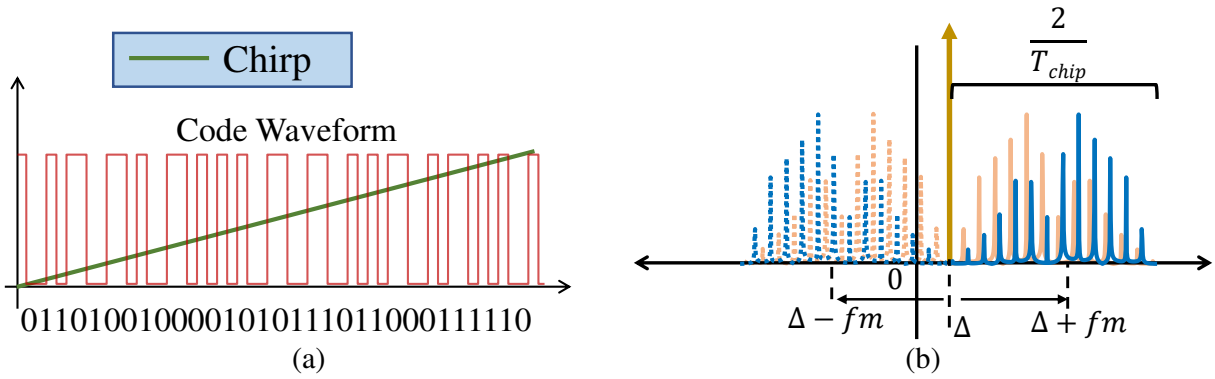


Figure 3.2. (a) Code multiplication with the incident radar chirp at the tag (b) Spectrum of the backscattered signal contains copies of the CDMA code spectrum centered at Δ frequency corresponding to the tag to radar separation.

ON state, resulting in a high Radar Cross Section (RCS). In the OFF state, the retro-reflectivity is reduced, providing lower RCS. AN On-OFF modulating signal modulates the RCS of the tag and provides a unique signature is provided to each tag.

3.2.2 R-Fiducial's novel Spreading Spectrum Modulation

We propose a tag modulation method to achieve unique identifiability of radar fiducials, allowing for the simultaneous operation of multiple tags while isolating their scattering from surrounding objects. Inspired by Code Division Multiple Access (CDMA) communication, our method employs orthogonal spreading codes associated with each tag, enabling multi-tag operation in the presence of background objects. This approach supports the detection of multiple tags in a single chirp period.

Each spreading code is an N-bit sequence that takes values in $\{+1,0\}$, with good auto-correlation and very low cross-correlation with other spreading codes. These codes continuously repeat on the tag and modulate the backscattered signal. Each tag has a unique spreading code, which serves as its characteristic property and helps isolate its backscatter signal from other tags and objects in the surroundings. Suppose two tags use spreading codes $C_1[n]$ and $C_2[n]$. Reflected chirps from the tags contain code1 $C_1[n]$ and code2 $C_2[n]$ periodically repeating on them, giving rise to $[C_1C_1....C_1] + [C_2C_2....C_2]$ at the receiver. Upon cross-correlating, the received signal with $C_1[n]$, the resulting correlation is dominated by code1 because $C_2[n]$ is very

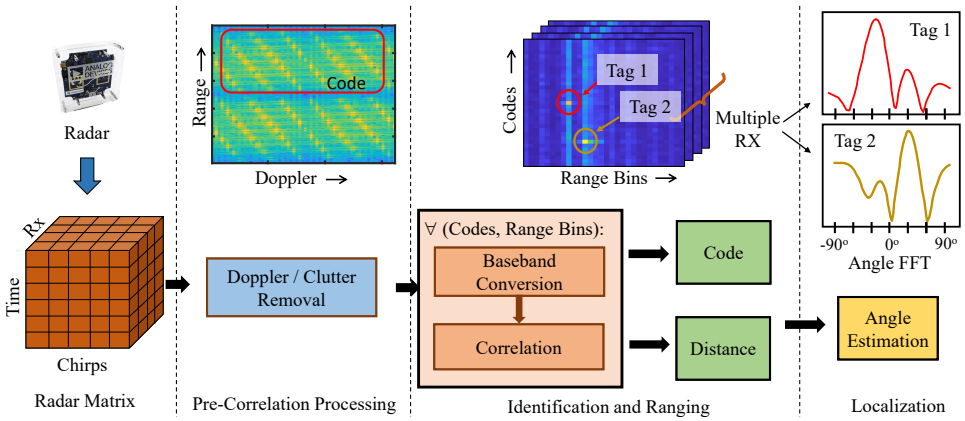


Figure 3.3. Radar Processing Overview: The received signal from the radar is preprocessed to remove doppler and static clutter. The next step jointly solves for identifying the correct code and distance from the radar. The correlation values for the identified tags across multiple receivers are used to estimate the angle of each tag, thus providing the locations.

weakly correlated with $C_1[n]$. The cross-correlation of the received signal with C_1 looks like a sequence of k impulses where k is the number of code repetitions in a chirp duration. Similar observations can be made when correlating with the code C_2. The periodicity of these peaks in the cross-correlation helps us identify the presence of a particular code, and the codes help us to distinguish between the multiple tags.

$$\begin{aligned} R[n] &= C_1[n] * [C_1C_1....C_1] + C_1[n] * [C_2C_2....C_2] \\ &\approx \sum_{l=0}^{l=k} \delta[n - lN] + 0 \end{aligned} \tag{3.1}$$

To apply a spreading code to the tag's backscattered signal, we utilize the N-bit spreading code to generate the modulation signal $m(t)$. Here, a binary value of 1 corresponds to an ON to OFF transition with a period of $\frac{1}{f_m}$, while a binary value of 0 corresponds to an OFF to ON transition with the same period. Denote $c(t)$ as the continuous time version of the code sequence $C[n]$. Now we concatenate the transitions for all N-bits to obtain the waveform $m(t)$, as illustrated in Figure 3.2a. This process is equivalent to multiplying $c(t)$ with a square wave of frequency f_m. Therefore, each bit has a duration of $T_{bit} = \frac{1}{f_m}$, and $c(t)$ occupies a bandwidth of $\frac{2}{T_{bit}} = 2f_m$. By considering only the first-order harmonics in the Fourier series decomposition of a square wave, we can express $m(t)$ in terms of $c(t)$ and f_m, as shown in equation 3.2.

$$m(t) \approx c(t)\frac{2}{\pi}cos(2\pi f_m t) \approx \frac{c(t)}{\pi}[e^{j2\pi f_m t} + e^{-j2\pi f_m t}] \tag{3.2}$$

Choice of Modulation frequency: Here, we explain how to choose the modulation frequency f_m in relation to the ADC's sampling frequency. From equation 3.2, it can be seen that the modulation signal $m(t)$ has 2 copies of $c(t)$ centered at $f_m, -f_m$. Each copy has a bandwidth of $2f_m$, and $m(t)$ occupies a total bandwidth of $4f_m$. The modulation constraints arise because of the following conditions: (a) Each backscattered chirp contains k repetitions of the code, and (b) the radar's Analog to Digital converter (ADC) has a finite sampling rate of f_s. To meet the first condition, the bit duration in each code has to be adjusted to ensure k code repetitions in a chirp duration T_c. Recall that a code contains N-bits each of duration $\frac{1}{f_m}$, so each code occupies a length of N/f_m. Hence the chirp duration T_c must be greater than k code durations leading to the timing constraint given by the inequality 3.3.

$$T_c \geq k\frac{N}{f_m} \implies f_m \geq k\frac{N}{T_c} \tag{3.3}$$

Also, for the radar to capture the modulated signal, we have to ensure the spectrum of the modulating signal lies within the bandwidth of the ADC. Since the modulating signal $m(t)$ occupies $4f_m$ bandwidth, it leads to bandwidth constraint as given by inequality 3.4.

$$4f_m \leq f_s \implies f_m \leq \frac{f_s}{4} \tag{3.4}$$

So, for a given chirp duration and the radar's sampling frequency, the modulation frequency f_m has to be chosen to satisfy both the timing and bandwidth constraints.

3.2.3 Detection and Localization of R-Fiducial

In this section, we describe our algorithm (overview in Figure 3.3) for identifying one or more tags and then localizing them accurately, i.e., finding the distance and angle of the tag w.r.t. the radar.

Identifying and ranging the tag using cross-correlation: The chirps reflected from the tag contain the modulating signal $m(t)$ in the radar's received signal. Upon dechirping the received signal at the radar, the dechirped signal contains the tag reflection $D_{mod}(t) = m(t)cos[2\pi\Delta t]$ where Δ is the frequency corresponding to the tag's distance from the radar. We decompose the tag reflection as shown in equation 3.5.

$$\begin{aligned} D_{mod}(t) &= c(t)\left(\frac{2}{\pi}cos[2\pi f_m t]cos[2\pi\Delta t]\right) \\ &= \frac{c(t)}{2\pi}(e^{j2\pi(f_m+\Delta)t} + e^{-j2\pi(f_m+\Delta)t} \\ &+ e^{j2\pi(f_m-\Delta)t} + e^{-j2\pi(f_m-\Delta)t}) \end{aligned} \tag{3.5}$$

In equation 3.5, it can be seen that the reflected signal of a tag at a radar contains multiple copies of a code $c(t)$ centered at different frequencies $f_m+\Delta, f_m-\Delta, -f_m+\Delta$, and $-f_m-\Delta$ and the tag's backscattered signal spectrum is illustrated in Figure 3.2b. Given the received samples at the radar $D_{mod}(t)$, there are two unknowns for each tag: the distance of the tag, which is captured in the term Δ, and the identity of the tag, which is captured in the code $c(t)$. Moreover, the received signal also consists of multiple environmental reflections.

A joint solution is employed to solve for the tag-radar separation and the tag's identity. The codes modulated by the tags are orthogonal, as described in section 3.2.2, and therefore, cross-correlating the received samples with the correct code reveals whether the code exists or not. Consequently, we should observe high cross-correlation with periodically repeating peaks when computing the cross-correlations of the received samples with the correct code. However, a direct cross-correlation of $D_{mod}(t)$ with the correct code $c(t)$ does not result in a good correlation

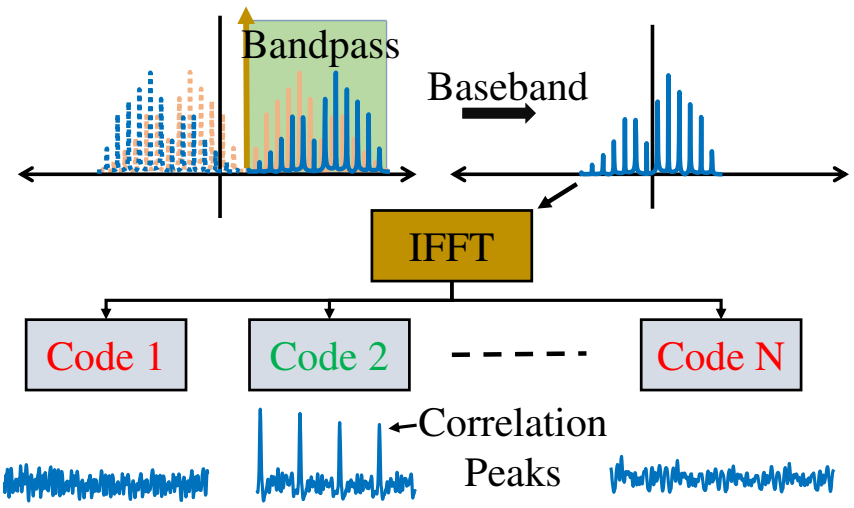


Figure 3.4. Correlation: The received backscattered signal from the tag contains 4 copies of the code in the frequency domain due to aliasing. The component at $\Delta + f_m$ is retrieved using bandpass filtering and then converted to baseband. The time-domain version of the signal is then correlated with all the codes in the codebook for identifying the tag.

due to the presence of four different copies of the code in $D_{mod}(t)$, each with a frequency offset term that corrupts the correlation. It is important to note that frequency offsets are unknown since we do not know the tag-radar distance captured in the term Δ.

To eliminate the unknown frequency offsets, the joint estimation of Δ and the code sequence in the backscatter signal is performed by iterating over the possible range of values that Δ can take and all the code sequences that can be present. Prior to cross-correlation, filtering of the samples of $D_{mod}(t)$ is applied to remove undesirable terms from Equation 3.5. Figure 3.4 illustrates that the code copies centered at $f_m + \Delta$ and $f_m - \Delta$ have distinct frequency ranges. A bandpass filter with a bandwidth of $2f_m$ is applied to the samples to preserve the copies at $f_m + \Delta$ and $f_m - \Delta$ and eliminate the copies at $-f_m + \Delta$ and $-f_m - \Delta$. This filtering process is shown in Equation 3.6.

$$\text{Filtered Signal} = c(t)[e^{j2\pi(f_m+\Delta)t} + e^{j2\pi(f_m-\Delta)t}] \tag{3.6}$$

In the next step, we remove the frequency offset factors in the filtered signal by shifting the code

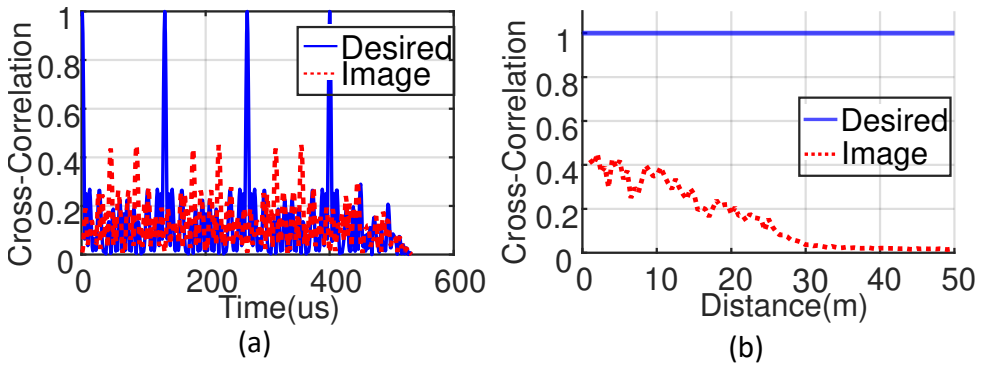


Figure 3.5. (a) Normalized cross-correlation for 2m tag to radar separation. (b) Peak value of cross-correlations as a function of tag-to-radar distance. Image signal's contribution to the cross-correlation reduces gradually as tag-to-radar separation increases

copy centered at $f_m + \Delta$ to zero frequency. This is achieved by multiplying the filtered signal with $e^{-j2\pi(f_m+\Delta)t}$, leading to $x(t)$ in equation 3.7.

$$x(t) = c(t) + c(t)e^{j2\pi(-2\Delta)t}. \tag{3.7}$$

After filtering out undesired frequency components, the resulting signal $x(t)$ is correlated with all the code sequences. The desired signal $c(t)$ in $x(t)$ generates a periodic impulse train in cross-correlation with correct code, while an additional image signal $c(t)e^{j2\pi(-2\Delta)t}$ with a frequency offset of 2Δ generates a smaller cross-correlation peak that decreases with increasing tag-to-radar separation Δ. In Figure 3.5, we empirically show that the cross-correlation contribution from the image signal becomes smaller as Δ increases. We perform the cross-correlations of $x(t)$ with all possible codes over all the range bins and store the correlation peaks in a 2D matrix. Then we identify the highest value in the 2D matrix to determine the code in the backscatter signal and its associated tag-to-radar separation, enabling joint determination of the tag-to-radar separation and code.

Performance in doppler: We analyze the impact of the doppler effect caused by the moving vehicle on the joint decoding. The fast movement of a vehicle during a chirp duration causes the received chirp to be slightly shifted in frequency proportional to the carrier frequency and the vehicle's velocity (doppler effect). This shift in frequency manifests as a shift in the code

spectrum by a few bins in the range FFT. Since our joint decoding algorithm iterates over all possible frequency shifts, we can still identify the tag. Still, the shift in the spectrum introduces the error in the estimated tag to radar separation. However, when we simulate the doppler scenario for up to 150kmph speeds (see evaluation Figure 3.9) we find that the doppler results only in a maximum error of 2 meters.

Combining across multiple chirps: To improve the accuracy and range of tag detection using radar, it is necessary to enhance the signal-to-noise ratio (SNR) of the received signal, as the strength of the tag's backscatter signal varies inversely with the fourth power of the tag-to-radar separation ($\propto \frac{1}{d^4}$). This can make it difficult to detect periodic peaks in the cross-correlation, as the reflected signal power can quickly drop to the receiver noise floor. To address this issue, we combine radar samples from multiple received chirps, allowing the cross-correlation periodic peaks to stand out from the noise. Typically, there is a gap time between consecutive chirps, which we fill with zeros and append to the chirp samples. Cross-correlation is then performed with the code sequence to detect the peaks and identify the code present. Cross-correlation on the appended samples is equivalent to averaging over the noise samples, which reduces the noise power and results in a processing gain. Combining L chirps reduces the noise power by a factor of L and results in a processing gain of $10\log_{10}(L)$ dB. By setting a threshold based on the peak-to-noise ratio, we can detect the presence of a particular code. Any correlation value greater than the threshold is marked as a positive detection.

3.3 Implementation

In our study, we implemented and evaluated an end-to-end detection system using the 24 GHz DEMORAD radar platform from Analog Devices. The platform is a MIMO radar with 2 transmit and 4 receiver antennas, and we used one transmit and 4 receive chains to collect data and estimate the angle of arrival for the tag's angular location. To ensure good autocorrelation and cross-correlation properties between codes, we used 31-bit Gold code sequences [41, 42], which

are widely used in GPS [106] and CDMA [31] standards. The code sequence was modulated at 250kHz modulation frequency to meet the bandwidth constraint. We configured the radar with an upchirp time of 496 μs and a gap time of 104 μs between two chirps, but our implementation of R-Fiducial is independent of these timing parameters and could work with other choices. Longer bit Gold sequences can also be used to support more tags.

Tag's RF circuit: The tag's RF circuitry is designed on a printed circuit board (PCB) that includes Van Atta Array antennas connected by transmission lines and RF switches for modulating backscattered signals. The directional patch antennas on the tag have a 50-ohm impedance over a frequency range of 24GHz to 24.3GHz, offering a 14 dBi peak gain and 100-degree horizontal field of view for long-range and wide-field coverage. To minimize RF trace losses, we designed the PCB on a 20 mil Rogers 4003C dielectric substrate. We chose the MASW-011105 RF switch from MACOM for its low insertion loss of 1.6 dB at 24GHz and low DC power consumption of 5 uW [107].

Tag's Control circuit: An onboard PSoC 6 MCU is used to control the switch on the tag. We generate 31-bit Gold code sequences using two Linear feedback shift registers (LFSR) on the MCU, which are outputted from a GPIO pin as the control signal. The LFSRs are configurable to use different length codes, and in our implementation, we use 5-bit LFSRs to generate 31-bit codes.

3.4 Evaluations

3.4.1 R-Fiducial's performance for a single tag

Detection rate and Range of operation: R-Fiducial is designed to meet the low-latency requirements of the application, as it can perform all its operations using a single chirp's data. However, it also allows for coherently processing multiple chirps during runtime to extend the operating range and improve reliability. In Figure 3.6a, we demonstrate the detection performance of R-Fiducial over a range of more than 30 m for different observation times, with 100 frame

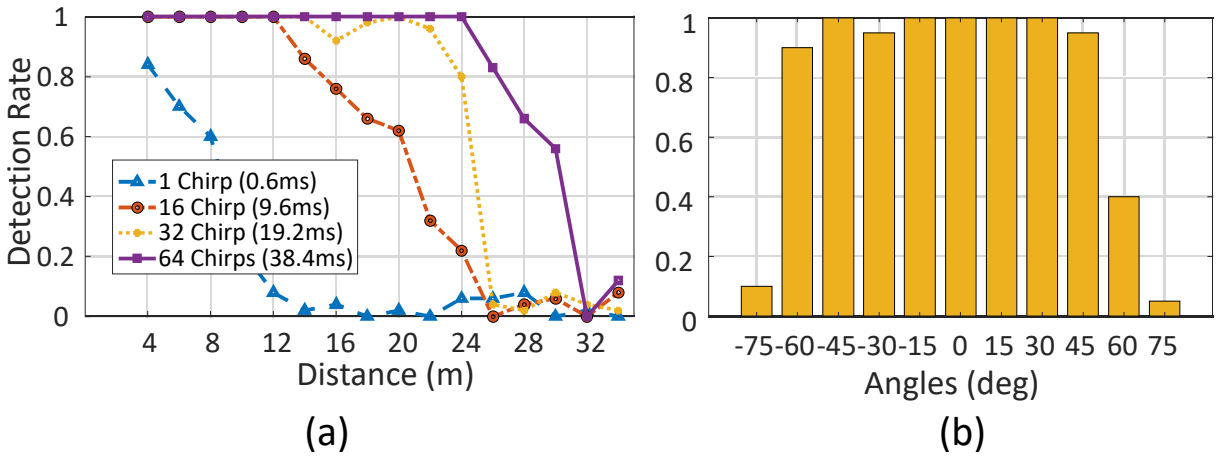
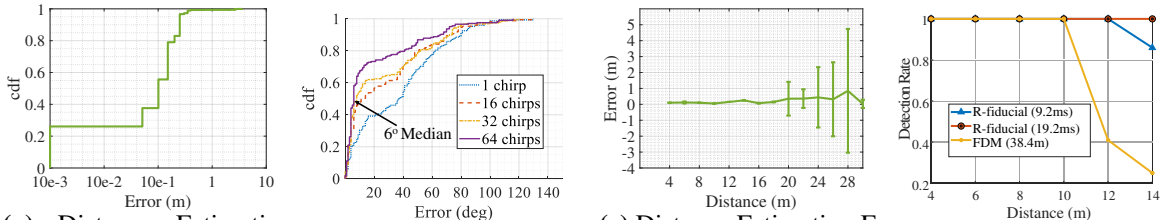


Figure 3.6. (a) Detection Rate for different observation times. (b) Detection rate for different angles of incidence showing a wide field of view of our tags.



(a) Distance Estimation CDF

(b) Angle Error CDF

(c) Distance Estimation Error

(d) Millimetro vs Ours

Figure 3.7. (a-c) Localization performance of R-Fiducial (d) Detection rate trend with the distance (the observation time is given in brackets) when comparing R-Fiducial with the FM scheme. R-Fiducial provides more reliable results in lesser latency.

measurements taken at each distance. Note that the maximum range of detection heavily depends on the antenna gain of the tag [105] and the radar hardware. For instance, the maximum EIRP allowed at the 24GHz band is 20dBm, at which the detection range would extend to around 45(130) m for 1(64) chirp combining.

Field of View: The wide Field of view is a crucial feature of R-Fiducial, and its detection rate for different angles of incidence is evaluated to understand its performance. At a distance of 5m from the tag, 100 frame measurements are taken for each angle, and the detection rate is measured. Figure 3.6b illustrates that R-Fiducial can maintain its reliable detection performance for more than 60 degrees of incidence on each side, making it suitable for challenging scenarios such as reading a stop sign on a curved road or an exit sign from a long distance in a corridor.

Localization: We evaluate R-Fiducial's performance in estimating the range and angle of a tag in an outdoor environment. The results, shown in Figure 7a-b, indicate that the median error

Table 3.2. Area under curve for multi-tag experiment

	AUC		
Latency →	**2.4 ms**	**9.6 ms**	**38.4 ms**
Config 1	0.985	0.998	0.999
Config 2	0.887	0.986	0.999
Config 3	0.627	0.937	0.999

in distance and angle estimation is 0.1m and 6 degrees, respectively, based on 100 measurements taken at various distances and angles. The error in distance estimation increases with distance, and its standard deviation increases as the signal-to-noise ratio (SNR) decreases.

3.4.2 R-Fiducial's scalability to multiple tags

The scalability of R-Fiducial is crucial, and we evaluate its ability to operate with multiple tags simultaneously. The correlation-based detector correlates all codes in the codebook with the received signal. To detect the tags, a detection threshold must be set. This threshold determines the number of detections made and the potential for false positives. We use the Area Under the Curve (AUC) metric to evaluate the detector's performance with multiple tags. The AUC measures the classifier's ability to distinguish between classes by plotting the true positive rate against the false-positive rate for different threshold values.

Detection performance and latency: In this experiment, we evaluate the performance of R-Fiducial in detecting and identifying multiple tags simultaneously. We place three tags at different locations within a range of 10m from the radar, covering different configurations of wide-angle separation (config 1) to closely spaced tags (config 3). We capture 100 frames for each configuration and plot the true positive rate against the false-positive rate to assess the performance. The AUC values are calculated for different chirp combinations and configurations, and the results are presented in Figure 3.8 and Table 3.2. Combining many chirps provides almost perfect AUC values, while a small number of chirps can yield good AUC values for smaller distances. The latency of R-Fiducial is independent of the number of tags, as all codes in all distance bins are checked for presence regardless of the number of tags. This experiment demonstrates that R-Fiducial can detect and identify multiple tags simultaneously

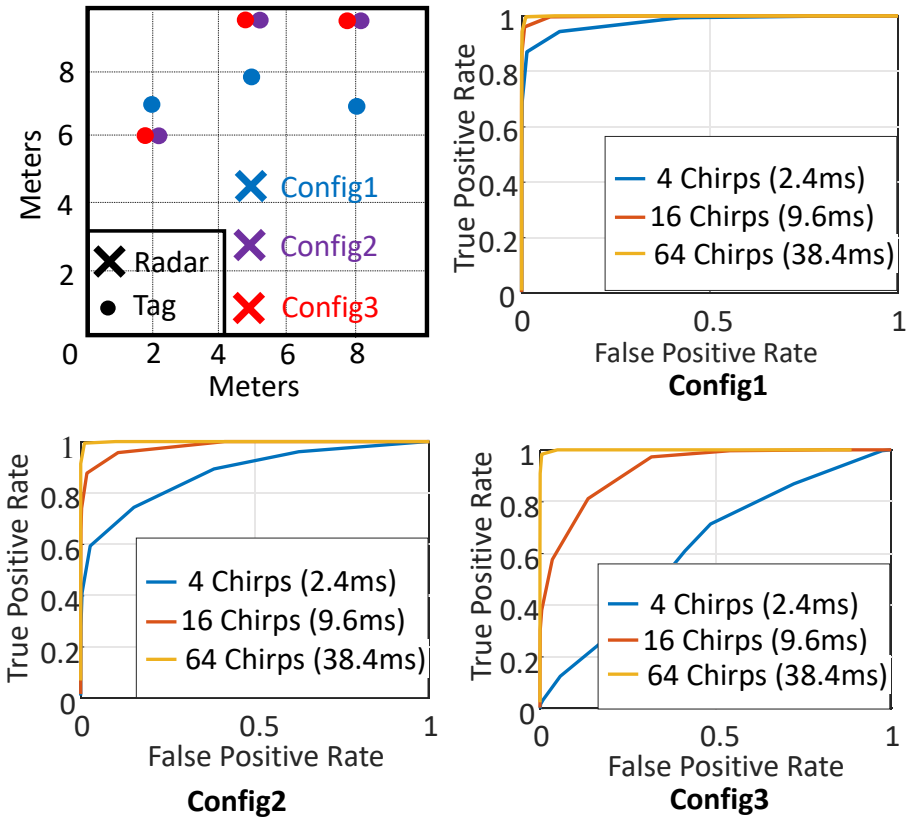


Figure 3.8. ROC curves for multi-tag experiments for different chirp combinations. More the curve towards the left, the better

while maintaining its high reliability.

3.4.3 R-Fiducial in the field

We conduct experiments to demonstrate R-Fiducial's ability to detect traffic infrastructure in low visibility conditions such as fog, where other systems like cameras and LiDAR may fail. Results (Figure 3.9a) show that R-Fiducial could reliably detect a stop sign at a distance of 10m, demonstrating its high reliability in adverse weather conditions, which is essential for both indoor and outdoor applications.

Mobility Experiments: To show the performance of R-Fiducial in more realistic settings, we perform the experiments by mounting the radar on a car moving at varying speeds to evaluate the detection performance of R-Fiducial. Figure 3.9b shows the deployment of radar on the car for mobility experiments. Figure 3.9c shows that R-Fiducial maintains its high detection rate

even under mobility scenarios. To further test our system at higher dopplers, we simulate speeds up to 150 kmph and plot the distance estimation error in Figure 3.9d to show that R-Fiducial can detect the tag even in case of high doppler and provide accurate distance estimation. R-Fiducial's spread spectrum codes provide resiliency to doppler by design. (section 3.2.3).

3.4.4 Microbenchmark: Reliability of Code-based modulation

In this study, we compare the performance of our code-based modulation (CDM) scheme with a frequency-based modulation (FM) baseline called Millimetro [105] in terms of detection rate. We use the same hardware for both schemes, and the FM scheme uses five modulation frequencies between 200-600 Hz. The experiment involves moving the tag from 2m to 14m and measuring the time required for detection. Our results (Figure 3.7d) show that the detection rate of the FM scheme starts to decrease after 10m, while our CDM scheme maintains reliable detection with low latency. We also note that the Millimetro [105] uses high-gain narrow beamwidth antennas, providing longer range but narrow azimuthal beamwidth, which limit its practicality for many radar fiducial applications. In contrast, our CDM scheme provides higher reliability, maintains low detection latency, and allows longer observation times for increased processing gain.

Chapter 3, in full, is a reprint of the material as it appears in "R-fiducial: Millimeter Wave Radar Fiducials for Sensing Traffic Infrastructure." by Kshitiz Bansal, Manideep Dunna, Sanjeev Anthia Ganesh, Eamon Patamasing and Dinesh Bharadia, which appears in 2023 IEEE 97th Vehicular Technology Conference (VTC2023-Spring). IEEE, 2023. The dissertation author was the primary investigator and author of this paper.

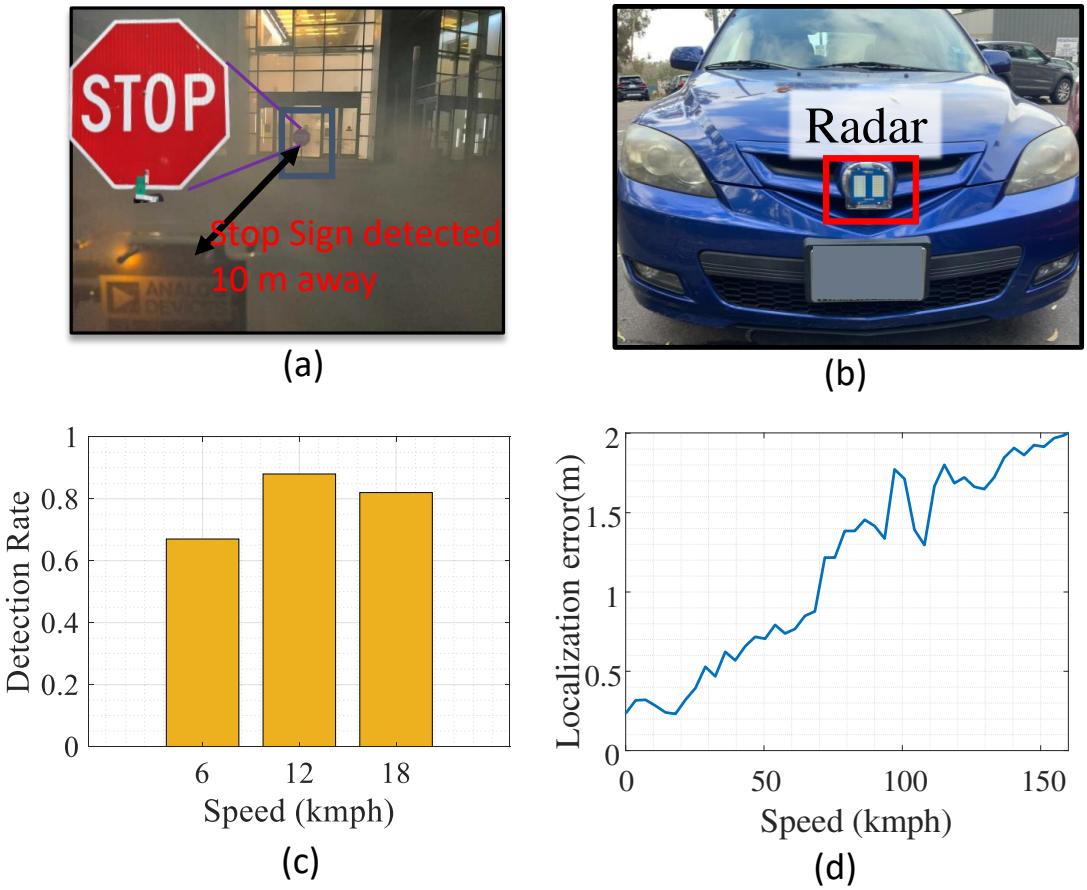


Figure 3.9. (a) R-Fiducial in bad-weather (b) Experimental setup with radar deployed on the car. (c) The detection rate of R-Fiducial in real-world experiments when the radar is mounted on a car that runs at different speeds. (d) Simulation results of distance estimation error in case of high doppler.

Chapter 4

SHENRON

4.1 Introduction

The adoption of high-resolution millimeter-wave (mmWave) radars in autonomous systems has increased due to their reliability in adverse conditions such as fog or smoke. However, radar algorithm development and widespread deployment for various applications have not kept pace with lidars and cameras, primarily due to the scarcity of available high-resolution radar data, which hinders rapid algorithm development and testing. Additionally, the performance of radar algorithms depends on the configuration used during radar data collection, and each application requires a unique set of configurations. For instance, detecting high-speed vehicles on the road over long distances requires a high maximum range and speed configuration, while tracking and identifying humans in a warehouse may require a lesser maximum range or speed but better resolution in range and speed. These different configurations in radar are obtained by choosing different values for parameters, such as the bandwidth of signal or the time between frames. Finding a good configuration for a particular application requires iterations through the vast space of radar parameters which in turn requires extensive real-world data collection. Still, these parameters are only good for the application where they have been tested. Due to this, deploying radars for new applications quite often suffer from what is known as a ”cold-start problem”, i.e. a significant real-world data collection effort is required to even find a good set of parameters before algorithm development can start [113].

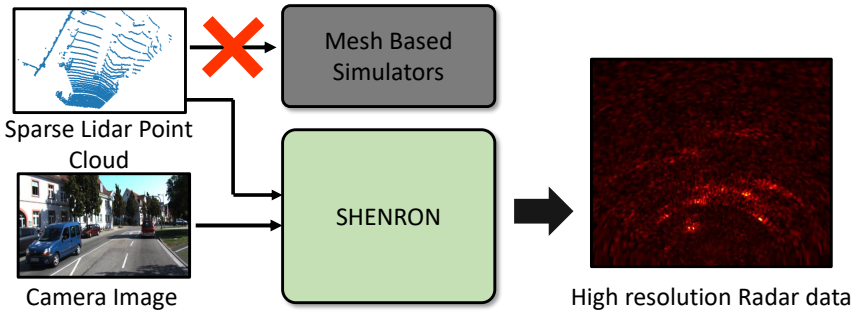


Figure 4.1. Current radar simulators need high-quality mesh input to simulate radars. We present SHENRON that uses only a sparse lidar and camera input to generate high resolution radar data. Above is an example of automotive radar data generation using a lidar and camera from Kitti dataset.

To tackle the problem of data scarcity, simulations provide a unique opportunity. For radars, having access to simulations can provide large-scale data for algorithm development and testing, and also allow quick iterations through the radar's vast parameter space to find a small set of good parameters for a particular application. Such a simulator should fulfill four crucial requirements; **R1** - *time and compute efficiency* for rapid iteration through vast parameter space, **R2** - *ease of defining the input* for wider applicability, **R3** - *scalability* for large-scale testing in diverse conditions and **R4** - *configurability* to test different kinds of radars and parameters.

The existing commercial radar simulation tools [23, 103] are not openly accessible and require users to input high-quality mesh and material definitions for defining a scene and its objects. This is simply not possible in many cases, as generating a detailed mesh requires expert knowledge and equipment, and a poorly designed mesh could significantly hurt the performance of simulations, violating requirement **R2**. Even if one generates a detailed mesh, scaling to new scenes again requires significant effort. This makes these tools difficult to use and hard to scale to new conditions, violating requirement **R3**. Further, for scenes with complex meshes, simulation time can reach an order of hours as it takes a long time to simulate the exact physics of electromagnetic interactions, violating **R1** and making it infeasible to scale up the simulations, violating **R3**. Moreover, most of these tools have only support limited radar configuration such as single transmitter and receiver, violating **R4**.

In this paper, we present SHENRON, a highly efficient, easy-to-use and scalable method

of simulating high-resolution radar data. Rather than requiring a detailed mapping of the environment generated by an expert, we design a simulator that only needs a sparse lidar point cloud and a camera image to generate high-fidelity radar data (Figure 4.1) (**R2**). This means that SHENRON can immediately convert all the existing large-scale labeled lidar datasets into radar datasets with user-defined parameters, allowing for rapid algorithm development and testing data large scale (**R3**). SHENRON's design is flexible and can be used to create simulation for any radar waveform or antenna configuration (**R4**). Most importantly, a user can simply scan any space with a sparse lidar and camera and use SHENRON to rapidly evaluate multiple radar configurations for their choice of application. Hence, SHENRON, for the first time, makes the rapid design, testing and parameter search for radar perception possible, without needing expensive data collection effort or compute heavy mesh-based sensor simulations.

A major challenge in using raw lidar data as the basis for radar simulations is that it lacks the crucial material information required to model RF reflections. In SHENRON, we solve this problem by using the semantic information present in the camera to identify objects present in the scene. We map this semantic information to accurate RF reflection profiles of different objects using real-world measurements and physics-based modelling. With these derived reflection profiles, we are able to significantly improve the runtime of the simulator compared to commercial software (**R1**), while keeping the simulations high-fidelity.

The main objective of any simulation is to accurately predict the performance an algorithm would achieve when deployed in real world [56]. In our evaluations, we show that SHENRON's power profiles achieves a very high correlation value of **0.8433** when compared to real-world data. Further, we choose the commonly used indoor and outdoor algorithms of mapping and object detection and show how SHENRON can accurately predict the performance obtained by different algorithms in real world, only by using simulated data. We demonstrate the generalization capability of our system by evaluating it using two different radars operating at 77 GHz and 24 GHz frequencies. SHENRON is open-source and available at https://wcsng.ucsd.edu/shenron/.

In summary, we make the following contributions:

- We provide an open-source radar simulation framework, that eliminates the need for high-quality meshes and provides large-scale radar data only using sparse lidar point clouds.
- We show how a camera image can be used to obtain the missing material information in lidar point clouds.
- We derive highly accurate RF reflection profiles using real-world measurements that make the simulations fast and high-fidelity.
- We perform indoor and outdoor case studies showing how SHENRON helps in algorithm development, training and testing models and parameter tuning.

4.2 Related Work

Wireless Channel Solvers: Commercial wireless solvers use ray-tracing and EM wave models to provide the wireless channel for the scenes. HFSS SBR+ (shooting and bouncing rays) [23] uses geometric optics (GO), uniform theory of diffraction (UTD) and creeping wave physics to simulate radar data and Sligar et al [104] uses HFSS SBR+ solver and generate range-doppler data. However, they do not account for the diffused scattering effects of radar signals. Remcom's Wavefarer [103, 27] further extend the framework to include diffused scattering using the models provided in [29]. The major shortcoming of these tools is the heavy dependence on the high-quality meshes that require expert equipment, are difficult to scale and have large time and memory requirements (requirements R1. R2 and R3). Furthermore, they do not provide the ability to simulate multi-antenna radar data, limiting their application for variety of use-cases (requirement R4). SHENRON presents a technique to accurately simulate high-resolution MIMO (multi-input multi-output) radar data without needing high-quality meshes, thereby reducing time and memory, as well as making it much more scalable.

Single channel radar: There have been efforts in the past to simulate radar either as a sensor to detect objects as point targets or simulate range-doppler data for single channel between

transmit and receive [117, 49, 108]. Li et al. [63] uses a visibility based model of targets and radar range equation to simulate received power. However, they do not consider the effect of extended targets and the scattering phenomenon of radar signals. Machida et al. [69] propose a new method to perform SBR in an efficient manner. However, they use Torrance-sparrow model of reflection which is only an extension of specular reflections to rough surfaces. They also do not model scattering of radar signals and do not provide validation against the real data.

MIMO radar: To the best of our knowledge, the only past work that performs simulation for MIMO radar is from Schussler et al. [97]. However, as they also use meshes as input, it requires expert modelling and can not be used to convert existing lidar based datasets into radar (requirement R2). Moreover, they do not provide any details about how to obtain materials for large scale simulations, limiting their applicability to define large scale scenes (requirement R3).

Lidar to Radar translation: Deep learning based approaches L2R-GAN [119] and There and Back again [122] use a neural network to learn the sensor model of radar. However, using deep learning based simulations limits the flexibility of simulations. These approaches have only been used to simulate mechanical radar and are not easily scalable to other types of radars such as MIMO radar or radars with different antenna geometries (requirement R4). This severely constraints the applicability of these simulation frameworks. SHENRON uses physics based models to achieve high-fidelity radar simulations, hence making possible the simulation of any kind of radar signal and parameters.

4.3 Background and Motivation

An automotive MIMO radar uses antenna arrays as transmitter and receivers. A radar antenna radiates energy in the space defined by its field of view in form of millimeter waves [16]. These waves interact with all the objects in the scene based on their visibility from radar. The interaction of the waves with objects is in the form of specular reflections and diffused scattering [28]. The amount of energy making its way back to the radar from that object surface

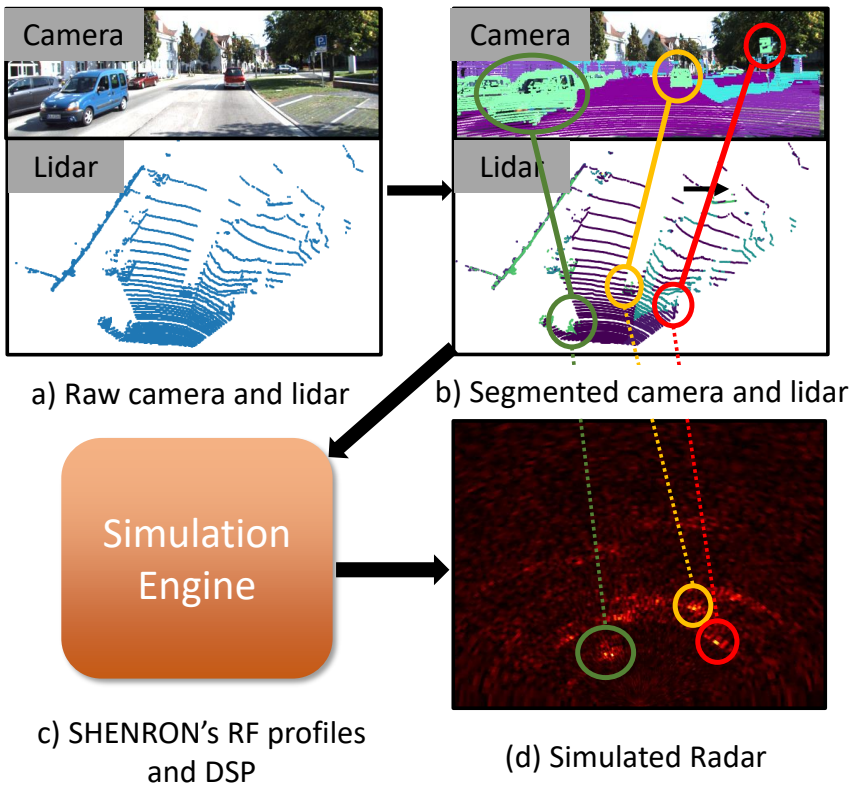


Figure 4.2. Overview of SHENRON. We first obtain the scene from the lidar point clouds and infer material properties using camera images. This information along with derived RF reflection profiles is used to model accurate reflection power and is finally converted to radar data using the DSP module. Green, Yellow: Cars ; Red: Traffic Sign

is determined based on the orientation of the surface and its material properties. During the reception in radar, all the reflections add up at the receiver and sampled by an ADC (analog to digital converter). Radars either directly provide raw ADC samples or filtered point clouds. The latter are obtained by using a detection technique like CFAR (Constant False Alarm Rate) filtering[90], which detects the energy corresponding to any object's reflection, by estimating the local noise power. For either data format to be accurate, it is extremely important to correctly model the power return from the surface of objects. Failure to do so may cause the object to be indistinguishable from noise.

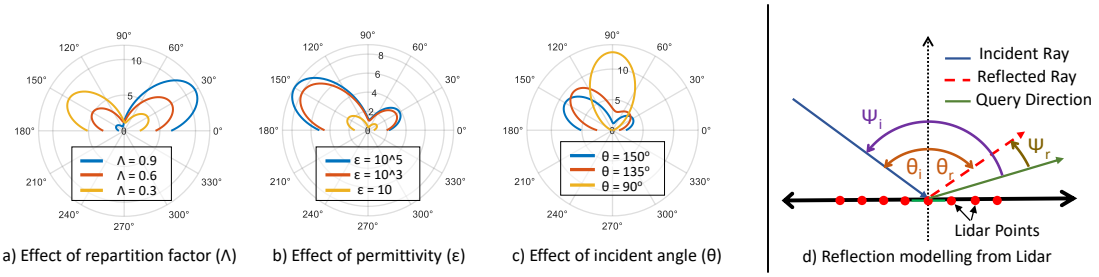
The objective of SHENRON is to accurately simulate the radar data for a given scene and simple object representations. In the next section, we first dive into the idea of how these object representations can be generated in a scalable way by piggybacking on top of how a

lidar perceives these objects, instead of requiring complicated meshes/environment geometries. Moving ahead, we would discuss how additional properties, like material types can be obtained to enhance the point clouds, as well as help model the reflections accurately for the radar signal. Figure 4.2 provides an overview of SHENRON.

4.4 SHENRON

4.4.1 Lidar point cloud as an input

Typically in a simulation environment, when a mesh defines a surface, a ray-tracer returns all the points of interaction of the signal with the surfaces of objects present in the scene. This provides an impulse response of the scene in the form of interaction points and their power of reflection. This impulse response can be convolved with a radar waveform to generate the radar data. If one can obtain these interaction points in an efficient way and model the power reflected from each point accurately, it would obviate the need of using a mesh input. We make a key observation that lidar sensors scan the real environment by shooting rays. These rays travel from the lidar sensor, and reflect back towards the lidar after interacting with any surface. Due to this nature of operation, lidar directly captures the impulse response of a real world scene in the form of a point cloud, where each point denotes a point of interaction. Modern lidars have high angular resolution, and they generate point clouds with uniform polar density allowing them to cover the scene in great detail. Hence, SHENRON directly uses these lidar point clouds as input to represent the scene geometry in a simple and efficient form. This approach allows any user to directly collect real-world data of the domain of interest along with its finer details and efficiently simulate radar data. Moreover, there are abundant open-source lidar point cloud datasets that have data from real-life traffic scenarios. Using these datasets in SHENRON, one can generate large amounts of radar data, which use real-life scenes as input geometry.



a) Effect of repartition factor (Λ) b) Effect of permittivity (ε) c) Effect of incident angle (θ) d) Reflection modelling from Lidar

Figure 4.3. Scattering Profiles: Figure shows the effect of changing parameters on the scattering profiles. For each plot, we only change one parameter, keeping others fixed.

4.4.2 Obtaining the power of reflections

Lidar point cloud provides us the points of interaction, but to faithfully capture the radar behavior, it is important to model the power of reflections very accurately. The two main factors that dictate the reflections from a surface are the surface geometry and its material properties. In simple words, the geometry or orientation of the surface determines the direction of reflection while the material properties such as permittivity, determines the amount of reflection power. For obtaining the orientation of the surface, we calculate the outward normal vectors from object surfaces using RANSAC plane fitting method in the neighbourhood of each lidar point. These normal vectors provide the orientation of the local surface of object at each lidar point. However, the information about the material properties is still absent from the lidar point cloud. In SHENRON, we provide a novel way to obtain these properties by using camera images, which are usually present in most large scale datasets along with the lidar point clouds.

Material from camera: Recent advancements in camera semantic scene segmentation have shown the robustness of identifying and segmenting objects in a camera image [93]. We use a pretrained neural network [22] for obtaining the semantic maps of the camera image. After obtaining these semantics, we project the lidar point cloud on the camera image and associate each point with its corresponding object class from the semantic map. In the next sections, we will list the classes we used and how they were mapped to radar usable object models.

4.4.3 Modeling Radar Reflections from Lidar

In this section, first, we will describe how we physically model the radar reflections from the lidar point cloud, given the material properties. Finally, towards the end, we will describe how we mapped the classes obtained from camera to the required material properties using these physical models.

Interaction of EM waves with surfaces

We treat the EM wave interaction with a surface as a combination of penetrated energy and reflected energy. The reflected energy is further divided into specular energy and diffuse scattered energy. For an ideal planar surface, laws of reflection state that all reflected energy should go into specular direction, but due to the roughness of surfaces a lot of energy actually scatters after reflection. In our experiments, we found that this split sufficiently models the real radar data.

First, we determine the total incident power on each point of the point cloud. Each point in the point cloud is expanded to a surfel. The surfel size is chosen based on the angular separation of lidar rays in vertical and horizontal directions. Incident power at each surfel in the point cloud is then estimated based on its distance and surface area. Now, the fraction of power reflected from each such surfel is determined by the material and the orientation of the local surface. The incident power is first split into penetrated and reflected power: $|\bar{E}_i|^2 = |\bar{E}_R|^2 + |\bar{E}_P|^2$, where $|\bar{E}_i|^2$ is the incident power, $|\bar{E}_R|^2$ is the total reflected power and $|\bar{E}_P|^2$ is the penetrated power. We use Fresnel coefficients of reflections and transmission to determine the power distribution among both terms: $\Gamma(\varepsilon,\theta) = |\bar{E}_R|^2/|\bar{E}_i|^2$ where $\Gamma(\varepsilon,\theta)$ is the Fresnel reflection co-efficient, which is dependent on the permittivity (ε) and incident angle with respect to normal ($\theta = \theta_i$). Note that the permittivity (ε) is the first material property we need to determine for any material. Next, the total reflected power is split into the scatter and specular components.

Specular Power

We refer to specular reflections as the energy governed by the laws of reflection where the incidence angle equals the angle of reflection. At mmwave frequencies, these types of reflections carry a significant amount of power. However, the surface roughness causes some power to be spent in scattering: $|\tilde{E}_{SP}|^2 = |\bar{E}_i|^2 * \Gamma^2 * R^2$ where R^2 is the ratio of specular power ($|\tilde{E}_{SP}|^2$) over total reflected power ($|\bar{E}_R|^2$). The value of reduction ratio R depends on the surface roughness of a material [62]: $R = e^{-0.5(4\pi\sigma_h \cos\theta/\Lambda)^2}$ where σ_h is the standard deviation of surface roughness and also the second material property we need to determine for a material. Λ is the wavelength of EM wave. Since in radars, the receiver is situated close to transmitter, we only consider the specular power return when the incident ray direction is within 2 deg threshold of the normal.

Scattered Power

Scattered power is composed of the non-specular component of the total reflected power which is caused by surface roughness. The angular distribution of the scattered power has been studied comprehensively in past work [29]. We use the double lobe back scatter model defined for scatter reflections, as it contributes prominently in the case of co-located transmitter and receivers which is the case in radars. The double lobe model is given by

$$L^2 = \Lambda * \Big(\frac{1+\cos\psi_r}{2}\Big)^{\alpha_R} + (1-\Lambda)\Big(\frac{1+\cos\psi_i}{2}\Big)^{\alpha_I} \tag{4.1}$$

$$|\tilde{E}_{DS}|^2 = P_i * \Gamma^2 * (1-R^2) * L^2 \tag{4.2}$$

where L is the power distribution of scattered power ($|\tilde{E}_{DS}|^2$) as the function of angle; ψ_r is the angle between query and reflected ray direction, ψ_i is the angle between query and incident ray direction, Λ is the repartition factor between the amplitudes of front and back scatter lobe (Figure 4.3d). The α_I, α_R parameter controls the lobe shape and is set from empirical evidence [29]. In case of radar, the query direction is same as the incident ray direction hence

Table 4.1. Mapping of camera semantic classes to material parameters used in simulation

Camera Class	Material	Permittivity (ε)	Roughness (mm)
Trees	Wood	2	1.7
Road, Walls	Concrete	5.24	1.7
Pedestrians	Humans	2	0.1
Cars, Trucks	Metals	100000	0.05

$\psi_r = 2\theta$ and $\psi_i = 0$, θ is the incident angle (also the query angle). Putting these values in equation 4.1:

$$L^2 = (\Lambda * (\cos^2\theta)^{\alpha_R} + (1-\Lambda)) \quad \alpha_R = 1,2,\ldots \tag{4.3}$$

Determining Material Parameters

With the above formulation, we find that each material can be described by using the correct parameter values in above model. Specifically, for each material, we determine σ (permittivity) and σ_H (roughness standard deviation) to calculate the reflected energy. To determine these values for each material, we run real-life experiments and empirically tune the value that best describes the real-world power patterns. Specifically, we collect data for different materials at different orientations and distance from both lidar and radar. Then we compare the power profiles generated by our simulations and real world radar. We compare the total power and correlation values. We use this comparison to set the parameters of our model. Figure 4.3 shows how the scattering profile changes by choosing different values for parameters in our model. Table 4.1 contains the values we used in our simulation.

4.4.4 Digital Signal Processing (DSP) implementation

Having interpreted the scene as a collection of points which reflect energy through specular or scatter means, the received signal is generated as a sum of individual reflections to get radar ADC samples. In SHENRON, the signal processing chain is modeled after a MIMO-FMCW [16] radar. This block is flexible which allows various radar system parameters including radar waveform to be changed.

To generate the ADC samples for a MIMO (Multiple Input Multiple Output) radar, reflected waves from each point are constructed with delay based on their time of flight (ToF), amplitude based on path loss and specular/scatter reflection power return. In the MIMO setup, the ToF at each receive antenna in the array is calculated, accounting for the difference in path lengths with respect to each other. Finally, the attenuated and phase shifted reflected waves are cumulatively added with AWGN (Additive White Gaussian Noise) noise to derive the complete radar-received signal. This creates a 2D matrix, with respect to time and antenna dimension, providing range and angle information of objects in the scene. The DSP block computes the reflected signal from each point *i* as an attenuated and ToF delayed FMCW waveform equation. The final received waveform is the sum of all individual reflections and additive gaussian noise (4.4):

$$Y = \sum_{i=1}^{N} \left(L_i^{refl} * L_i^{path} * e^{j2\pi(fc+0.5*k*(t-\tau_i))(t-\tau_i)} \right) + n_{AWGN} \tag{4.4}$$

Where L_i^{refl} is the specular/scatter loss, L_i^{path} is the path loss, τ_i is the ToF, f_c is the carrier frequency and k is the frequency change rate of the FMCW waveform. n_{AWGN} is the AWGN noise whose amplitude is obtained from real data.

Radar doppler estimation: The doppler information of radar data provides valuable insights into the movement of objects relative to the radar. It is measured using the phase changes across multiple radar chirps within a frame. To simulate accurate doppler characteristics, lidar point clouds are needed at a frame rate that matches that of the chirps, which is typically of the order of a few microseconds. However, lidar datasets often lack such high frame rates, necessitating the use of labels present in the datasets to track objects through frames. By determining the position of an object in different frames and the time between frames, the instantaneous speed can be calculated. Additionally, the relative doppler values are obtained by using the ego speed of radar measured by an IMU sensor. The relative speed of each point with respect to the radar is then used to calculate the change in phase generated due to that speed across chirps, thus simulating the doppler.

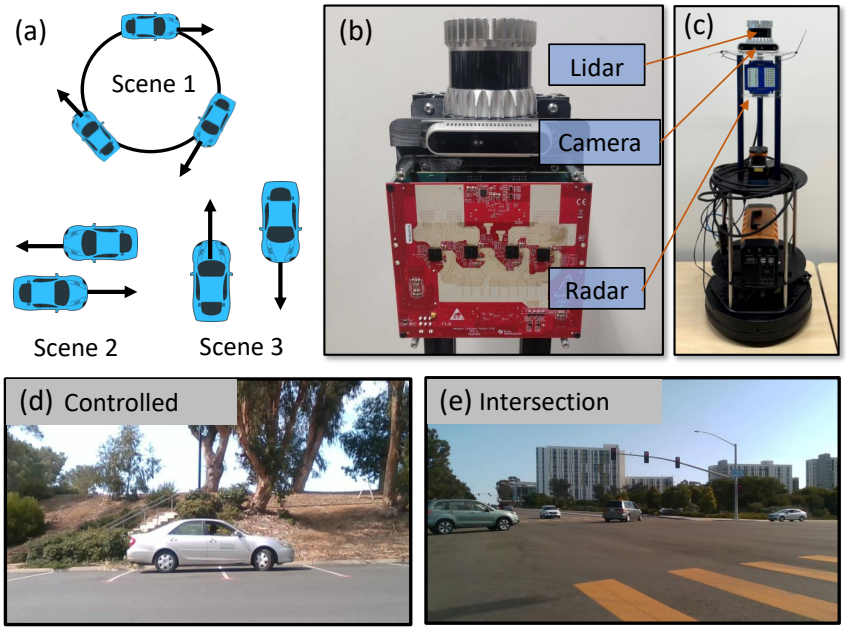


Figure 4.4. a) Different maneuvers for controlled experiments. b) Sensor setup with TI 77GHz radar for outdoor experiments. c) Sensor setup for 24GHz radar for the indoor case study. d) Example scene from controlled experiments. e) Example scene from a busy intersection.

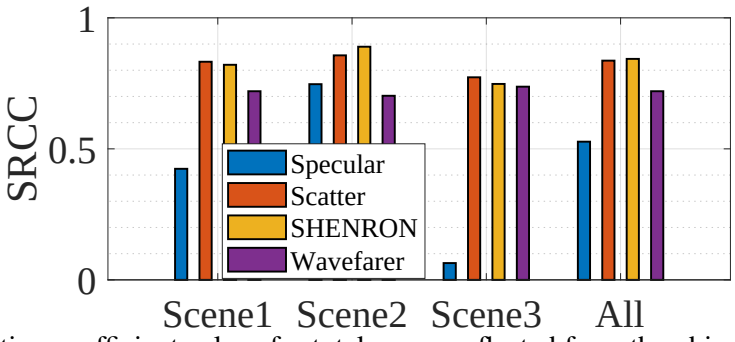


Figure 4.5. Correlation coefficient values for total power reflected from the object. We also show the effect of considering specular and scattering profiles on the correlation.

4.5 Evaluation

4.5.1 Evaluating Power Profiles of SHENRON

Experimental setup: We collect data using a co-located setup of 64-channel ouster lidar, Intel Real Sense Camera and a TI imaging radar as shown in figure 4.4). To exhaustively cover all the views of an object a radar would see, we consider 3 types of scenes in a controlled experiment. For *scene 1* the object moves in a circle in front of the sensor setup. In this scene, the sensor sees all the orientations of the object allowing us to evaluate quality of simulations

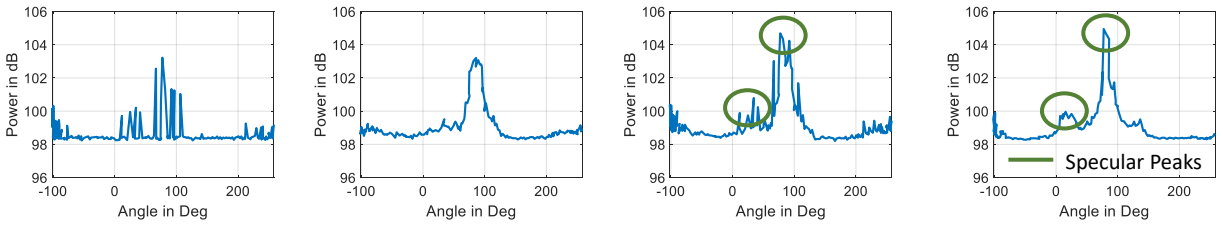


Figure 4.6. Power profile of a car compared against real data for different reflection models. Specular model (a) provides correct power at specular incidence (+- 90), while scattering model (b) gets the power at almost all angles of incidence. SHENRON (c) accurately models the power profile by suitably combining the specular and scattering reflection models.

with respect to different orientations. For *scene 2* the object moves in cross-range direction and at multiple distances. For Scene 3, the object approaches or moves away from the setup. Scene 2 and 3 represent a more common view of the objects as generally encountered while driving. Figure 4.4 shows a schematic of each of these scenes. For each scene we collect 8 different runs. Each run consists of 200 frames of sensor data from each sensor, collected at a rate of 10Hz. We repeat all these experiments for both a car and a pedestrian.

Correlation of Power Return: For each scene with car, we calculate the pearson's correlation between the power profile generated by simulated and real data. For SHENRON's simulated power profile, we also show the effect of using the specular reflection model and scattering model separately to better understand the contribution of each component. While comparing with real data, commercial simulator Wavefarer [103] obtains a correlation value of 0.7199, while SHENRON maintains a high correlation value of **0.8433** when averaged among all scenarios (figure 4.5). One thing to note is that scattering alone gets a high correlation value, but only by using both specular and scatter profiles, SHENRON can obtain the correct absolute power values. This is also shown in Figure 4.6 that shows the power profile for a single run from scene 1. When the car is at 90 degree orientation, the contribution of specular component is significant, while in all other orientations, scattering accounts for most of the power. This shows the importance of considering both specular and scattering components in radar simulations, to accurately model the power reflected by an object.

Qualitative comparison We present a qualitative analysis of the contributions of each

Figure 4.7. Qualitative ablation study for SHENRON. The data corresponds to the image shown in figure 4.4 scene 1. The figure shows how the different components (a-c) of SHENRON (d), models different characteristics of the real world scene (e).

component of SHENRON in Figure 4.7. Four different cases are considered: no material (where everything is set as metal), specular only, scattering only, and complete SHENRON. The scene includes a car in front of the radar, with trees and concrete structures in the background (Figure 4.4). In the no material case, the simulator cannot differentiate the power of reflections based on materials, resulting in similar power returns for all points. In the specular only case, the specular reflection peak is accurately captured, but the power distribution across the car's body and the background objects (trees and concrete structures) is not represented. In the scattering only case, the power distribution is properly estimated for both the car and background objects, but the strong specular peak is missing. Finally, SHENRON accurately models both specular and scattering power returns, producing simulations that closely resemble real-world data for both the car and background objects.

4.5.2 Predicting Object Detection Performance

We assess the effectiveness of SHENRON in a real-world automotive task, specifically object detection, which is crucial for the perception stack in autonomous driving. Here, we consider a case of a general automotive scenario, i.e., a busy intersection. We collect data for multiple vehicles passing through the intersection in all directions. Figure 4.4e shows the environment where we collect data. We divide the dataset into an 8:2 ratio for training and testing. The train and test data comes from different recordings at the intersection.

Table 4.2. Average Precision (AP) performance: The results show that object detection trained on simulated data works almost as good as one trained on actual real data.

Dataset →	**Real**	**Simulated**	**Noise**	**RadarSimPy**
Average Precision	57.5	57.8	15.0	16.38

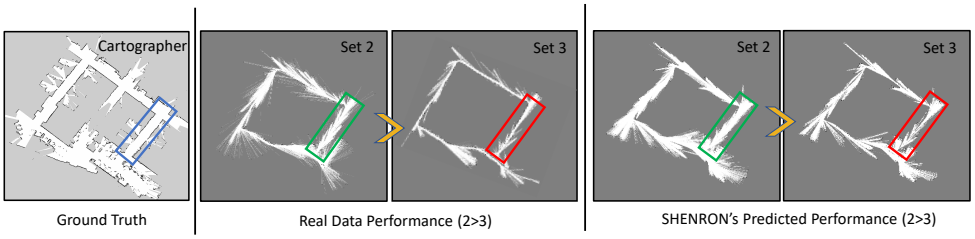


Figure 4.8. Sim2Real Predictivity. The figure displays occupancy maps for parameter sets 2 and 3 using both real radar data and simulated data. A highlighted box indicates a corridor mapped differently under these parameter sets. Notably, SHENRON accurately predicts the relative performance of these sets, indicating that set 2 yields superior mapping results.

Experiment Details and Results: We train three networks, one each on the real and simulated data and one on Gaussian noise. For the noise experiment, the test set is from real data. We use U-NET [91] backbone with detection heads for object detection. We report average precision (AP) metric with IOU threshold of 0.5. Table 4.2 shows the result of this experiment. The network trained only on simulated data achieves a very close performance to the network trained on the real data. Only by accurately modeling the environment and vehicle radar reflections SHENRON creates data that can be used to test deep learning networks for deployment in the real world. The marginally higher value in simulated data can be explained by the fact that both the ground truth (GT) boxes and simulated data are generated using lidar data and hence there is no misalignment between GT and simulated radar data, which is not the case for real data. Further, we see that the network trained solely on Gaussian noise performs significantly worse than the one trained on simulated or real data. This underscores the importance of accurate object representations in the input data for effective model training, validating that the simulations successfully mimic the real-world data. Notably, our approach SHENRON, enables the direct use of real-world lidar data for simulations, whereas using existing MIMO simulator *RadarSimPy* [Link] with sparse lidar data converted to mesh using poisson surface reconstruction predicts performance close to that of noise (Table 4.2)

4.5.3 Case Study - Dataset Augmentation Object Detection

In this case study, we demonstrate the effectiveness of using SHENRON for data generation to enhance the available training data. We employ a pre-existing radar-based object detection network called Radatron [70] for our task. To conduct our experiments, we utilize all the data collected in scene 1-3 for the controlled experiment as described previously (Figure 4.4a & 4.4d). The collected data is divided equally into training and testing datasets. The performance of Radatron on this data is presented in Table 4.3.

Subsequently, we leverage the lidar data from the training set to generate simulated radar data. This simulated data is then incorporated into the training process. We evaluate the trained model on the original test set from real data. The outcomes reveal a notable 15.023 average precision (IoU 0.5) improvement in performance. This indicates the successful application of SHENRON in augmenting radar datasets, leading to better generalization and enhanced performance.

Table 4.3. Object detection performance using Radatron [70]: the network trained by augmenting the dataset using simulated data gives better performance on real data.

Training Data	Test Data	Average Precision
Real	Real	31.78
Real + Simulated	Real	47.033

4.5.4 Case Study - Radar Parameter tuning

We perform a case study to show how SHENRON can be used to solve the ”cold-start” problem in radars. The aim is to understand if SHENRON can predict the relative performance while using radar with different parameters. We consider the crucial application of indoor mapping using radar that can aid in navigation during low visibility scenarios such as fire. To perform this study, we collect data for an office space of around 50m x 50m, using a system of lidar and camera sensors mounted on a turtle bot, as illustrated in Figure 4.4. We use the cartographer algorithm [48] to obtain a ground truth map which will be used to assess the quality

of map generated using radar data. We collected approximately 1500 frames at a rate of 10 frames per second.

We first use lidar and camera data as input to SHENRON to generate simulated radar point clouds. We performed simulations for different parameter sets as listed in Table 4.4. Specifically, we changed the bandwidth used (which determines range resolution), sampling rate and number of samples per chirp (determining maximum range), time between chirps (determining maximum doppler) and number of chirps per frame (determining doppler resolution). We convert the simulated radar data into an occupancy map using the matlab's *buildMap()* function. We then compare the generated map against the ground truth map created by cartographer, using the Intersection over Union (IoU) metric. As expected, different parameter sets generate varying quality of maps, shown by the IoU values obtained in Table 4.4. We find that the parameter sets 1, 2 and 4 provide the most optimal values.

Now, to validate the predicted performance of different parameters by SHENRON, we also collect data using a real radar mounted on the same turtle bot. We perform 6 different data collection runs in the same office space, one for each parameter set from Table 4.4. Remarkably, the relative performance between different parameter sets follows the same pattern as that predicted by SHENRON (Table 4.4). It is clear from this evaluation that parameter set 1, 2 and 4 are the optimal sets for this application compared to sets 3,5 and 6. Note that the purpose of this evaluation is to show how SHENRON can predict the performance of different parameter sets before even collecting real data. A more exhaustive parameter search would actually reveal the most optimal parameter sets for use. We also calculate the Sim-vs-Real Correlation Coefficient (SRCC) which was proposed by Kadian et al. [56] to measure the predictive ability of a simulation. The SRCC of this study comes out to be **0.93443**, demonstrating the effectiveness of using SHENRON in predicting the relative performance while using radar with different parameters, in a quick and efficient manner.The difference in IoU values between real and simulated data is primarily due to IoU's sensitivity to map alignment with lidar ground truth data. Since SHENRON utilizes lidar as input, it achieves higher IoU values because of perfect

Table 4.4. *Parameter sets used for indoor study*. BW: Bandwidth (MHz); N: samples per chirp; F_s: sampling rate (MHz); N_{chirps}: Chirps per frame; T_c: chirp repetition interval (ms); R: Range; D: Doppler; *res*: resolution; *max*: maximum

Set No.	Set 1	Set 2	Set 3	Set 4	Set 5	Set 6
BW	250	250	250	125	125	250
N	256	256	256	256	128	64
F_s	2.00	2.00	2.00	2.00	1.00	0.50
N_{chirps}	128	128	32	128	64	128
T_c	0.250	0.750	3.000	0.750	0.750	0.750
R_{res}	0.60	0.60	0.60	1.20	1.20	0.60
R_{max}	153.60	153.60	153.60	307.20	153.60	38.40
D_{res}	0.194	0.065	0.065	0.065	0.129	0.065
D_{max}	12.44	4.15	1.04	4.16	4.16	4.15
IoU_{sim}	**0.64**	**0.65**	0.47	**0.65**	0.58	0.62
IoU_{real}	**0.57**	**0.51**	0.27	**0.50**	0.37	0.48

alignment. Figure 4.8 shows the qualitative output for set 2 and 3. The total time spent in going through all the parameter sets in simulations was around 30 minutes, while the same real world test took an entire working day. This is only for 6 parameter sets, scaling this for 100+ sets will clearly be infeasible for real world tests. Hence, this case study clearly shows how SHENRON can significantly help in the ubiquitous deployment of radars for several indoor/outdoor applications, which otherwise would be a daunting task.

4.5.5 Time and RAM usage:

We also compare the compute resource utilization of SHENRON against commercial solvers. We run all the simulations in a Windows PC with 32 GB RAM and 3.19 GHz processor. For SHENRON, we run a multi-antenna multi-chirp simulation with the entire scene given by lidar point cloud. For Wavefarer, we run a simulation for single channel radar and a car mesh with all the material properties. The total memory used in the simulation and the time taken for the simulation is provided in Table 4.5. SHENRON uses **20x** time less memory and **1000x** times less time for each simulation. Hence, by using lidar point clouds as input, SHENRON can provide highly accurate results in a more efficient and scalable manner.

Table 4.5. Time and RAM usage

	Memory Usage	Time per simulation
Wavefarer	≈10GB	3146s
SHENRON	≈0.5GB	3s

4.6 Limitation and future works

The radar data generated by SHENRON can be used for various applications, such as object detection and radar parameter tuning, as shown in the paper. However, it should be noted that the simulator's reliance on lidar and camera data means that it cannot use datasets with distorted lidar or camera data due to poor weather conditions. The primary objective of SHENRON is to provide researchers with access to vast amounts of radar data to facilitate the development and testing of radar algorithms. Once developed, these algorithms can simply be employed in challenging weather conditions, as radar data is impervious to weather-related interference. A potential avenue for future research is to expand the range of materials considered by the system using the method proposed. This could enable the subdivision of different parts of the same semantic object into sub-classes, such as car windows and tires, or address material interaction challenges, particularly for materials like thin cardboard that allow mmwave to pass but yield different results for lidar signals. Secondly, SHENRON only considers direct reflections of radar signals. In the future, we can use splat-based techniques to perform ray-tracing directly on lidar point clouds and even capture multipath reflections.

Chapter 4, in full, is a reprint of the material as it appears in "SHENRON-Scalable, High fidelity and EfficieNt Radar simulatiON" by Kshitiz Bansal, Gautham Reddy and Dinesh Bharadia, which appears in IEEE Robotics and Automation Letters (2023). The dissertation author was the primary investigator and author of this paper.

Chapter 5

Conclusion

This dissertation has delved into the transformative potential of radar-based perception technologies for autonomous navigation and driving systems. Recognizing the limitations of traditional light-based sensors in adverse weather conditions, our research has proposed and explored a paradigm shift towards radar as the primary sensor. The four fundamental aspects investigated—high-quality perception, semantic understanding, contextual information, and scalable deployment—have collectively contributed to the advancement of autonomous perception.

Our work Pointillism addresses challenges associated with specularity and sparsity in radar point clouds, creating a high-quality perception region through strategic combinations of low-resolution radars. Spatial averaging techniques filter out noise, and a deep learning system accurately converts radar data into bounding boxes, enhancing the perception of the environment.

Semantic understanding is improved through a novel camera radar fusion system, RadSegNet, integrating radar point clouds with missing texture and color information from cameras. The proposed fusion philosophy breaks feature dependence, ensuring robust performance even in challenging conditions such as adverse weather or occlusion.

Recognizing the importance of environmental context, the dissertation addresses the challenge of identifying traffic infrastructure using radar alone. The introduction of specialized millimeter-wave backscatter tags, R-Fiducial, enables COTS radars to obtain this environmental

context. These tags are made possible by incorporating the Van-Atta array and RF switches which enhances radar visibility and enables precise localization without modifying radar hardware.

To facilitate practical deployment, an open-source radar simulation framework has been introduced, allowing for the efficient simulation of high-fidelity Multiple Input Multiple Output (MIMO) radar data using lidar point clouds and camera images. This framework serves as a transformative tool for researchers, enabling rapid iterations across the extensive radar parameter space and aiding in the identification of optimal configurations for diverse applications.

In summary, the comprehensive exploration of multi-radar systems, radar-camera fusion, smart infrastructure augmentation, and the innovative radar simulation framework collectively contributes to the progress of autonomous systems, particularly in navigating challenging environmental conditions. By addressing current limitations and proposing novel solutions, this dissertation not only advances the field of autonomous perception but also sets the stage for future breakthroughs and continued innovation in this rapidly evolving domain.

The systems presented in this study open up avenues for numerous future applications. Pointillism and RadSegNet, initially designed for road-related tasks, have potential applications in fields such as precision agriculture and construction. Additionally, R-fiducial, initially developed as a marker for radars, has broader use cases, including calibrating radars with other sensors like cameras and lidar, which already have specific fiducial markers. SHENRON contributes by creating a radar sensor model that can seamlessly integrate with existing simulators. An example of such integration is with the widely used CARLA simulator, offering a comprehensive evaluation of radar-based algorithms and their impact on downstream tasks like navigating through traffic. This integration also enables the assessment of sensing algorithms in specific accident-prone scenarios.

Bibliography

[1] Closer than you think: Know your stopping distances. https://www.theaa.com/breakdown-cover/advice/stopping-distances.

[2] FMCW fundamentals in TI radar. https://training.ti.com/mmwave-training-series.

[3] Intel Real Sense D415 depth camera. https://www.intelrealsense.com/.

[4] LiDAR vs. RADAR. https://www.sensorsmag.com/components/lidar-vs-radar.

[5] OS-1, Ouster 16 channel LiDAR. https://ouster.com/products/os1-lidar-sensor/.

[6] Perception matters: How deep learning enables autonomous vehicles to understand their environment. https://blogs.nvidia.com/blog/2018/08/10/autonomous-vehicles-perception-layer/.

[7] Report of traffic collision involving an autonomous vehicle (OL 316). https://www.dmv.ca.gov/portal/dmv/detail/vr/autonomous/autonomousveh_ol316.

[8] Scalabel - bekley deepdrive. https://www.scalabel.ai/.

[9] Self-driving cars can handle neither rain nor sleet nor snow. https://www.bloomberg.com/news/articles/2018-09-17/self-driving-cars-still-can-t-handle-bad-weather.

[10] Uber self-driving car crash: What really happened. https://www.forbes.com/sites/meriameberboucha/2018/05/28/uber-self-driving-car-crash-what-really-happened/#11619a944dc4.

[11] Waymo open dataset: An autonomous driving dataset, 2019.

[12] Fadel Adib, Zachary Kabelac, and Dina Katabi. Multi-person localization via RF body reflections. In *12th {USENIX} Symposium on Networked Systems Design and Implementation ({NSDI} 15)*, pages 279–292, 2015.

[13] Fadel Adib, Hongzi Mao, Zachary Kabelac, Dina Katabi, and Robert C Miller. Smart homes that monitor breathing and heart rate. In *Proceedings of the 33rd annual ACM conference on human factors in computing systems*, pages 837–846. ACM, 2015.

[14] Yasin Almalioglu, Mehmet Turan, Chris Xiaoxuan Lu, Niki Trigoni, and Andrew Markham. Milli-rio: Ego-motion estimation with low-cost millimetre-wave radar. *IEEE Sensors Journal*, 2020.

[15] Kang Min Bae, Namjo Ahn, Yoon Chae, Parth Pathak, Sung-Min Sohn, and Song Min Kim. Omniscatter: Extreme sensitivity mmwave backscattering using commodity fmcw radar. In *Proceedings of the 20th Annual International Conference on Mobile Systems, Applications and Services*, MobiSys '22, page 316–329, New York, NY, USA, 2022. Association for Computing Machinery.

[16] Kshitiz Bansal, Keshav Rungta, Siyuan Zhu, and Dinesh Bharadia. Pointillism: Accurate 3d bounding box estimation with multi-radars. In *Proceedings of the 18th Conference on Embedded Networked Sensor Systems*, pages 340–353, 2020.

[17] 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 *2020 IEEE International Conference on Robotics and Automation (ICRA)*, pages 6433–6438. IEEE, 2020.

[18] Markus Bühren and Bin Yang. Automotive radar target list simulation based on reflection center representation of objects. In *Proc. Intern. Workshop on Intelligent Transportation (WIT)*, pages 161–166, 2006.

[19] 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.

[20] Simon Chadwick, Will Maddern, and Paul Newman. Distant vehicle detection using radar and vision. In *2019 International Conference on Robotics and Automation (ICRA)*, pages 8311–8317. IEEE, 2019.

[21] Shuo Chang, Yifan Zhang, Fan Zhang, Xiaotong Zhao, Sai Huang, Zhiyong Feng, and Zhiqing Wei. Spatial attention fusion for obstacle detection using mmwave radar and vision sensor. *Sensors*, 20(4):956, 2020.

[22] Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. Rethinking atrous convolution for semantic image segmentation. *arXiv preprint arXiv:1706.05587*, 2017.

[23] Ushemadzoro Chipengo, Peter M Krenz, and Shawn Carpenter. From antenna design to high fidelity, full physics automotive radar sensor corner case simulation. *Modelling and Simulation in Engineering*, 2018.

[24] L Daniel, D Phippen, E Hoare, A Stove, M Cherniakov, and M Gashinova. Low-thz radar, lidar and optical imaging through artificially generated fog. 2017.

[25] Andreas Danzer, Thomas Griebel, Martin Bach, and Klaus Dietmayer. 2D car detection in radar data with pointnets, 2019.

[26] Michael Darms, Paul Rybski, and Chris Urmson. Classification and tracking of dynamic objects with multiple sensors for autonomous driving in urban environments. In *2008 IEEE Intelligent Vehicles Symposium*, pages 1197–1202. IEEE, 2008.

[27] Rene Degen, Harry Ott, Fabian Overath, Christian Schyr, Mats Leijon, and Margot Ruschitzka. Methodical approach to the development of a radar sensor model for the detection of urban traffic participants using a virtualreality engine. *Journal of Transportation Technologies*, 11:179–195, 2011.

[28] Vittorio Degli-Esposti and Henry L Bertoni. Evaluation of the role of diffuse scattering in urban microcellular propagation. In *Gateway to 21st Century Communications Village. VTC 1999-Fall. IEEE VTS 50th Vehicular Technology Conference (Cat. No. 99CH36324)*, volume 3, pages 1392–1396. IEEE, 1999.

[29] Vittorio Degli-Esposti, Franco Fuschini, Enrico M. Vitucci, and Gabriele Falciasecca. Measurement and modelling of scattering from buildings. *IEEE Transactions on Antennas and Propagation*, 55(1):143–153, 2007.

[30] Maurizio di Bisceglie and Carmela Galdi. Cfar detection of extended objects in high-resolution sar images. *IEEE Transactions on geoscience and remote sensing*, 43(4):833–843, 2005.

[31] Jabbari Dinan. Spreading codes for direct sequence cdma and wideband cdma cellular networks. *IEEE Communications Magazine*, 36(9):48–54, 1998.

[32] Xinxin Du, Marcelo H Ang, Sertac Karaman, and Daniela Rus. A general pipeline for 3D detection of vehicles. In *2018 IEEE International Conference on Robotics and Automation (ICRA)*, pages 3194–3200. IEEE, 2018.

[33] Martin Ester, Hans-Peter Kriegel, Jörg Sander, Xiaowei Xu, et al. A density-based algorithm for discovering clusters in large spatial databases with noise. In *Kdd*, volume 96, pages 226–231, 1996.

[34] Kevin Eykholt, Ivan Evtimov, Earlence Fernandes, Bo Li, Amir Rahmati, Chaowei Xiao, Atul Prakash, Tadayoshi Kohno, and Dawn Song. Robust physical-world attacks on deep learning visual classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1625–1634, 2018.

[35] Lorenzo Favalli, Paolo Gamba, T Gatti, and Alessandro Mecocci. Multi-radar data fusion for object tracking and shape estimation. *Signal processing*, 48(3):235–239, 1996.

[36] Florian Fembacher, Farhan Bin Khalid, Gabor Balazs, Dian Tresna Nugraha, and André Roger. Real-time synthetic aperture radar for automotive embedded systems. In *2018 15th European Radar Conference (EuRAD)*, pages 517–520. IEEE, 2018.

[37] 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*, 22(3):1341–1360, 2020.

[38] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the KITTI vision benchmark suite. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012.

[39] Thomas Gisder, Marc-Michael Meinecke, and Erwin Biebl. Synthetic aperture radar towards automotive applications. In *2019 20th International Radar Symposium (IRS)*, pages 1–10. IEEE, 2019.

[40] Shahzad Gishkori, David Wright, Liam Daniel, Marina Gashinova, and Bernard Mulgrew. Imaging moving targets for a forward-scanning automotive sar. *IEEE Transactions on Aerospace and Electronic Systems*, 56(2):1106–1119, 2019.

[41] Robert Gold. Optimal binary sequences for spread spectrum multiplexing (corresp.). *IEEE Transactions on Information Theory*, 13(4):619–621, 1967.

[42] Robert Gold. Maximal recursive sequences with 3-valued recursive cross-correlation functions (corresp.). *IEEE transactions on Information Theory*, 14(1):154–156, 1968.

[43] Christopher Grimm, Tai Fei, Ernst Warsitz, Ridha Farhoud, Tobias Breddermann, and Reinhold Haeb-Umbach. Warping of radar data into camera image for cross-modal supervision in automotive applications. *IEEE Transactions on Vehicular Technology*, 71(9):9435–9449, 2022.

[44] Junfeng Guan, Sohrab Madani, Suraj Jog, and Haitham Hassanieh. High resolution millimeter wave imaging for self-driving cars, 2019.

[45] Shirsendu Sukanta Halder, Jean-François Lalonde, and Raoul de Charette. Physics-based rendering for improving robustness to rain. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10203–10212, 2019.

[46] Lars Hammarstrand, Lennart Svensson, Fredrik Sandblom, and Joakim Sorstedt. Extended object tracking using a radar resolution model. *IEEE Transactions on Aerospace and Electronic Systems*, 48(3):2371–2386, 2012.

[47] 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.

[48] Wolfgang Hess, Damon Kohler, Holger Rapp, and Daniel Andor. Real-time loop closure in 2d lidar slam. In *2016 IEEE international conference on robotics and automation (ICRA)*, pages 1271–1278. IEEE, 2016.

[49] Nils Hirsenkorn, Paul Subkowski, Timo Hanke, Alexander Schaermann, Andreas Rauch, Ralph Rasshofer, and Erwin Biebl. A ray launching approach for modeling an fmcw radar system. In *2017 18th International Radar Symposium (IRS)*, pages 1–10. IEEE, 2017.

[50] Martin Holder, Philipp Rosenberger, Hermann Winner, Thomas D'hondt, Vamsi Prakash Makkapati, Michael Maier, Helmut Schreiber, Zoltan Magosi, Zora Slavik, Oliver Bringmann, et al. Measurements revealing challenges in radar sensor modeling for virtual validation of autonomous driving. In *2018 21st International Conference on Intelligent Transportation Systems (ITSC)*, pages 2616–2622. IEEE, 2018.

[51] Chen-Yu Hsu, Aayush Ahuja, Shichao Yue, Rumen Hristov, Zachary Kabelac, and Dina Katabi. Zero-effort in-home sleep and insomnia monitoring using radio signals. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 1(3):59, 2017.

[52] Chen-Yu Hsu, Rumen Hristov, Guang-He Lee, Mingmin Zhao, and Dina Katabi. Enabling identification and behavioral sensing in homes using radio reflections. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, page 548. ACM, 2019.

[53] Chen-Yu Hsu, Yuchen Liu, Zachary Kabelac, Rumen Hristov, Dina Katabi, and Christine Liu. Extracting gait velocity and stride length from surrounding radio signals. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, pages 2116–2126. ACM, 2017.

[54] Jyh-Jing Hwang, Henrik Kretzschmar, Joshua Manela, Sean Rafferty, Nicholas Armstrong-Crews, Tiffany Chen, and Dragomir Anguelov. Cramnet: Camera-radar fusion with ray-constrained cross-attention for robust 3d object detection. In *European Conference on Computer Vision*, pages 388–405. Springer, 2022.

[55] I Jouny. Target identification using multi-radar fusion. In *Automatic Target Recognition XVII*, volume 6566, page 65660M. International Society for Optics and Photonics, 2007.

[56] Abhishek Kadian, Joanne Truong, Aaron Gokaslan, Alexander Clegg, Erik Wijmans, Stefan Lee, Manolis Savva, Sonia Chernova, and Dhruv Batra. Sim2real predictivity: Does evaluation in simulation predict real-world performance? *IEEE Robotics and Automation Letters*, 5(4):6670–6677, 2020.

[57] Jinhyeong Kim, Youngseok Kim, and Dongsuk Kum. Low-level sensor fusion network for 3d vehicle detection using radar range-azimuth heatmap and monocular image. In *Proceedings of the Asian Conference on Computer Vision*, 2020.

[58] Youngseok Kim, Jun Won Choi, and Dongsuk Kum. Grif net: Gated region of interest fusion network for robust 3d object detection from radar point cloud and monocular image. In *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 10857–10864. IEEE, 2020.

[59] Jason Ku, Melissa Mozifian, Jungwook Lee, Ali Harakeh, and Steven L Waslander. Joint 3d proposal generation and object detection from view aggregation. In *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 1–8. IEEE, 2018.

[60] Matti Kutila, Pasi Pyykönen, Hanno Holzhüter, Michèle Colomb, and Pierre Duthon. Automotive lidar performance verification in fog and rain. In *2018 21st International Conference on Intelligent Transportation Systems (ITSC)*, pages 1695–1701. IEEE, 2018.

[61] Alex H Lang, Sourabh Vora, Holger Caesar, Lubing Zhou, Jiong Yang, and Oscar Beijbom. Pointpillars: Fast encoders for object detection from point clouds. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12697–12705, 2019.

[62] B. Langen, G. Lober, and W. Herzig. Reflection and transmission behaviour of building materials at 60 ghz. In *5th IEEE International Symposium on Personal, Indoor and Mobile Radio Communications, Wireless Networks - Catching the Mobile Future.*, volume 2, pages 505–509 vol.2, 1994.

[63] Xin Li, Weiwen Deng, Sumin Zhang, Yaxin Li, Shiping Song, Shanshan Wang, and Guanyu Wang. Research on millimeter wave radar simulation model for intelligent vehicle. *International Journal of Automotive Technology*, 21(2):275–284, 2020.

[64] Zhuqi Li, Can Wu, Sigurd Wagner, James C Sturm, Naveen Verma, and Kyle Jamieson. Reits: Reflective surface for intelligent transportation systems. In *Proceedings of the 22nd International Workshop on Mobile Computing Systems and Applications*, pages 78–84, 2021.

[65] Teck-Yian Lim, Amin Ansari, Bence Major, Daniel Fontijne, Michael Hamilton, Radhika Gowaikar, and Sundar Subramanian. Radar and camera early fusion for vehicle detection in advanced driver assistance systems. In *Machine Learning for Autonomous Driving Workshop at the 33rd Conference on Neural Information Processing Systems*, 2019.

[66] 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.

[67] Jakob Lombacher, Markus Hahn, Jürgen Dickmann, and Christian Wöhler. Potential of radar for static object classification using deep learning methods. In *2016 IEEE MTT-S International Conference on Microwaves for Intelligent Mobility (ICMIM)*, pages 1–4. IEEE, 2016.

[68] Chris Xiaoxuan Lu, Stefano Rosa, Peijun Zhao, Bing Wang, Changhao Chen, John A Stankovic, Niki Trigoni, and Andrew Markham. See through smoke: robust indoor mapping with low-cost mmwave radar. In *MobiSys*, pages 14–27, 2020.

[69] Takashi Machida and Takashi Owaki. Rapid and precise millimeter-wave radar simulation for adas virtual assessment. In *2019 IEEE Intelligent Transportation Systems Conference (ITSC)*, pages 431–436. IEEE, 2019.

[70] Sohrab Madani, Jayden Guan, Waleed Ahmed, Saurabh Gupta, and Haitham Hassanieh. Radatron: Accurate detection using multi-resolution cascaded mimo radar. In *Computer*

Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXIX, pages 160–178. Springer, 2022.

[71] 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*, pages 0–0, 2019.

[72] Daniel Maturana and Sebastian Scherer. VoxNet: A 3D convolutional neural network for real-time object recognition.

[73] Mohammad Hossein Mazaheri, Alex Chen, and Omid Abari. mmtag: a millimeter wave backscatter network. In *Proceedings of the 2021 ACM SIGCOMM 2021 Conference*, pages 463–474, 2021.

[74] Petar Medeđović, Mladen Veletić, and Željko Blagojević. Wireless insite software verification via analysis and comparison of simulation and measurement results. In *2012 Proceedings of the 35th International Convention MIPRO*, pages 776–781. IEEE, 2012.

[75] Amanda Metzner and Thanuka Wickramarathne. On multi-sensor radar configurations for vehicle tracking in autonomous driving environments. In *2018 21st International Conference on Information Fusion (FUSION)*, pages 1–8. IEEE, 2018.

[76] 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.

[77] 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.

[78] Michael Meyer and Georg Kuschk. Deep learning based 3D object detection for automotive radar and camera. In *2019 16th European Radar Conference (EuRAD)*, pages 133–136. IEEE, 2019.

[79] Ramin Nabati and Hairong Qi. Rrpn: Radar region proposal network for object detection in autonomous vehicles. In *2019 IEEE International Conference on Image Processing (ICIP)*, pages 3093–3097. IEEE, 2019.

[80] Ramin Nabati and Hairong Qi. Centerfusion: Center-based radar and camera fusion for 3d object detection. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1527–1536, 2021.

[81] Felix Nobis, Maximilian Geisslinger, Markus Weber, Johannes Betz, and Markus Lienkamp. A deep learning-based radar and camera sensor fusion architecture for object detection. In *2019 Sensor Data Fusion: Trends, Solutions, Applications (SDF)*, pages 1–7. IEEE, 2019.

[82] John Nolan, Kun Qian, and Xinyu Zhang. Ros: passive smart surface for roadside-to-vehicle communication. In *Proceedings of the 2021 ACM SIGCOMM 2021 Conference*, pages 165–178, 2021.

[83] Andras Palffy, Julian FP Kooij, and Dariu M Gavrila. Occlusion aware sensor fusion for early crossing pedestrian detection. In *2019 IEEE Intelligent Vehicles Symposium (IV)*, pages 1768–1774. IEEE, 2019.

[84] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[85] Scott Drew Pendleton, Hans Andersen, Xinxin Du, Xiaotong Shen, Malika Meghjani, You Hong Eng, Daniela Rus, and Marcelo H Ang. Perception, planning, control, and coordination for autonomous vehicles. *Machines*, 5(1):6, 2017.

[86] Michail N Petsios, Emmanouil G Alivizatos, and Nikolaos K Uzunoglu. Solving the association problem for a multistatic range-only radar target tracker. *Signal Processing*, 88(9):2254–2277, 2008.

[87] Charles R. Qi, Wei Liu, Chenxia Wu, Hao Su, and Leonidas J. Guibas. Frustum pointnets for 3D object detection from RGB-D data. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Jun 2018.

[88] Charles R. Qi, Hao Su, Matthias NieBner, Angela Dai, Mengyuan Yan, and Leonidas J. Guibas. Volumetric and multi-view cnns for object classification on 3D data. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun 2016.

[89] Charles R. Qi, Li Yi, Hao Su, and Leonidas J. Guibas. PointNet++: Deep hierarchical feature learning on point sets in a metric space, 2017.

[90] Hermann Rohling. Radar cfar thresholding in clutter and multiple target situations. *IEEE transactions on aerospace and electronic systems*, (4):608–621, 1983.

[91] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*, pages 234–241. Springer, 2015.

[92] Christos Sakaridis, Dengxin Dai, and Luc Van Gool. Semantic foggy scene understanding with synthetic data. *International Journal of Computer Vision*, 126(9):973–992, 2018.

[93] Christos Sakaridis, Dengxin Dai, and Luc Van Gool. Acdc: The adverse conditions dataset with correspondences for semantic driving scene understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10765–10775, 2021.

[94] Alexander Scheel, Christina Knill, Stephan Reuter, and Klaus Dietmayer. Multi-sensor multi-object tracking of vehicles using high-resolution radars. In *2016 IEEE Intelligent Vehicles Symposium (IV)*, pages 558–565. IEEE, 2016.

[95] Karin Schuler, Denis Becker, and Werner Wiesbeck. Extraction of virtual scattering centers of vehicles by ray-tracing simulations. *IEEE Transactions on Antennas and Propagation*, 56(11):3543–3551, 2008.

[96] Ole Schumann, Markus Hahn, Jürgen Dickmann, and Christian Wöhler. Semantic segmentation on radar point clouds. *2018 21st International Conference on Information Fusion (FUSION)*, pages 2179–2186, 2018.

[97] Christian Schüßler, Marcel Hoffmann, Johanna Bräunig, Ingrid Ullmann, Randolf Ebelt, and Martin Vossiek. A realistic radar ray tracing simulator for large mimo-arrays in automotive environments. *IEEE Journal of Microwaves*, 1(4):962–974, 2021.

[98] E Sharp and M Diab. Van atta reflector array. *IRE Transactions on Antennas and Propagation*, 8(4):436–438, 1960.

[99] Marcel Sheeny, Emanuele De Pellegrin, Saptarshi Mukherjee, Alireza Ahrabian, Sen Wang, and Andrew Wallace. Radiate: A radar dataset for automotive perception in bad weather. In *2021 IEEE International Conference on Robotics and Automation (ICRA)*, pages 1–7. IEEE, 2021.

[100] Shaoshuai Shi, Chaoxu Guo, Li Jiang, Zhe Wang, Jianping Shi, Xiaogang Wang, and Hongsheng Li. PV-RCNN: Point-voxel feature set abstraction for 3D object detection. *arXiv preprint arXiv:1912.13192*, 2019.

[101] Shaoshuai Shi, Xiaogang Wang, and Hongsheng Li. PointRCNN: 3D object proposal generation and detection from point cloud, 2018.

[102] Xian Shuai, Yulin Shen, Yi Tang, Shuyao Shi, Luping Ji, and Guoliang Xing. millieye: A lightweight mmwave radar and camera fusion system for robust object detection. In *Proceedings of the International Conference on Internet-of-Things Design and Implementation*, pages 145–157, 2021.

[103] Gregory Skidmore, Tarun Chawla, and Gary Bedrosian. Combining physical optics and method of equivalent currents to create unique near-field propagation and scattering technique for automotive radar applications. In *2019 IEEE International Conference on Microwaves, Antennas, Communications and Electronic Systems (COMCAS)*. IEEE, 2019.

[104] Arien P Sligar. Machine learning-based radar perception for autonomous vehicles using full physics simulation. *IEEE Access*, 8:51470–51476, 2020.

[105] Elahe Soltanaghaei, Akarsh Prabhakara, Artur Balanuta, Matthew Anderson, Jan M Rabaey, Swarun Kumar, and Anthony Rowe. Millimetro: mmwave retro-reflective tags for accurate, long range localization. In *Proceedings of the 27th Annual International Conference on Mobile Computing and Networking*, pages 69–82, 2021.

[106] James J Spilker Jr. Gps signal structure and performance characteristics. *Navigation*, 25(2):121–146, 1978.

[107] MACOM Technologies. Masw-011105: Macom 17-31 ghz rf switch. https://cdn.macom.com/datasheets/MASW_011105.pdf.

[108] Jörn Thieling, Susanne Frese, and Jürgen Roßmann. Scalable and physical radar sensor simulation for interacting digital twins. *IEEE Sensors Journal*, 21(3):3184–3192, 2020.

[109] Yonglong Tian, Guang-He Lee, Hao He, Chen-Yu Hsu, and Dina Katabi. RF-based fall monitoring using convolutional neural networks. 2018.

[110] Lester Clare Van Atta. Electromagnetic reflector, October 6 1959. US Patent 2,908,002.

[111] Jessica Van Brummelen, Marie O'Brien, Dominique Gruyer, and Homayoun Najjaran. Autonomous vehicle perception: the technology of today and tomorrow. *Transportation research part C: emerging technologies*, 89:384–406, 2018.

[112] Tristan Visentin, Andrej Sagainov, Jürgen Hasch, and Thomas Zwick. Classification of objects in polarimetric radar images using CNNs at 77 GHz. *2017 IEEE Asia Pacific Microwave Conference (APMC)*, pages 356–359, 2017.

[113] Shelly Vishwakarma, Wenda Li, Chong Tang, Karl Woodbridge, Raviraj Adve, and Kevin Chetty. Simhumalator: An open-source end-to-end radar simulator for human activity recognition. *IEEE Aerospace and Electronic Systems Magazine*, 37(3):6–22, 2021.

[114] Jacquelyn A Vitaz. Enhanced discrimination techniques for radar-based on-metal identification tags. 2011.

[115] Jacquelyn A Vitaz, Amelia M Buerkle, and Kamal Sarabandi. Tracking of metallic objects using a retro-reflective array at 26 ghz. *IEEE transactions on antennas and propagation*, 58(11):3539–3544, 2010.

[116] Sourabh Vora, Alex H Lang, Bassam Helou, and Oscar Beijbom. Pointpainting: Sequential fusion for 3d object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4604–4612, 2020.

[117] Stefan O Wald and Frank Weinmann. Ray tracing for range-doppler simulation of 77 ghz automotive scenarios. In *2019 13th European Conference on Antennas and Propagation (EuCAP)*, pages 1–4. IEEE, 2019.

[118] LeiChen Wang, Simon Giebenhain, Carsten Anklam, and Bastian Goldluecke. Radar ghost target detection via multimodal transformers. *IEEE Robotics and Automation Letters*, 6(4):7758–7765, 2021.

[119] Leichen Wang, Bastian Goldluecke, and Carsten Anklam. L2r gan: Lidar-to-radar translation. In *Proceedings of the Asian Conference on Computer Vision*, 2020.

[120] Yan Wang, Wei-Lun Chao, Divyansh Garg, Bharath Hariharan, Mark Campbell, and Kilian Q Weinberger. Pseudo-lidar from visual depth estimation: Bridging the gap in 3d object detection for autonomous driving. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8445–8453, 2019.

[121] Zhixin Wang and Kui Jia. Frustum convnet: Sliding frustums to aggregate local point-wise features for amodal 3D object detection, 2019.

[122] Rob Weston, Oiwi Parker Jones, and Ingmar Posner. There and back again: Learning to simulate radar data for real-world applications. In *2021 IEEE International Conference on Robotics and Automation (ICRA)*, pages 12809–12816. IEEE, 2021.

[123] Svante Wold, Kim Esbensen, and Paul Geladi. Principal component analysis. *Chemometrics and intelligent laboratory systems*, 2(1-3):37–52, 1987.

[124] Jiong Xu, Ming Dong, Dazhi Ding, and Rushan Chen. Multi-radar data fusion imaging based on iterative adaptive algorithm. In *2017 International Applied Computational Electromagnetics Society Symposium (ACES)*, pages 1–2. IEEE, 2017.

[125] Ritu Yadav, Axel Vierling, and Karsten Berns. Radar+ rgb fusion for robust object detection in autonomous vehicle. In *2020 IEEE International Conference on Image Processing (ICIP)*, pages 1986–1990. IEEE, 2020.

[126] Bin Yang, Runsheng Guo, Ming Liang, Sergio Casas, and Raquel Urtasun. Radarnet: Exploiting radar for robust perception of dynamic objects. In *European Conference on Computer Vision*, pages 496–512. Springer, 2020.

[127] Shichao Yue, Hao He, Hao Wang, Hariharan Rahul, and Dina Katabi. Extracting multi-person respiration from entangled RF signals. 2018.

[128] Xiao Zhang, Wenda Xu, Chiyu Dong, and John M Dolan. Efficient l-shape fitting for vehicle detection using laser scanners. In *2017 IEEE Intelligent Vehicles Symposium (IV)*, pages 54–59. IEEE, 2017.

[129] Mingmin Zhao, Fadel Adib, and Dina Katabi. Emotion recognition using wireless signals. In *Proceedings of the 22nd Annual International Conference on Mobile Computing and Networking*, pages 95–108. ACM, 2016.

[130] Mingmin Zhao, Tianhong Li, Mohammad Abu Alsheikh, Yonglong Tian, Hang Zhao, Antonio Torralba, and Dina Katabi. Through-wall human pose estimation using radio signals. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7356–7365, 2018.

[131] Mingmin Zhao, Yonglong Tian, Hang Zhao, Mohammad Abu Alsheikh, Tianhong Li, Rumen Hristov, Zachary Kabelac, Dina Katabi, and Antonio Torralba. RF-based 3D skeletons. In *Proceedings of the 2018 Conference of the ACM Special Interest Group on Data Communication*, pages 267–281. ACM, 2018.

[132] Mingmin Zhao, Shichao Yue, Dina Katabi, Tommi S Jaakkola, and Matt T Bianchi. Learning sleep stages from radio signals: A conditional adversarial architecture. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 4100–4109. JMLR. org, 2017.

[133] Xingyi Zhou, Dequan Wang, and Philipp Krähenbühl. Objects as points. *arXiv preprint arXiv:1904.07850*, 2019.