Reconfigurable Intelligent Surfaces

for Next-Generation Communication and Sensing Systems

by

Abdelrahman Aly Hassan Anis Taha

A Dissertation Presented in Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy

Approved January 2023 by the
Graduate Supervisory Committee:

Ahmed Alkhateeb, Chair
Daniel Bliss
Cihan Tepedelenlioglu
Nicolò Michelusi

ARIZONA STATE UNIVERSITY

May 2023

ABSTRACT

With the rapid development of reflect-arrays and software-defined meta-surfaces, reconfigurable intelligent surfaces (RISs) have been envisioned as promising technologies for next-generation wireless communication and sensing systems. These surfaces comprise massive numbers of nearly-passive elements that interact with the incident signals in a smart way to improve the performance of such systems. In RIS-aided communication systems, designing this smart interaction, however, requires acquiring large-dimensional channel knowledge between the RIS and the transmitter/receiver. Acquiring this knowledge is one of the most crucial challenges in RISs as it is associated with large computational and hardware complexity. For RIS-aided sensing systems, it is interesting to first investigate scene depth perception based on millimeter wave (mmWave) multiple-input multiple-output (MIMO) sensing. While mmWave MIMO sensing systems address some critical limitations suffered by optical sensors, realizing these systems possess several key challenges: communication-constrained sensing framework design, beam codebook design, and scene depth estimation challenges. Given the high spatial resolution provided by the RISs, RIS-aided mmWave sensing systems have the potential to improve the scene depth perception, while imposing some key challenges too. In this dissertation, for RIS-aided communication systems, efficient RIS interaction design solutions are proposed by leveraging tools from compressive sensing and deep learning. The achievable rates of these solutions approach the upper bound, which assumes perfect channel knowledge, with negligible training overhead. For RIS-aided sensing systems, a mmWave MIMO based sensing framework is first developed for building accurate depth maps under the constraints imposed by the communication transceivers. Then, a scene depth estimation framework based on RIS-aided sensing is developed for building high-resolution accurate depth maps. Numerical simulations illustrate the promising performance of the proposed solutions, highlighting their potential for next-generation communication and sensing systems.

To my family.

ACKNOWLEDGMENTS

First, I would like to express my sincere gratitude to my supervisor, Prof. Ahmed Alkhateeb for his patience, motivation and technical guidance throughout this work. I would also like to thank my supervisory committee members, Prof. Prof. Daniel Bliss, Prof. Cihan Tepedelenlioglu, and Prof. Nicolò Michelusi for their time and valuable feedback. Finally and most importantly, my family deserves a special mention. I would like to thank my parents for their continuous support. I am eternally grateful.

TABLE OF CONTENTS

CHAPTER Page

CHAPTER Page

CHAPTER Page

LIST OF TABLES

LIST OF FIGURES

Figure Page

Figure Page

Chapter 1

INTRODUCTION

Wireless communications can be arguably considered as one of the main technological revolutions. With the massive number of devices that are wirelessly connected at all times, it is hard not to perceive the impact of wireless communications on the contemporary society. In the recent period, reconfigurable intelligent surfaces (RISs) have been envisioned as integral technologies for next-generation wireless communication and sensing systems. From a conceptual design perspective, by stacking a huge number of sensing or radiating elements, the RIS ideally aims to effectuate a continuous electromagnetically active surface. As depicted in Fig. 1.1, these RIS elements are expected to interact in a smart way with the incident signals in order to enhance the spectral efficiency and coverage of wireless systems [10, 11]. These surfaces could also be developed with energy-efficient implementations, e.g., using nearly-passive elements with reconfigurable parameters [12–14], which deems them more promising for next-generation wireless systems.

Similar to wireless communications, radar systems are deployed worldwide, with a variety of applications including air traffic control, geophysical monitoring, weather observation as well as surveillance for defense and security. Recently, a special focus has been assigned on radar systems operating in the millimeter-wave (mmWave) frequency band to provide high-accuracy environment details in a short range. mmWave radar systems that are used in short-range applications are commonly referred to as mmWave sensing systems. mmWave sensing is widely adopted in various mobility and imaging applications, e.g. autonomous vehicle applications. The goal of mmWave sensing is to acquire information about the surrounding environment using mmWave

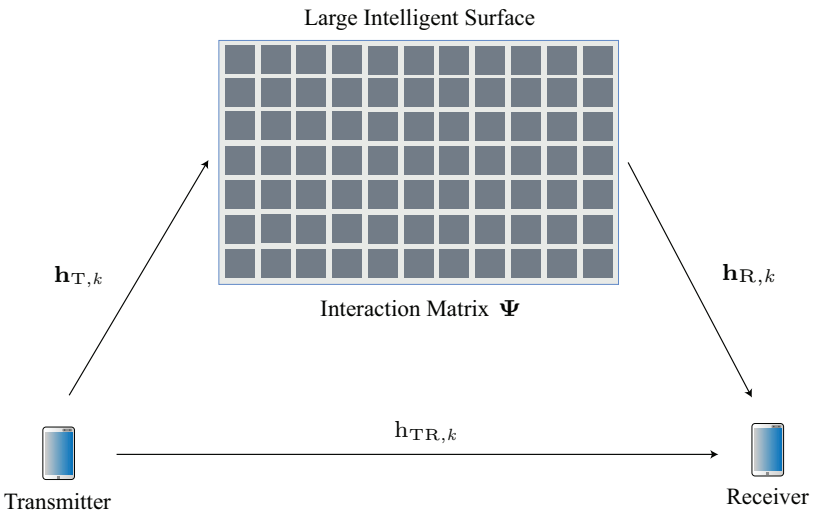


Figure 1.1: The Transmitter-Receiver Communication Is Assisted by a Reconfigurable Intelligent Surface (RIS). The RIS Is Interacting with the Incident Signal Through an Interaction Matrix $\mathbf{\Psi}$.

radar sensors.

Conventional radar and wireless communication systems operate separately at predefined frequency bands to avoid interference. Given the rapid growth of connected devices and services, the frequency spectrum is becoming increasingly congested, with almost all wireless services having a need for a greater access to it. As a result, spectrum authority entities are seeking opportunities to reuse some bands of the frequency spectrum currently restricted to other applications. The radar bands are among the best candidate to be shared with various communication systems. For these reasons, joint sensing and communication systems, sharing the same spectral band, have captured a great deal of attention in recent years.

Augmented and virtual reality (AR/VR) systems are rapidly becoming key components of the wireless landscape. Enabling immersive wireless AR/VR experience, however, requires high resolution and accurate scene depth perception. This can potentially allow the wireless AR/VR users to move freely within their indoor or outdoor environment. Current scene depth perception approaches rely mainly on

optical sensing for constructing the depth maps. Previous depth map construction approaches focused on leveraging: (i) monocular images using RGB cameras [8], (ii) passive/active stereo images using either RGB-D depth cameras [15, 16] or infrared (IR) stereo cameras [17, 18], and (iii) gated images using active gated imaging cameras [19, 20]. The performance of these depth cameras, however, has clear limitations in several scenarios in AR/VR applications, which motivates the need for RIS-aided wireless sensing systems in constructing accurate scene depth perception.

Enabling the aforementioned systems in practice suffers from some critical challenges. In the upcoming sections, we discuss the key research challenges in (a) RIS aided wireless communication systems and (b) RIS aided wireless sensing systems for scene depth estimation. We then introduce our contributions to addressing these research challenges.

1.1 RIS Aided Communication Systems

In reconfigurable intelligent surfaces aided communication systems, prior work focused on designing the RIS interaction matrices and evaluating their spectral efficiencies and coverage gains while assuming the availability of the global channel knowledge. For example, all the prior work in [13, 21–23] assumed that the knowledge about the channels between the RIS and the transmitters/receivers is available at the base station, either perfectly or with some error. Obtaining this channel knowledge, however, is one of the most crucial challenges for RIS systems because of the massive number of RIS elements and the hardware constraints on these elements. More specifically, if all RIS elements are passive, channel estimation or beam training solutions yields huge-and possible prohibitive, training overhead, as the number of pilots (or codebook beams) in this case will be in the order of the number of RIS elements. To reduce this training overhead, the RIS is required to employ a complex

hardware architecture that connects all the antenna elements to a baseband processing unit either through a fully-digital or hybrid analog/digital architectures [24, 25]. This approach, though, can lead to high hardware complexity and power consumption given the massive number of RIS elements.

This challenge motivates the need to develop an interaction design framework for the RIS systems with no prior challenge knowledge. This interaction design framework needs to approach the upper bound on the achievable rate — which assumes perfect channel knowledge — with low-training overhead and with energy-efficient hardware.

1.2 RIS Aided Sensing Systems for Scene Depth Estimation

Current scene depth estimation approaches rely heavily on using optical sensing systems. While optical sensors can generally provide good accuracy, they suffer from critical limitations. These limitations stem from the fundamental properties of the way visible light propagates and interacts with the elements of an environment. Adopting the depth map construction approaches in [8, 15–20, 26] have the following important complications. (i) First, the accuracy of optical sensors normally degrades in scenarios with unfavorable light conditions, in the presence of shiny, dark, or transparent objects/surfaces, and in the presence of non-line-of-sight (NLoS) objects/surfaces. While there are some attempts to solve some of these challenges using IR stereo cameras [17] or excessive processing of the RGB-D images [27], there is no complete and general solution yet to this problem. (ii) Second, optical sensors suffer from key privacy concerns and depth/velocity estimation ambiguity for distant objects/surfaces. For example, the depths for distant surfaces can not be resolved by the algorithms in [17, 27]. (iii) The field of view coverage is also a key challenge. The depth map coverage is limited by the camera field of view. The camera field of view is constrained by the camera lens and by the light sensor. The field of view in wire-

less sensing systems, however, can be constrained by the antenna radiation pattern. By comparison, the typical field of view in mmWave wireless sensing systems can be larger than the typical optical camera field of view. (iv) In addition, for AR/VR systems, another key challenge is the additional bill of materials (BOM) cost incurred from integrating the IR stereo camera systems into the wireless AR/VR device architectures. By contrast, the existing mmWave systems in the wireless AR/VR device architectures incur no additional BOM cost when leveraged for depth map estimation purposes jointly with the primary purpose of wireless communications. These challenges motivate the research for other technologies to complement or replace the optical sensors in accurately estimating the scene depth of the surrounding environment.

1.3 Overview of Contributions

The key challenges discussed in Sections 1.1 and 1.2 need to be addressed in this dissertation. With this motivation, the research problems addressed in this dissertation serves as a good start to resolve these challenges. The primary contributions of this dissertation can be summarized as follows.

1. We propose efficient RIS interaction design approaches for RIS-aided wireless communication systems with negligible training overhead. First, we propose a new RIS architecture, where all the elements are passive except for a few randomly distributed active channel sensors. Only those few active sensors are connected to the baseband of the RIS controller to enable the efficient design of the RIS interaction matrices. Using this new architecture, we develop three solutions that design the RIS interaction matrices: (a) compressive sensing based solution, (b) supervised deep learning based solution, and (c) deep reinforcement learning based solution. Simulation results show that the developed

solutions can all approach the optimal upper bound, which assumes perfect channel knowledge, when only a few RIS elements are active and with almost no training overhead [28–30].

2. We propose a mmWave MIMO based scene depth estimation framework for wireless AR/VR systems, under the constraints imposed by mmWave communication hardware and frame structure. We define the characteristics of the desirable mmWave sensing beamforming codebook for efficient depth map construction and develop a depth-map suitable sensing beamforming codebook that meets these characteristics. Given the designed beamforming codebook, we propose a signal processing approach for jointly processing the signals received by the sensing beams and building accurate depth maps. Simulation results show the promise of mmWave MIMO sensing in becoming a viable depth estimation solution for communication-constrained sensing systems, either as a standalone approach or as an integrated approach with RGB-D depth cameras [31]. This contribution point represents an important step towards developing RIS-aided wireless sensing systems for scene depth estimation.

3. We propose a sensing framework for scene depth estimation using RIS aided wireless sensing systems. This framework comprises two key elements, namely the RIS interaction codebook design and the scene depth estimation solution. We propose a novel RIS interaction codebook design capable of creating a sensing grid of reflected beams that meets the desirable characteristics of efficient scene depth map construction. Given the designed RIS interaction codebook, we develop a post-processing solution on the receive signals to build high-resolution accurate depth maps. Simulation results highlight the potential of leveraging RIS aided mmWave sensing in achieving accurate depth perception of the sur-

rounding environment.

1.4 Notations

We use the following notation throughout this dissertation: $\mathbf{A}$ is a matrix, $\mathbf{a}$ is a vector, a is a scalar, $\mathcal{A}$ is a set of scalars, and $\boldsymbol{\mathcal{A}}$ is a set of vectors. $\|\mathbf{a}\|_p$ is the p-norm of $\mathbf{a}$. $|\mathbf{A}|$ is the determinant of $\mathbf{A}$, $\|\mathbf{A}\|_F$ is its Frobenius norm, whereas $\mathbf{A}^T$, $\mathbf{A}^H$, $\mathbf{A}^*$, $\mathbf{A}^{-1}$, $\mathbf{A}^\dagger$ are its transpose, Hermitian (conjugate transpose), conjugate, inverse, and pseudo-inverse respectively. $[\mathbf{A}]_{r,c}$ is the element in the r^{th} row and c^{th} column of the matrix $\mathbf{A}$. $[\mathbf{A}]_{r,:}$ and $[\mathbf{A}]_{:,c}$ are the r^{th} row and c^{th} column of the matrix $\mathbf{A}$ respectively. $[\mathbf{a}]_k$ is the k^{th} element of the vector $\mathbf{a}$. $\text{diag}(\mathbf{a})$ is a diagonal matrix with the entries of $\mathbf{a}$ on its diagonal. $\mathbf{I}$ is the identity matrix. $\mathbf{1}_N$ and $\mathbf{0}_N$ are the N-dimensional all-ones and all-zeros vector, respectively. $\mathbf{A} \otimes \mathbf{B}$ is the Kronecker product of $\mathbf{A}$ and $\mathbf{B}$, $\mathbf{A} \circ \mathbf{B}$ is their Khatri-Rao product, and $\mathbf{A} \odot \mathbf{B}$ is their Hadamard product. $\text{vec}(\mathbf{A})$ is a vector whose elements are the stacked columns of matrix $\mathbf{A}$. $\mathcal{N}(\mathbf{m}, \mathbf{R})$ is a complex-valued Gaussian random vector with mean $\mathbf{m}$ and covariance $\mathbf{R}$. $|\boldsymbol{\mathcal{A}}|$ is the cardinality of the set $\boldsymbol{\mathcal{A}}$. $\mathbb{E}[\cdot]$ is used to denote expectation. $\text{Re}(z)$, $\text{Im}(z)$, and $\arg(z)$ are the real part, the imaginary part, and the phase angle of the complex number z. $f(t) * g(t)$ is the continuous-time convolution of two signals $f(t)$ and $g(t)$. $\text{FFT}_m(\cdot)$ is the 1D FFT operation on the input matrix along its column dimension of index m.

1.5 Organization

The rest of the dissertation is organized as follows. In Chapter 2, we first consider the challenge of adopting reconfigurable intelligent surfaces to assist the wireless communication systems with no prior channel knowledge, and we propose efficient refelection beamforming design approaches for such challenge. Then, we consider wireless AR/VR systems in Chapter 3 and we propose a communication-constrained

mmWave MIMO based wireless sensing framework for scene depth estimation. In Chapter 4, we then investigate RIS aided wireless sensing systems for scene depth estimation. Last, concluding remarks and future work are presented in Chapter 5.

Chapter 2

INTERACTION DESIGN FOR RECONFIGURABLE INTELLIGENT SURFACES

2.1 Abstract

Employing Reconfigurable Intelligent Surfaces (RISs) is a promising solution for improving the coverage and rate of future wireless systems. These surfaces comprise massive numbers of nearly-passive elements that interact with the incident signals, for example by reflecting them, in a smart way that improves the wireless system performance. Prior work focused on the design of the RIS reflection matrices assuming full channel knowledge. Estimating these channels at the RIS, however, is a key challenging problem. With the massive number of RIS elements, channel estimation or reflection beam training will be associated with (i) huge training overhead if all the RIS elements are passive (not connected to a baseband) or with (ii) prohibitive hardware complexity and power consumption if all the elements are connected to the baseband through a fully-digital or hybrid analog/digital architecture. This chapter [1] proposes efficient solutions for these problems by leveraging tools from compressive sensing and deep learning. First, a novel RIS architecture based on *sparse channel sensors* is proposed. In this architecture, all the RIS elements are passive except for a few elements that are active (connected to the baseband). We then develop three solutions that design the RIS reflection matrices with negligible training overhead. In the first approach, we leverage compressive sensing tools to construct the channels

[1] This chapter is based on the work published in the journal paper: A. Taha, M. Alrabeiah and A. Alkhateeb, "Enabling Large Intelligent Surfaces With Compressive Sensing and Deep Learning," in IEEE Access, vol. 9, pp. 44304-44321, 2021. This work was supervised by Prof. Ahmed Alkhateeb. Dr. Muhammad Alrabeiah provided important ideas for the large intelligent surface aided system design that greatly improved the work.

at all the RIS elements from the channels seen only at the active elements. In the second approach, we develop a deep-learning based solution where the RIS learns how to interact with the incident signal given the channels at the active elements, which represent the state of the environment and transmitter/receiver locations. We show that the achievable rates of the proposed solutions approach the upper bound, which assumes perfect channel knowledge, with negligible training overhead and with only a few active elements, making them promising for future RIS systems.

2.2 Introduction

reconfigurable intelligent surfaces (RISs) have been envisioned as integral constituents of beyond-5G wireless systems [10–14, 21–23, 32–39]. From a conceptual design perspective, by stacking a huge number of sensing or radiating elements, the RIS ideally aims to effectuate a continuous electromagnetically active surface. These RIS elements are expected to interact in a smart way with the incident signals in order to enhance the spectral efficiency and coverage of wireless systems [10, 11]. What adds to the appeal of such surfaces is that their function could be performed with energy-efficient implementations, e.g., using nearly-passive elements such as analog phase shifters [12–14]. Prior work focused on designing the RIS interaction matrices and evaluating their spectral efficiencies and coverage gains *while assuming the availability of global channel knowledge*. **But how can these extremely large-dimensional channels be estimated if the RIS is implemented using only reflecting elements?** Obtaining this channel knowledge may require huge—and possibly prohibitive—training overhead, which represents the main challenge for the RIS system operation. To overcome that, this work proposes a novel RIS hardware architecture along with three solutions based on compressive sensing and deep learning. These solutions utilize the novel architecture of the surface and design the interaction matrix with very negligible

training overhead.

2.2.1 Prior Work

Under various names such as large intelligent surfaces, intelligent reflecting surfaces, and smart reflect-arrays, RIS-assisted wireless communications have been drawing increasing interest in recent years. From an implementation perspective, RIS can be built using nearly-passive elements with reconfigurable parameters [13]. Various RIS designs have been proposed in the literature with more prominence given to software-defined metamaterials [32, 33] and conventional reflect-arrays [12, 14] among others. For all those designs, different signal processing solutions have been proposed for optimizing the design of the RIS interaction matrices. An RIS-assisted downlink multiuser setup was considered in [13] with single-antenna users. computational low-complexity algorithms were then proposed for optimizing the design of the RIS interaction matrices, using quantized phase shifters/reflectors for modeling the RIS elements. In [21], an RIS-assisted downlink scenario was considered, where both the RIS interaction matrix and the base station precoder matrix were designed, assuming the case where a line-of-sight (LOS) may exist between the base station and the RIS. In [22], a new transmission strategy combining RIS with index modulation was proposed to improve the system spectral efficiency.

In terms of the overall system performance, an uplink multiuser scenario was considered in [34] and the data rates were formulated for the case where channel estimation errors exist in the available channel knowledge. A downlink RIS-assisted multiple-input multiple-output (MIMO) non-orthogonal multiple access (NOMA) framework is proposed in [23] for achieving higher system spectrum efficiency gains. The RIS can be leveraged for wireless localization purposes as well; in [39], an RIS-assisted downlink millimeter wave (mmWave) positioning problem was analyzed from

the Fisher Information perspective. Based on this analysis, an algorithm was developed for improving the positioning quality.

Deep learning solutions have been proposed in the literature for addressing design challenges in mmWave and massive MIMO systems [40–42]. In [40], a deep learning based beam prediction solution was proposed for distributed mmWave MIMO systems to serve highly mobile users with negligible training overhead and high data rate gains, compared to coordinated beamforming strategies that do not leverage machine learning. In [41], a deep learning based blockage prediction solution was proposed to address the reliability and latency challenges of sudden blockage of the line-of-sight link in mmWave MIMO systems. A channel covariance prediction solution using generative adversarial networks was proposed in [42] for mmWave Massive MIMO systems to reduce the training overhead associated with acquiring the channel knowledge.

The Critical Challenge: All the prior work in [13, 14, 21–23, 34] assumed that the knowledge about the channels between the RIS and the transmitters/receivers is available at the base station, either perfectly or with some error. Obtaining this channel knowledge, however, is one of the most crucial challenges for RIS systems because of the massive number of antennas (RIS elements) and the hardware constraints on these elements. More specifically, if the RIS elements are implemented using phase shifters that just reflect the incident signals, then there are two main approaches for designing the RIS reflection matrix. The first approach is to estimate the RIS-assisted channels at the transmitter/receiver by training all the RIS elements, normally one by one, and then use the estimated channels to design the reflection matrix. This yields a massive channel training overhead because of the very large number of elements at the RIS. Instead of the explicit channel estimation, the RIS reflection matrix can be selected from quantized codebooks via online beam/reflection training. This is similar

to the common beam training techniques in mmWave systems that employ similar phase shifter architectures [43, 44]. To sufficiently quantize the space, however, the size of the reflection codebooks needs normally to be in the order of the number of antennas, which leads to huge training overhead. To avoid this training overhead, a trivial solution is to employ fully-digital or hybrid analog/digital architectures at the RIS, where every antenna element is connected somehow to the baseband where channel estimation strategies can be used to obtain the channels [24, 25, 45]. This solution, however, leads to high hardware complexity and power consumption because of the massive number of RIS elements.

2.2.2 Contribution

In this chapter, we consider an RIS-assisted wireless communication system and propose a novel RIS architecture as well as compressive sensing and deep learning based solutions that design the RIS reflection matrix with negligible training overhead. More specifically, the contributions of this chapter can be summarized as follows.

- *Novel RIS hardware architecture:* We introduce a new RIS architecture where all the elements are passive except a few randomly distributed active channel sensors. Only those few active sensors are connected to the baseband of the RIS controller and are used to enable the efficient design of the RIS reflection matrices with low training overhead.

- *Compressive sensing based RIS reflection matrix design:* Given the new RIS architecture with randomly distributed active elements, we develop a compressive sensing based solution to recover the full channels between the RIS and the transmitters/receivers from the *sampled* channels sensed at the few active elements. Using the constructed channels, we then design the RIS reflection

matrices with no training overhead. We show that the proposed solution can efficiently design the RIS reflection matrices when only a small fraction of the RIS elements are active, yielding a promising solution for RIS systems from both energy efficiency and training overhead perspectives.

- *Deep learning based RIS reflection matrix design:* By leveraging deep learning tools, we propose three solutions that learns the direct mapping from the sampled channels seen at the active RIS elements and the optimal RIS reflection matrices that maximize the system achievable rate. Essentially, the proposed approaches teach the RIS system how to interact with the incident signal given the knowledge of the sampled channel vectors, that we call *environment descriptors*. The RIS learns that when it observes these environment descriptors, it should reflect the incident signal using this reflection matrix. Different from the compressive sensing solution, the deep learning approaches leverage the prior observations at the RIS and does not require any knowledge of the array structure. It is worth mentioning that a conference version of this work is presented in [29].

- *A novel deep reinforcement learning (DRL) based solution* is proposed for predicting the best RIS interaction, where the RIS learns how to reflect the incident signals in the best possible way by adjusting its reflection matrix. This solution eliminates the need for collecting large training dataset, hence requires almost no beam training overhead. The proposed framework is directed more towards *standalone RIS operation*, where the RIS architecture is not controlled/assisted by any base station, but rather operating on its own while interacting with the environment, and without any initial training phase requirement. A conference version of this work is presented in [30].

The proposed solutions are extensively evaluated using the accurate ray-tracing based DeepMIMO dataset [2]. The results show that the developed compressive sensing and deep learning solutions can all approach the optimal upper bound, which assumes perfect channel knowledge, when only a few RIS elements are active and with almost no training overhead.

The rest of the chapter is organized as follows. Section 2.3 presents the system and channel models adopted. Section 2.4 presents the formal description of the main problem — the design of the RIS interaction matrix. Section 2.5 proposes and discusses the novel sparse RIS architecture. Sections 2.6, 2.7, and 2.8 present, respectively, the proposed compressive sensing, supervised deep learning, and deep reinforcement learning solutions to the problem of designing the interaction matrix. Section 2.9 puts the proposed architecture and solutions to test by investigating the performance of each solution and the effect of various design parameters. Finally, Section 2.10 concludes this chapter with a summary of the findings and a few concluding remarks.

2.3 System and Channel Models

The adopted system and channel models for reconfigurable intelligent surfaces (RISs) are described in this section.

2.3.1 System Model

Consider a communication system where a transmitter is communicating with a receiver, and this communication is aided by a reconfigurable intelligent surface (RIS), as depicted in Fig. 1.1. These transmitters/receivers can represent either base stations or user equipment. As shown in Fig. 1.1, the RIS is interacting with the incident signal through an interaction matrix $\mathbf{\Psi}$. Let the RIS be equipped with M reconfigurable elements and assume that both the transmitter and receiver have a

single-antenna. It is worth noting here that such an assumption is only adopted for simplicity of exposition and the proposed solutions and the results in this chapter can be readily extended to multi-antenna transceivers. To put that description in formal terms, we adopt an OFDM-based system with K subcarriers. We define $h_{\mathrm{TR},k} \in \mathbb{C}$ as the direct channel between the transmitter and receiver at the k^{th} subcarrier, $\mathbf{h}_{\mathrm{T},k}, \mathbf{h}_{\mathrm{R},k} \in \mathbb{C}^{M\times 1}$ as the $M \times 1$ uplink channels from the transmitter and receiver to the RIS at the k^{th} subcarrier, and by reciprocity, $\mathbf{h}_{\mathrm{T},k}^T, \mathbf{h}_{\mathrm{R},k}^T$ as the downlink channels. The received signal at the receiver side could be expressed as

$$y_k = \underbrace{\mathbf{h}_{\mathrm{R},k}^T \boldsymbol{\Psi}_k \mathbf{h}_{\mathrm{T},k} s_k}_{\text{RIS-assisted link}} + \underbrace{h_{\mathrm{TR},k} s_k}_{\text{Direct link}} + n_k, \tag{2.1}$$

where the matrix $\boldsymbol{\Psi}_k \in \mathbb{C}^{M\times M}$, that we call the RIS interaction matrix, characterizes the *interaction* of the RIS with the incident (impinging) signal from the transmitter. s_k represents the transmitted signal over the k^{th} subcarrier, and satisfies the per-subcarrier power constraint $\mathbb{E}[|s_k|^2] = \frac{P_{\mathrm{T}}}{K}$, with P_{T} being the total transmit power. The receive noise is denoted by $n_k \sim \mathcal{N}_{\mathbb{C}}(0, \sigma_n^2)$.

The overall objective of the RIS is then to interact with the incident signal (via adjusting $\boldsymbol{\Psi}_k$) in a way that optimizes a certain performance metric such as the system achievable rate or the network coverage. To simplify the design and analysis of the algorithms in this work, we will focus on the case where the direct link does not exist. This represents the scenarios where the direct link is either blocked or has negligible receive power compared to that received through the RIS-assisted link. With this assumption, the receive signal can be expressed as

$$y_k = \mathbf{h}_{\mathrm{R},k}^T \boldsymbol{\Psi}_k \mathbf{h}_{\mathrm{T},k} s_k + n_k, \tag{2.2}$$

$$\overset{(a)}{=} \left(\mathbf{h}_{\mathrm{R},k} \odot \mathbf{h}_{\mathrm{T},k}\right)^T \boldsymbol{\psi}_k s_k + n_k, \tag{2.3}$$

where (a) follows from the diagonal structure of the interaction matrix $\boldsymbol{\Psi}_k$, whose diagonal entries could be stacked in a vector $\boldsymbol{\psi}_k \in \mathbb{C}^{M\times 1}$ such that $\boldsymbol{\Psi}_k = \mathrm{diag}\left(\boldsymbol{\psi}_k\right)$.

This diagonal structure results from the RIS operation where every element $m, m \in \{1, 2, \ldots, M\}$, reflects only its incident signal after multiplying it with an interaction factor $[\boldsymbol{\psi}_k]_m$. Now, we make two important notes on these interaction vectors. First, while the interaction factors, $[\boldsymbol{\psi}_k]_m, \forall m, k$, can generally have different magnitudes (amplifying/attenuation gains), it is more practical to assume that the RIS elements are implemented using only phase shifters. Second, since the implementation of the phase shifters is done in the analog domain (using RF circuits), the same phase shift will be applied to the signals on all subcarriers, i.e., $\boldsymbol{\psi}_k = \boldsymbol{\psi}, \forall k$. Accounting for these practical considerations, we assume that every interaction factor is just a phase shifter, i.e., $[\boldsymbol{\psi}]_m = e^{j\phi_m}$. Further, we will call the interaction vector $\boldsymbol{\psi}$ in this case the *reflection beamforming* vector.

2.3.2 Channel Model

In this work, we adopt a wideband geometric channel model for the channels $\mathbf{h}_{\mathrm{T},k}, \mathbf{h}_{\mathrm{R},k}$ between the transmitter/receiver and the RIS [29, 30, 40]. Consider an uplink transmitter-RIS channel, $\mathbf{h}_{\mathrm{T},k} \in \mathbb{C}^{M\times 1}$, consisting of L clusters, each of which (i.e., ℓ^{th} cluster) contributes a single ray with a time delay $\tau_\ell \in \mathbb{R}$; azimuth/elevation angles of arrival, $\phi_\ell \in [0, 2\pi), \theta_\ell \in [0, \pi)$; an uplink path loss ρ_{T}; and a complex coefficient $\alpha_\ell \in \mathbb{C}$. Let $p(\tau)$ denotes the pulse shaping function for T_S-spaced signaling evaluated at τ seconds. Let the array response vector of the RIS at the angles of arrival, ϕ_ℓ, θ_ℓ, be defined as $\mathbf{a}(\phi_\ell, \theta_\ell) \in \mathbb{C}^{M\times 1}$. The delay-$d$ channel vector, $\boldsymbol{h}_{\mathrm{T},d} \in \mathbb{C}^{M\times 1}$, between the transmitter and the RIS can then be formulated as

$$\boldsymbol{h}_{\mathrm{T},d} = \sqrt{\frac{M}{\rho_{\mathrm{T}}}} \sum_{\ell=1}^{L} \alpha_\ell \; p(dT_S - \tau_\ell) \; \mathbf{a}\left(\theta_\ell, \phi_\ell\right), \tag{2.4}$$

Given this delay-d channel, the channel vector at subcarrier k, $\mathbf{h}_{\mathrm{T},k}$, can be defined

in the frequency domain as

$$\mathbf{h}_{\mathrm{T},k} = \sum_{d=0}^{D-1} \boldsymbol{h}_{\mathrm{T},d}\; e^{-j\frac{2\pi k}{K}d}. \tag{2.5}$$

where D is the channel tap length. The downlink RIS-receiver channel $\mathbf{h}_{\mathrm{R},k}$ can be defined similarly. The channel vectors, $\{\mathbf{h}_{\mathrm{T},k}\}_{k=1}^{K}$ and $\{\mathbf{h}_{\mathrm{R},k}\}_{k=1}^{K}$, are assumed constant within the period of one coherence time, T_C, which mainly depends on the dynamics of the environment and the user mobility. It is worth noting that the number of channel paths L depends highly on the operational frequency band and the propagation environment. For example, mmWave channels normally consist of a few channel paths, ∼3-5 paths, [46–48], while sub-6 GHz signal propagation generally experiences rich scattering resulting in channels with more multi-path components.

2.4 Problem Formulation

Given the system and channel models in Section 2.3, our objective is to design the RIS interaction vector (reflection beamforming vector), $\boldsymbol{\psi} \in \mathbb{C}^{M\times 1}$, in order to maximize the achievable rate at the receiver, which can be formulated as

$$R = \frac{1}{K}\sum_{k=1}^{K} \log_2\left(1 + \mathsf{SNR}\left|\mathbf{h}_{\mathrm{R},k}^{T}\boldsymbol{\Psi}\mathbf{h}_{\mathrm{T},k}\right|^2\right), \tag{2.6}$$

$$= \frac{1}{K}\sum_{k=1}^{K} \log_2\left(1 + \mathsf{SNR}\left|\left(\mathbf{h}_{\mathrm{T},k}\odot\mathbf{h}_{\mathrm{R},k}\right)^{T}\boldsymbol{\psi}\right|^2\right), \tag{2.7}$$

where $\mathsf{SNR} = P_{\mathrm{T}}/(K\sigma_n^2)$ represents the signal-to-noise ratio. As mentioned in Section 2.3.1, every element in the RIS reflection beamforming vector, $\boldsymbol{\psi}$, is implemented using an RF phase shifter. These phase shifters, however, normally have a quantized set of angles and can not shift the signal with any phase. To capture this constraint, we assume that the reflection beamforming vector $\boldsymbol{\psi}$ can only be picked from a pre-defined codebook $\boldsymbol{\mathcal{P}}$. Every candidate reflection beamforming codeword in $\boldsymbol{\mathcal{P}}$ is assumed to be implemented using quantized phase shifters. With this assumption,

our objective is then to find the optimal reflection beamforming vector $\boldsymbol{\psi}^\star$ that solves

$$\boldsymbol{\psi}^\star = \arg\max_{\boldsymbol{\psi}\in\boldsymbol{\mathcal{P}}} \sum_{k=1}^{K} \log_2\left(1+\mathsf{SNR}\left|\left(\mathbf{h}_{\mathrm{T},k}\odot\mathbf{h}_{\mathrm{R},k}\right)^T\boldsymbol{\psi}\right|^2\right), \tag{2.8}$$

to result in the optimal rate $R^\star$ defined as

$$R^\star = \max_{\boldsymbol{\psi}\in\boldsymbol{\mathcal{P}}} \frac{1}{K}\sum_{k=1}^{K} \log_2\left(1+\mathsf{SNR}\left|\left(\mathbf{h}_{\mathrm{T},k}\odot\mathbf{h}_{\mathrm{R},k}\right)^T\boldsymbol{\psi}\right|^2\right). \tag{2.9}$$

The optimization problem in (2.8), unfortunately, has no close-form solution. This is a consequence of (a) the time-domain implementation of the reflection beamforming vector, i.e., using only one vector $\boldsymbol{\psi}$ for all subcarriers, and (b) the quantized codebook constraint.

The main challenge: As characterized in (2.8), finding the optimal RIS interaction vector $\boldsymbol{\psi}^\star$ and achieving the optimal rate $R^\star$ requires an exhaustive search over the codebook $\boldsymbol{\mathcal{P}}$. Note that the codebook size should normally be in the same order as the number of antennas to make use of these antennas. This means that a reasonable reflection beamforming codebook for RIS systems will probably have thousands of candidate codewords. With such huge codebooks, solving the exhaustive search in (2.8) is very challenging. More specifically, there are two main approaches for performing the search in (2.8).

- **Full channel estimation with offline exhaustive search:** In this approach, we need to estimate the full channels between the RIS and the transmitter/receiver, $\mathbf{h}_{\mathrm{T},k}, \mathbf{h}_{\mathrm{R},k}$ and use it to find the best reflection beamforming vector by the offline calculation of (2.8). Estimating these channel vectors, however, requires the RIS to employ a complex hardware architecture that connects all the antenna elements to a baseband processing unit either through a fully-digital or hybrid analog/digital architectures [24, 25]. Given the massive numbers of antennas at reconfigurable intelligent surfaces, this approach can yield **prohibitive hardware**

complexity in terms of routing and power consumption among others. If the RIS is operated and controlled via a base station or an access point [13], then this channel estimation process can be done at these communication ends. This, however, assumes an orthogonal training over the RIS antennas, for example by activating one RIS antenna at a time, which leads to **prohibitive training overhead** given the number of antennas at the RIS.

- **Online exhaustive beam training:** Instead of the explicit channel estimation, the best RIS beam reflection vector $\boldsymbol{\psi}^\star$ can be found through an over-the-air beam training process. This process essentially solves the exhaustive search in (2.8) by testing the candidate interaction vectors $\boldsymbol{\psi} \in \boldsymbol{\mathcal{P}}$ one by one. This exhaustive beam training process, however, incurs again **very large training overhead** at the RIS systems.

Our objective in this chapter is to enable reconfigurable intelligent surfaces by addressing this main challenge. More specifically, our objective is to enable RIS systems to approach the optimal achievable rate in (2.9) by adopting **low-complexity hardware architectures** and requiring **low training overhead**. For this objective, we first propose a novel energy-efficient RIS transceiver architecture in Section 2.5. Then, we show in Sections 2.6-2.7 how to employ this RIS architecture to achieve near-optimal achievable rates with negligible training overhead via leveraging tools from compressive sensing and deep learning.

2.5 Reconfigurable Intelligent Surfaces with Sparse Sensors: A Novel Architecture

As discussed in Section 2.4, a main challenge for the RIS system operation lies in the high hardware complexity and training overhead associated with designing the RIS interaction (reflection beamforming) vector, $\boldsymbol{\psi}$. To overcome this challenge

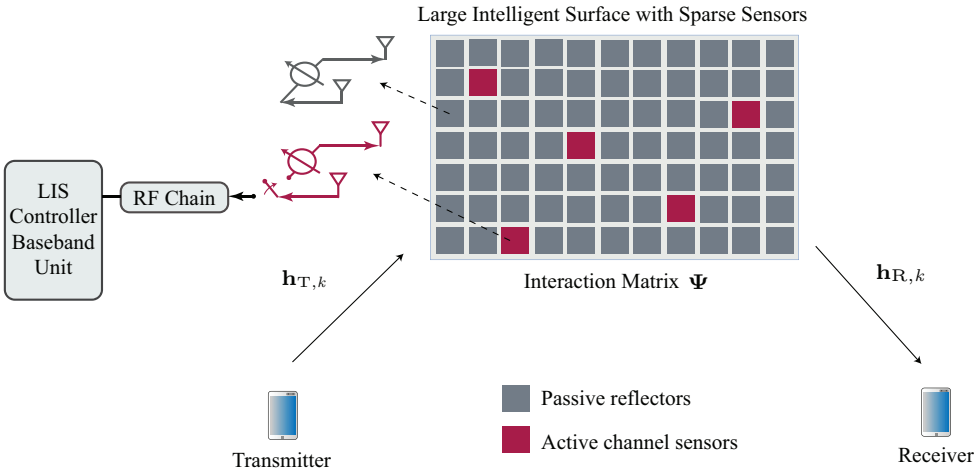


Figure 2.1: This Figure Illustrates the Proposed RIS Architecture Where $\overline{M}$ Active Channel Sensors Are Randomly Distributed Over the RIS. These Active Elements Have Two Modes of Operation (I) a Channel Sensing Mode Where It Is Connected to the Baseband and Is Used to Estimate the Channels and (Ii) a Reflection Mode Where It Just Reflects the Incident Signal by Applying a Phase Shift. The Rest of the RIS Elements Are Passive Reflectors and Are Not Connected to the Baseband.

and enable RIS systems in practice, we adopt a novel RIS architecture that relies on sparsely embedded active sensors. To further illustrate this architecture, consider the RIS depicted in Fig. 2.1, which consists of (i) a set of M passive reflecting elements and (ii) another set of $\overline{M}$ *active* channel sensors such that $\overline{M} \ll M$. The M passive elements are all implemented using RF phase shifters, and they are not connected to the baseband unit. On the other hand, the $\overline{M}$ active sensors are assumed to be selected from the passive sensors in the RIS. In particular, those sensors are designed to have two modes of operation (as shown in Fig. 2.1): (i) A channel sensing mode where they work as receivers with full RF chains and baseband processing, and (ii) a reflection mode where they act just like the rest of the passive elements that reflect the incident signal.

Before proceeding further, we need to emphasize two important points. First, while we describe the M phase-shifting elements as passive elements, they are normally implemented using reconfigurable active RF circuits [12, 49]. We just adopt

that terminology to differentiate them from the active channel sensors, i.e., they are passive in the sense that they do not provide any sensing information to the baseband. Second, our proposed architecture is different from the one proposed in [50], where an all-passive RIS assists the multiuser communication systems enabled by an all-active access point. Next, we define the channels from the transmitter/receiver to the active channel sensors of the RIS, and then we discuss how to leverage this energy-efficient RIS architecture for designing the RIS interaction vector $\boldsymbol{\psi}$.

Sampled channel vectors: We define the $\overline{M} \times 1$ uplink *sampled* channel vector, $\overline{\mathbf{h}}_{\mathrm{T},k} \in \mathbb{C}^{\overline{M}\times 1}$, as the channel vector from the transmitter to the $\overline{M}$ active elements at the RIS. This vector can then be expressed as

$$\overline{\mathbf{h}}_{\mathrm{T},k} = \mathbf{G}_{\mathrm{RIS}}\ \mathbf{h}_{\mathrm{T},k}, \tag{2.10}$$

where $\mathbf{G}_{\mathrm{RIS}}$ is an $\overline{M} \times M$ selection matrix that selects the entries of the original channel vector, $\mathbf{h}_{\mathrm{T},k}$, that correspond to the active RIS elements. If $\mathcal{A}$ defines the set of indices of the active RIS antenna elements, $|\mathcal{A}| = \overline{M}$, then $\mathbf{G}_{\mathrm{RIS}} = [\mathbf{I}]_{\mathcal{A},:}$, i.e., $\mathbf{G}_{\mathrm{RIS}}$ includes the rows of the $M \times M$ identity matrix, $\mathbf{I}$, that correspond to the indices of the active elements. The sampled channel vector, $\overline{\mathbf{h}}_{\mathrm{R},k} \in \mathbb{C}^{\overline{M}\times 1}$, from the receiver to the $\overline{M}$ active sensors of the RIS is similarly defined. Finally, $\overline{\mathbf{h}}_k = \overline{\mathbf{h}}_{\mathrm{T},k} \odot \overline{\mathbf{h}}_{\mathrm{R},k}$ is defined as the overall RIS sampled channel vector at the k^{th} subcarrier.

Designing the RIS interaction vector: In the system model and the proposed RIS architecture in, respectively, Section 2.3.1 and Fig. 2.1, the sampled channel vectors $\overline{\mathbf{h}}_{\mathrm{T},k}, \overline{\mathbf{h}}_{\mathrm{R},k}$ can easily be estimated. This is done by, for example, using an uplink training approach, in which the transmitter can send a single pilot that is simultaneously processed with all active elements to get $\overline{\mathbf{h}}_{\mathrm{T},k}$. The same approach could also be followed to estimate $\overline{\mathbf{h}}_{\mathrm{R},k}$. With the knowledge of these two sampled channels, the critical question now becomes: can we use them to select the optimal

reflection beamforming vector $\boldsymbol{\psi}^\star$ that solves (2.9)? The next three sections propose th three approaches for addressing this problem, by leveraging compressive sensing (in Section 2.6), supervised deep learning (in Section 2.7), and deep reinforcement learning (in Section 2.8).

2.6 Compressive Sensing Based RIS Interaction Design

As shown in Section 2.4, finding the optimal RIS interaction (reflection beamforming) vector $\boldsymbol{\psi}^\star$ that maximizes the achievable rate with no beam training overhead requires the availability of the full channel vectors $\mathbf{h}_{\mathrm{T},k}, \mathbf{h}_{\mathrm{R},k}$. Estimating these channel vectors at the RIS, however, normally requires that every RIS antenna gets connected to the baseband processing unit through a fully-digital or hybrid architecture [25, 45, 51]. This can massively increase the hardware complexity with the large number of antennas at the RIS systems. In this section, and adopting the low-complexity RIS architecture proposed in Section 2.5, we show that it is possible to recover the full channel vectors $\mathbf{h}_{\mathrm{T},k}, \mathbf{h}_{\mathrm{R},k}$ from the sampled channel vectors $\overline{\mathbf{h}}_{\mathrm{T},k}, \overline{\mathbf{h}}_{\mathrm{R},k}$ when the channels experience sparse scattering. This is typically the case in mmWave and LOS-dominant sub-6 GHz systems.

2.6.1 Recovering Full Channels from Sampled Channels:

With the proposed RIS architecture in Fig. 2.1, the RIS can easily estimate the *sampled* channel vectors $\overline{\mathbf{h}}_{\mathrm{T},k}, \overline{\mathbf{h}}_{\mathrm{R},k}$ through uplink training from the transmitter and receiver to the RIS with a few pilots. Next, we explain how to use these sampled channel vectors to estimate the full channel vectors $\mathbf{h}_{\mathrm{T},k}, \mathbf{h}_{\mathrm{R},k}$. First, note that the

$\mathbf{h}_{\mathrm{T},k}$ in (2.4), (2.5) (and similarly for $\mathbf{h}_{\mathrm{R},k}$) can be written as

$$\mathbf{h}_{\mathrm{T},k} = \sqrt{\frac{M}{\rho_{\mathrm{T}}}} \sum_{d=0}^{D-1} \sum_{\ell=1}^{L} \alpha_\ell \, p(dT_S - \tau_\ell) \, \mathbf{a}\left(\theta_\ell, \phi_\ell\right) e^{-j\frac{2\pi k}{K}d}, \tag{2.11}$$

$$= \sum_{\ell=1}^{L} \beta_{\ell,k} \; \mathbf{a}\left(\theta_\ell, \phi_\ell\right), \tag{2.12}$$

where $\beta_{\ell,k} = \sqrt{\frac{M}{\rho_{\mathrm{T}}}} \alpha_\ell \sum_{d=0}^{D-1} p(dT_S - \tau_\ell) e^{-j\frac{2\pi k}{K}d}$. Further, by defining the array response matrix $\mathbf{A}$ and the k^{th} subcarrier path gain vector $\boldsymbol{\beta}_k$ as

$$\mathbf{A} = \left[\mathbf{a}\left(\theta_1, \phi_1\right), \mathbf{a}\left(\theta_2, \phi_2\right) ..., \mathbf{a}\left(\theta_L, \phi_L\right)\right], \tag{2.13}$$

$$\boldsymbol{\beta}_k = \left[\beta_{1,k}, \beta_{2,k}, ..., \beta_{L,k}\right]^T, \tag{2.14}$$

we can write $\mathbf{h}_{\mathrm{T},k}$ in a more compact way as $\mathbf{h}_{\mathrm{T},k} = \mathbf{A} \; \boldsymbol{\beta}_k$. Now, we note that in several important scenarios, such as mmWave and LOS-dominant sub-6 GHz, the channel experiences sparse scattering, which results is a small number of paths L [24, 47]. In order to leverage this sparsity, we follow [45] and define the dictionary of array response vectors $\mathbf{A}_{\mathrm{D}}$, where every column constructs an array response vector in one quantized azimuth and elevation direction. For example, if the RIS adopts a uniform planar array (UPA) structure, then we can define $\mathbf{A}_{\mathrm{D}}$ as

$$\mathbf{A}_{\mathrm{D}} = \mathbf{A}_{\mathrm{D}}^{\mathrm{Az}} \otimes \mathbf{A}_{\mathrm{D}}^{\mathrm{El}} \tag{2.15}$$

with $\mathbf{A}_{\mathrm{D}}^{\mathrm{Az}}$ and $\mathbf{A}_{\mathrm{D}}^{\mathrm{El}}$ being the dictionaries of the azimuth and elevation array response vectors. Every column in $\mathbf{A}_{\mathrm{D}}^{\mathrm{Az}}$ (and similarly for $\mathbf{A}_{\mathrm{D}}^{\mathrm{El}}$) constructs an azimuth array response in one quantized azimuth (elevation) direction. If the number of grid points in the azimuth and elevation dictionaries is $N_{\mathrm{D}}^{\mathrm{Az}}$ and $N_{\mathrm{D}}^{\mathrm{El}}$, respectively, and the number of horizontal and vertical elements of the UPA is $M_{\mathrm{H}}, M_{\mathrm{V}}$, where $M = M_{\mathrm{H}} M_{\mathrm{V}}$, then $\mathbf{A}_{\mathrm{D}}$ has dimensions $M \times N_{\mathrm{D}}^{\mathrm{Az}} N_{\mathrm{D}}^{\mathrm{El}}$. Now, assuming that size of the grid is large enough such that the azimuth and elevation angles $\theta_\ell, \phi_\ell, \forall \ell$ matches exactly L points in

this grid (which is a common assumption in the formulations of the sparse channels estimation approaches [24, 45, 52]), then we can rewrite $\mathbf{h}_{\mathrm{T},k}$ as

$$\mathbf{h}_{\mathrm{T},k} = \mathbf{A}_{\mathrm{D}}\ \mathbf{x}_{\boldsymbol{\beta},k}, \tag{2.16}$$

where $\mathbf{x}_{\boldsymbol{\beta},k}$ is an $N_{\mathrm{D}}^{\mathrm{Az}} N_{\mathrm{D}}^{\mathrm{El}}$ sparse vector with $L \ll N_{\mathrm{D}}^{\mathrm{Az}} N_{\mathrm{D}}^{\mathrm{El}}$ non-zero entries equal to the elements of $\boldsymbol{\beta}_k$. Further, these non-zero entries are in the positions that correspond to the channel azimuth/elevation angles of arrival. Next, let $\widehat{\widehat{\mathbf{h}}}_{\mathrm{T},k}$ denote the noisy sampled channel vectors, then we can write

$$\widehat{\widehat{\mathbf{h}}}_{\mathrm{T},k} = \mathbf{G}_{\mathrm{RIS}}\mathbf{h}_{\mathrm{T},k} + \mathbf{v}_k, \tag{2.17}$$

$$= \mathbf{G}_{\mathrm{RIS}}\mathbf{A}_{\mathrm{D}}\ \mathbf{x}_{\boldsymbol{\beta},k} + \mathbf{v}_k, \tag{2.18}$$

$$= \boldsymbol{\Phi}\ \mathbf{x}_{\boldsymbol{\beta},k} + \mathbf{v}_k, \tag{2.19}$$

where $\mathbf{v}_k \sim \mathcal{N}_{\mathbb{C}}\left(\mathbf{0}, \sigma_n^2 \mathbf{I}\right)$ represent the receive noise vector at the RIS active channel sensors and $\mathbf{G}_{\mathrm{RIS}}$ is the selection matrix defined in (2.10). Now, given the equivalent sensing matrix, $\boldsymbol{\Phi}$ and the noisy sampled channel vector $\widehat{\widehat{\mathbf{h}}}_{\mathrm{T},k}$, the objective is to estimate the sparse vector $\mathbf{x}_{\boldsymbol{\beta},k}$ that solves the non-convex combinatorial problem

$$\min \left\|\mathbf{x}_{\boldsymbol{\beta},k}\right\|_0 \quad \text{s.t.} \ \left\|\widehat{\widehat{\mathbf{h}}}_{\mathrm{T},k} - \boldsymbol{\Phi}\ \mathbf{x}_{\boldsymbol{\beta},k}\right\|_2 \leq \sigma. \tag{2.20}$$

Given the sparse formulation in (2.20), several compressive sensing reconstruction algorithms, such as orthogonal matching pursuit (OMP) [53, 54], can be employed to find an approximate solution for $\mathbf{x}_{\boldsymbol{\beta},k}$. With this solution for $\mathbf{x}_{\boldsymbol{\beta},k}$, the full channel vector $\mathbf{h}_{\mathrm{T},k}$ can be constructed according to (2.16). Finally, the constructed full channel vector can be used to find the best RIS reflection beamforming vector, $\boldsymbol{\psi}_{n^{\mathrm{CS}}} \in \boldsymbol{\mathcal{P}}$, out of the codebook $\boldsymbol{\mathcal{P}}$, via an offline search using (2.8).

In this work, we assume for simplicity that the $\overline{M}$ active channel sensors are randomly selected from the M RIS elements, assuming that all the elements are

equally likely to be selected. It is important, however, to note that the specific selection of the active elements designs the compressive sensing matrix $\mathbf{\Phi}$ and decides its properties. Therefore, it is interesting to explore the optimization of the active element selection, leveraging tools from nested arrays [55], co-prime arrays [56, 57], incoherence frames [58], and difference sets [51, 59].

2.6.2 Simulation Results and Discussion:

To evaluate the performance of the proposed compressive sensing based solution, we consider a simulation setup at two different carrier frequencies, namely 3.5GHz and 28GHz. The simulation setup consists of one reconfigurable intelligent surface with a uniform planar array (UPA) in the y-z plane, which reflects the signal coming from one transmitter to another receiver, as depicted in Fig. 2.6. This UPA consists of 16×16 antennas at 3.5GHz and 64×64 antennas at 28GHz. We generate the channels using the publicly available ray-tracing based DeepMIMO dataset [2], with the 'O1' scenario that consists of a street and buildings on the sides of the street. Please refer to Section 2.9.1 for a detailed description of the simulation setup and its parameters.

Given this described setup, and adopting the novel RIS architecture in Fig. 2.1, we apply the proposed compressive-sensing based solution described in Section 2.6.1 as follows: (i) We obtain the channel vectors $\mathbf{h}_{\mathrm{T},k}, \mathbf{h}_{\mathrm{R},k}$ using the ray-tracing based DeepMIMO dataset, and add noise with the noise parameters described in Section 2.9.1. (ii) Adopting the RIS architecture in Fig. 2.1, we randomly select $\overline{M}$ elements to be active and construct the sampled channel vectors $\widehat{\overline{\mathbf{h}}}_{\mathrm{T},k}, \widehat{\overline{\mathbf{h}}}_{\mathrm{R},k}$. (iii) Using OMP with a grid of size $N_{\mathrm{D}}^{\mathrm{Az}} N_{\mathrm{D}}^{\mathrm{El}}, N_{\mathrm{D}}^{\mathrm{Az}} = 2M_{\mathrm{H}}, N_{\mathrm{D}}^{\mathrm{El}} = 2M_{\mathrm{V}}$, we recover an approximate solution of the full channel vectors and use this to search for the optimal RIS interaction vector using (2.8). The achievable rate using this proposed compressive sensing based solu-

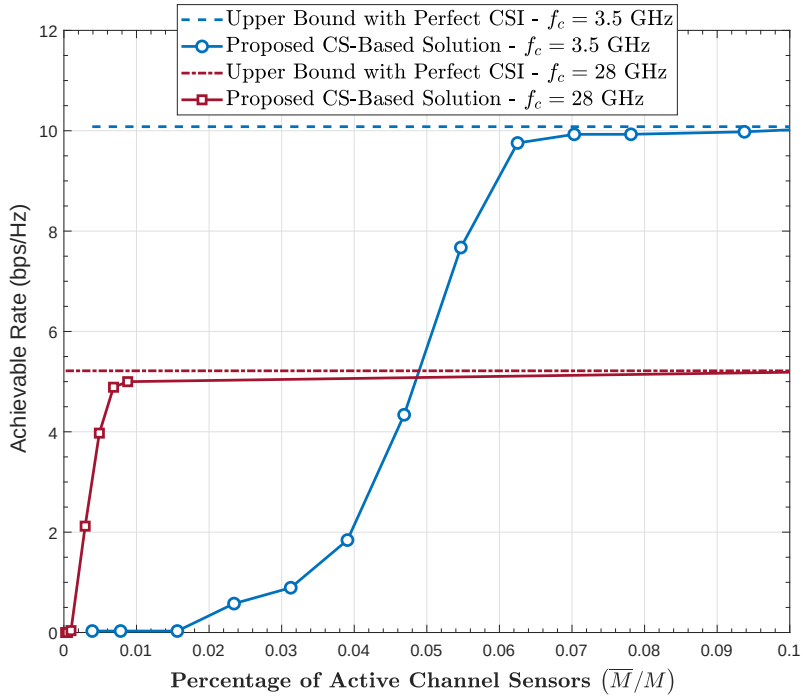


Figure 2.2: This Figure Plots the Achievable Rates Using the Proposed Compressive Sensing Based Solution for Two Scenarios, Namely a mmWave 28GHz Scenario and a Low-frequency 3.5GHz One. These Achievable Rates Are Compared to the Optimal Rate $R^\star$ in (2.9) That Assumes Perfect Channel Knowledge. This Figure Illustrates the Potential of the Proposed Solutions That Approach the Upper Bound, While Requiring Only a Small Fraction of the Total RIS Elements to Be Active.

tion is shown in Fig. 2.2 compared to the upper bound with perfect *full* channel state information (CSI), $\mathbf{h}_{\mathrm{T},k}$ and $\mathbf{h}_{\mathrm{R},k}$, calculated according to (2.9).

Gains and Limitations: In Fig. 2.2, we plot the achievable rates of the proposed compressive sensing based solution and upper bound versus the ratio of the active elements to the total number of antennas, i.e., $\overline{M}/M$. As shown in this figure, the proposed novel RIS architecture with the compressive sensing based solution can achieve almost the optimal rate with a small fraction of the RIS antennas being active. This illustrates the significant saving in power consumption that can be achieved using the RIS architecture in Fig. 2.1 that includes a few active channel sensors. Further, since the RIS reflection beamforming vector $\boldsymbol{\psi}$ is obtained through an offline search with no beam training, the proposed solution approaches the optimal rate

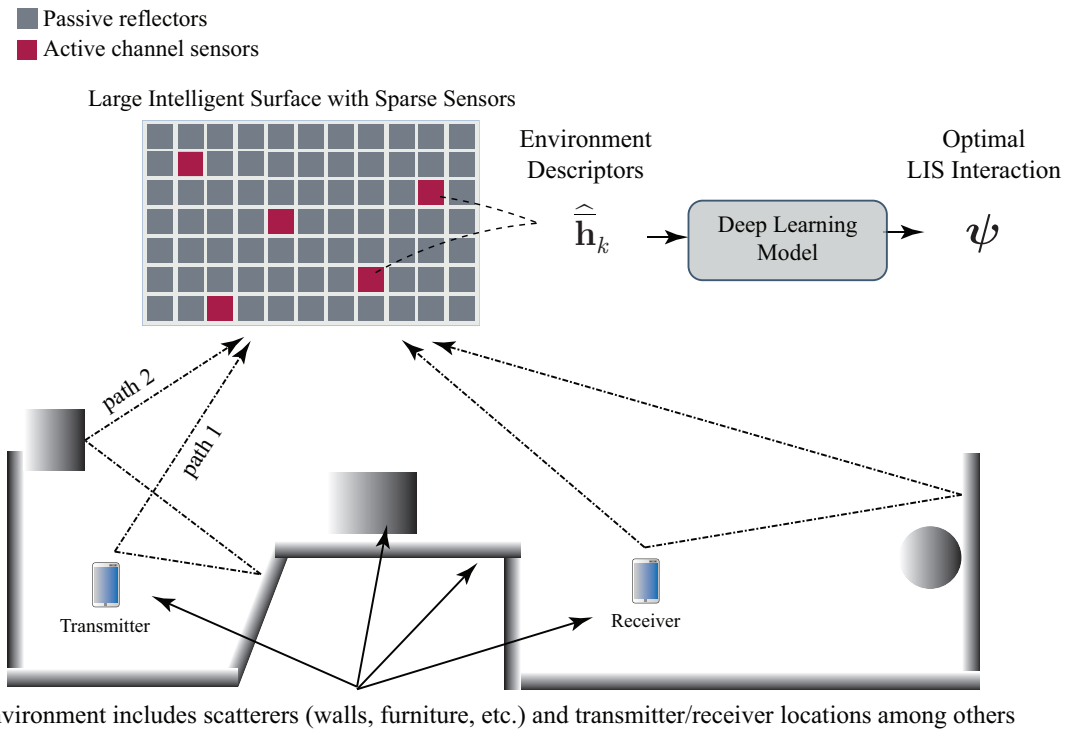


Figure 2.3: This Figure Summarizes the Key Idea of the Proposed Supervised Deep Learning (SL) Solution. The Sampled Channel Vectors Are Considered as Environment Descriptors as They Define, with Some Resolution, the Transmitter/Receiver Locations and the Surrounding Environment. The Deep Learning Model Learns How to Map the Observed Environment Descriptors to the Optimal RIS Reflection Vector.

with negligible training overhead, ideally with two uplink pilots to estimate $\widehat{\overline{\mathbf{h}}}_{\mathrm{T},k}, \widehat{\overline{\mathbf{h}}}_{\mathrm{R},k}$. This enables the proposed RIS systems to support highly mobile applications such as vehicular communications and wireless virtual/augmented reality.

Despite this interesting gain of the proposed compressive sensing based solution, it has some limitations. First, recovering the full channel vectors from the sampled ones according to Section 2.6.1 requires the knowledge of the array geometry and is hard to extend to RIS systems with unknown array structures. Second, the compressive sensing solution relies on the sparsity of the channels and its performance becomes limited in scenarios with rich NLOS scattering. This is shown in Fig. 2.2 as the compressive sensing based solution requires a higher ratio of the RIS elements to be active to approach the upper bound in the 3.5GHz scenario that has more scattering

than the mmWave 28GHz case. Further, the compressive sensing solution does not leverage previous observations to improve the current channel recovery. These limitations motivate the deep learning based solutions that we propose in the following sections.

2.7 Supervised Deep Learning Based RIS Interaction Design

In this section, we introduce a *novel* application of deep learning in the reflection beamforming design problem of reconfigurable intelligent surfaces. The section is organized as follows: First, the key idea of the proposed supervised deep learning (SL) based reflection beamforming design is explained. Then, the system operation and the adopted deep learning model are diligently described. We refer the interested reader to [60] for a brief background on deep learning.

2.7.1 Key Idea

The reconfigurable intelligent surfaces are envisioned as key components of future networks [13]. These surfaces will interact with the incident signals, for example by reflecting them, in a way that improves the wireless communication performance. To decide on this interaction, however, the RIS systems or their operating base stations and access points need to acquire some knowledge about the channels between the RIS and the transmitter/receiver. As we explained in Section 2.4, the massive number of antennas at these surfaces makes obtaining the required channel knowledge associated with (i) prohibitive training overhead if all the RIS elements are passive or (ii) infeasible hardware complexity/power consumption in the case of fully-digital or hybrid based RIS architectures.

The channel vectors/matrices, however, are intuitively some functions of the various elements of the surrounding environment such as the geometry, scatterer materi-

als, and the transmitter/receiver locations among others. Unfortunately, the nature of this function—its dependency on the various components of the environment—makes its mathematical modeling very hard and infeasible in many cases. This dependence, though, means that the interesting role the RIS is playing could be enabled with some form of awareness about the surrounding environment. With this motivation, and adopting the proposed RIS architecture in Fig. 2.1, we propose to utilize the sampled channels seen by the few active elements of the RIS as *environment descriptors*. These descriptors are expected to capture some information about the multi-path signature [40–42], as shown in Fig. 2.3. By tapping into the environment-specific information in those descriptors, a prediction on the optimal RIS interaction vector could be made using a deep learning algorithm. The algorithm is simply expected to learn a mapping function that relates the descriptor vector space with that of the RIS interaction vector. **In an abstract sense, this could be seen as teaching the RIS system how to interact with the wireless signal given the knowledge of the environment descriptors.** This is a desirable ability for the RIS to have, especially considering that the sampled channel vectors can be obtained with negligible training overhead as explained in Section 2.5. Ideally, the algorithm will learn a perfect prediction function that maps an environment descriptor to the optimal interaction vector, which means the RIS can approach the optimal rate in (2.9) with negligible training overhead and with low-complexity architectures (as only a few elements of the RIS are active).

2.7.2 Proposed System Operation

In this section, we describe the system operation of the proposed deep learning based RIS interaction solution. The proposed system operates in two phases, namely (I) the learning phase and (II) the prediction phase.

Learning phase: In this phase, the RIS employs an exhaustive search reflection beamforming approach, as will be explained shortly, while it is collecting the dataset for the deep learning model. Once the dataset is fully acquired, the RIS trains the deep learning model, which in turn will be leveraged in the prediction phase. Let the term "data sample" indicates the data point captured in one coherence block, and define the concatenated *sampled* channel vector as $\overline{\mathbf{h}} = \text{vec}\left(\left[\overline{\mathbf{h}}_1, \overline{\mathbf{h}}_2, \ldots, \overline{\mathbf{h}}_K\right]\right)$. Further, let $\overline{\mathbf{h}}(s)$ denotes the concatenated *sampled* channel vector at the s^{th} coherence block, where $s = 1, ..., S$ and S is the total number of data samples used to construct the learning dataset. As depicted in Algorithm 1, at every coherence block s, the proposed RIS system operation consists of four steps, namely (1) estimating the sampled channel vector, (2) exhaustive beam training, (3) constructing a new data point for the learning dataset, and (4) data transmission. After collecting the whole dataset with S data samples, the deep learning model is trained. We describe these steps in detail as follows.

1. Sampled channel estimation (lines 1,2): For every channel coherence block s, the transmitter and receiver transmit two orthogonal uplink pilots. The RIS active elements will receive these pilots and estimate the *sampled* channel vectors to construct the multipath signature, which is expressed as

$$\widehat{\overline{\mathbf{h}}}_{\text{T},k}(s) = \overline{\mathbf{h}}_{\text{T},k}(s) + \mathbf{v}_k, \widehat{\overline{\mathbf{h}}}_{\text{R},k}(s) = \overline{\mathbf{h}}_{\text{R},k}(s) + \mathbf{w}_k, \tag{2.21}$$

$$\widehat{\overline{\mathbf{h}}}_k(s) = \widehat{\overline{\mathbf{h}}}_{\text{T},k}(s) \odot \widehat{\overline{\mathbf{h}}}_{\text{R},k}(s), \tag{2.22}$$

$$\widehat{\overline{\mathbf{h}}}(s) = \text{vec}\left(\left[\widehat{\overline{\mathbf{h}}}_1(s), \widehat{\overline{\mathbf{h}}}_2(s), \ldots, \widehat{\overline{\mathbf{h}}}_K(s)\right]\right). \tag{2.23}$$

where $\mathbf{v}_k, \mathbf{w}_k \sim \mathcal{N}_{\mathbb{C}}\left(\mathbf{0}, \sigma_n^2 \mathbf{I}\right)$ are the receive noise vectors at the RIS active channel sensors.

2. Exhaustive beam training (lines 3-6): In this step, the RIS performs an

exhaustive search over reflection codewords using the reflection codebook $\boldsymbol{\mathcal{P}}$. Specifically, the RIS attempts every candidate reflection beamforming vector, $\boldsymbol{\psi}_n, n = 1, ..., |\boldsymbol{\mathcal{P}}|$, and receives a feedback from the receiver indicating the achievable rate attained by using this interaction vector, $R_n(s)$, which is defined as

$$R_n(s) = \frac{1}{K}\sum_{k=1}^{K}\log_2\left(1+\mathsf{SNR}\left|\left(\mathbf{h}_{\mathrm{T},k}(s)\odot\mathbf{h}_{\mathrm{R},k}(s)\right)^T\boldsymbol{\psi}_n\right|^2\right). \tag{2.24}$$

Note that, in practice, the computation and feedback of the achievable rate $R_n(s)$ will have some error compared to (2.24) because of the limitations in the pilot sequence length and feedback channel, which are neglected in this work. For the rest of this chapter, we define the achievable rate vector at the s^{th} coherence block as $\mathbf{r}(s) = \left[R_1(s), R_2(s), ..., R_{|\boldsymbol{\mathcal{P}}|}(s)\right]^T$.

3. Learning dataset update (line 7): The new data entry comprised of the sampled channel vector $\widehat{\overline{\mathbf{h}}}(s)$, estimated in step (1), and the corresponding rate vector $\mathbf{r}(s)$, constructed in step (2), is added to the deep learning dataset $\mathcal{D}$, such that $\mathcal{D} \leftarrow (\widehat{\overline{\mathbf{h}}}(s), \mathbf{r}(s))$.

4. Data transmission (line 8): After the beam training task, given the constructed achievable rate vector $\mathbf{r}(s)$, the best reflection beamforming vector, $\boldsymbol{\psi}_{n^\star}$, that corresponds to the highest achievable rate, where $n^\star = \arg\max_n \left[\mathbf{r}(s)\right]_n$, is used to reflect the transmitted data from the transmitter for the rest of the coherence block.

5. Deep learning model training (line 9): After acquiring the data entries for all S coherence blocks, the deep learning model is trained using the entire dataset $\mathcal{D}$. This model learns how to map an input (the *sampled* channel vector $\widehat{\overline{\mathbf{h}}}$) to an output (predicted achievable rate with every candidate interaction vector $\widehat{\mathbf{r}} = \left[\widehat{R}_1, \widehat{R}_2, \ldots, \widehat{R}_{|\boldsymbol{\mathcal{P}}|}\right]$, as shown in Fig. 2.4. It is worth mentioning here that while

Algorithm 1 Supervised Deep Learning Based Reflection Beamforming Prediction

Inputs: Reflection beamforming codebook $\boldsymbol{\mathcal{P}}$.

Phase I: Learning phase

1: **for** $s = 1$ **to** S **do** ▷ *For every channel coherence block*

2: RIS receives two pilots to estimate $\widehat{\overline{\mathbf{h}}}(s)$.

3: **for** $n = 1$ **to** $|\boldsymbol{\mathcal{P}}|$ **do** ▷ *Beam training*

4: RIS reflects using $\boldsymbol{\psi}_n$ beam.

5: RIS receives the feedback $R_n(s)$.

6: Construct $\mathbf{r}(s) = \left[R_1(s), R_2(s), ..., R_{|\boldsymbol{\mathcal{P}}|}(s)\right]^T$.

7: Store new entry in the learning dataset, $\mathcal{D} \leftarrow \left(\widehat{\overline{\mathbf{h}}}(s), \mathbf{r}(s)\right)$.

8: RIS reflects using $\boldsymbol{\psi}_{n^\star}$ beam, $n^\star = \arg\max_n \left[\mathbf{r}(s)\right]_n$.

9: Train the SL model using the learning dataset $\mathcal{D}$.

Phase II: Prediction phase

10: **while** True **do** ▷ *Repeat for every channel coherence block*

11: RIS receives two pilots to estimate $\widehat{\overline{\mathbf{h}}}$.

12: Predict the rate vector $\widehat{\mathbf{r}}$ using the trained SL model.

13: RIS reflects using $\boldsymbol{\psi}_{n^{\mathrm{SL}}}$ beam, $n^{\mathrm{SL}} = \arg\max_n \left[\widehat{\mathbf{r}}\right]_n$.

we assume that the system will switch one time to Phase II after the deep learning model is trained, the system will need to retrain and refine the model frequently to account for the changes in the environment.

Prediction phase: Following the deep learning model training in the learning phase, the RIS leverages the trained model to predict the reflection beamforming vector directly from the estimated *sampled* channel vector, $\widehat{\overline{\mathbf{h}}}$. As shown in Algorithm 1, Phase II performs the following steps repeatedly for every channel coherence block.

1. Sampled channel estimation (line 11): This step is the same as the first

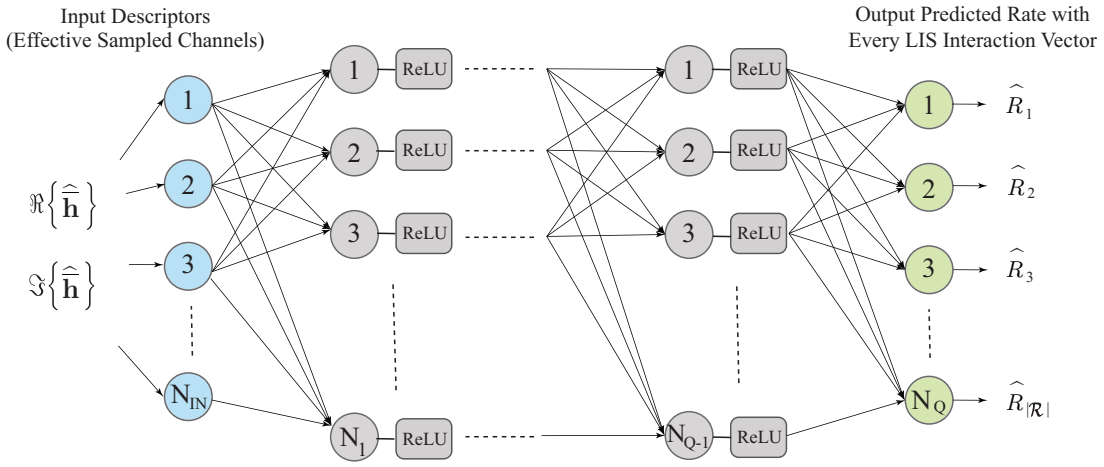


Figure 2.4: The Adopted Neural Network Architecture Consists of Q Fully Connected Layers. Each Layer Is Followed by a Non-linear ReLU Activation Layer. The Deep Learning Model Learns How to Map the Observed Sampled Channel Vectors to the Predicted Achievable Rate Using Every RIS Interaction Vector.

step in the learning phase. The active elements of the RIS receive uplink pilots to estimate and construct the concatenated *sampled* channel vector, $\widehat{\overline{\mathbf{h}}}$.

2. Achievable rate prediction (line 12): In this step, the estimated *sampled* channel vector, $\widehat{\overline{\mathbf{h}}}$, is fed into the trained deep learning model. It predicts the achievable rate vector, $\widehat{\mathbf{r}}$, which is used to identify the best SL-based reflection beamforming vector.

3. Data transmission (line 13): In this step, the predicted deep learning reflection beamforming vector, $\boldsymbol{\psi}_{n^{\text{SL}}}$, that corresponds to the highest predicted achievable rate, where $n^{\text{SL}} = \arg\max_n [\widehat{\mathbf{r}}]_n$, is used to reflect the transmitted data from the transmitter for the rest of the coherence block. Note that instead of selecting only the interaction vector with the highest predicted achievable rate, the RIS can generally select the k_{B} beams corresponding to the k_{B} highest predicted achievable rates. It can then refine this set of beams online with the receiver to select the one with the highest achievable rate. In Section 2.9.6, we evaluate the performance gain if more than one reflection beam, i.e. k_{B} reflection beams, are selected.

2.7.3 Deep Learning Model

Recent advances in machine learning have proven deep learning to be one of the most successful learning paradigms [61]. With this motivation, a deep neural network is chosen in this work to be the model with which the desired RIS interaction function is learned. In the following, the elements of this model are described.

Input Representation: A single input to the neural network model is defined as a stack of environment descriptors at K sub-carrier frequencies, i.e., the *sampled* channel vector $\widehat{\overline{\mathbf{h}}}$. This sets the dimensionality of a single input vector to $K\overline{M}$. A common practice in machine learning is the normalization of the input data. This guarantees a stable and meaningful learning process [62]. The normalization method of choice here is a simple per-dataset scaling; all samples are normalized by one constant value over the whole input data,

$$\widehat{\overline{\mathbf{h}}}_{\text{norm}}(s) = \frac{\widehat{\overline{\mathbf{h}}}(s)}{\max_s \left\| \widehat{\overline{\mathbf{h}}}(s) \right\|_\infty}, \; s = 1, \ldots, S. \tag{2.25}$$

Besides helping the learning process, this normalization choice preserves distance information encoded in the environment descriptors. This way the model learns to become more aware of the surroundings, which is the bedrock for proposing a machine-learning-powered RIS.

The last pre-processing step of input data is to convert them into real-valued vectors without losing the imaginary-part information. This is done by splitting each complex entry into real and imaginary values, doubling the dimensionality of each input vector. The main reason behind this step is the modern implementations of SL models, which mainly use real-valued computations.

Target Representation: The learning approach used in this work is supervised learning. This means the model is trained with input data that are accompanied by their so-called *target responses* [60]. They are the desired responses the model is

expected to approximate when it encounters inputs like those in the input training data. Since the target of the training process is to learn a function mapping descriptors to reflection vectors, the model is designed to output a set of predictions on the achievable rates of every possible reflection beamforming vector in the codebook $|\boldsymbol{\mathcal{P}}|$. Hence, the training targets are real-valued vectors, $\mathbf{r}(s), s = 1, ..., S$, with the desired rate for each possible reflection vector.

For the same training-efficiency reason expressed for the input representation, the labels are usually normalized. The normalization used in this work is per-sample scaling where every vector of rates $\mathbf{r}(s)$ is normalized using its maximum rate value $\max_n [\mathbf{r}(s)]_n$. The output of the normalization process is denoted by $\widehat{\mathbf{r}}(s)$. The choice of normalizing each vector independently guards the model against being biased towards some strong responses. In terms of our RIS application, it gives the receivers equal importance regardless of how close or far they are from the RIS.

Neural Network Architecture: The SL model is designed as a Multi-Layer Perceptron (MLP) network, sometimes referred to as a feedforward Fully Connected network. It is well-established that MLP networks are universal function approximators [63]. This motivates adopting an MLP network to capture the relation between the environment descriptors and the RIS interaction (reflection beamforming) vectors. As depicted in Fig. 2.4, the proposed MLP model consists of Q layers. The first $Q-1$ of them alternate between fully connected and non-linearity layers and the last layer (output layer) is a fully connected layer. For the fully connected layers, each Layer q in the network has a stack of N_q neurons, each of which sees all the outputs of the previous layer. For the non-linearity layers, they all employ Rectified Linear Units (ReLUs) [60].

Training Loss Function: The model training process aims at minimizing a loss function that measures the quality of the model predictions. Given the objective of

predicting the best reflection beamforming vector, $\boldsymbol{\psi}_{n^{\mathrm{SL}}}$, having the highest achievable rate estimate, $\max_n \widehat{R}_n$, the model is trained using a regression loss function. At every coherence block, the neural network is trained to make its output, $\widehat{\mathbf{r}}$, as close as possible to the desired output, the normalized achievable rates, $\overline{\mathbf{r}}$. Specifically, the training is guided through minimizing the loss function, $L(\boldsymbol{\theta})$, expressed as

$$L(\boldsymbol{\theta}) = \mathsf{MSE}(\overline{\mathbf{r}}, \widehat{\mathbf{r}}), \tag{2.26}$$

where $\boldsymbol{\theta}$ represents the set of all the neural network parameters and $\mathsf{MSE}(\overline{\mathbf{r}}, \widehat{\mathbf{r}})$ indicates the mean-squared-error between $\overline{\mathbf{r}}$ and $\widehat{\mathbf{r}}$.

From the training overhead standpoint, the SL based reflection beamforming solution, however, still demands a large dataset collection phase before training, in the learning phase. Given an end goal of achieving harmonic co-existence between all the heterogeneous wireless systems, setting an objective of developing *standalone* RIS architectures with no beam training overhead appears like the next step forward for reaching that end goal. To reach this objective, we propose a deep reinforcement Learning (DRL) based reflection beamforming design approach in the upcoming section.

2.8 Deep Reinforcement Learning Based RIS Interaction Design

In this section, we introduce a *novel* application of deep reinforcement learning in predicting the RIS reflection coefficients without requiring any prior training overhead, as detailed in [30]. The section is organized as follows: First, the key idea of the proposed deep reinforcement learning (DRL) based reflection beamforming design is explained. Then, the system operation and the deep learning model are detailed.

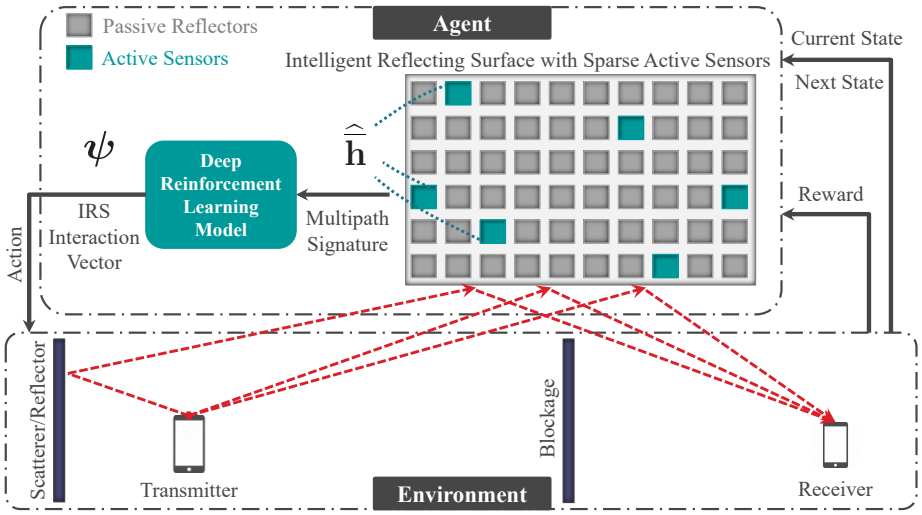


Figure 2.5: This Figure Summarizes the Key Idea of the Proposed Deep Reinforcement Learning (DRL) Solution. The Transmitter-receiver Communication Is Assisted by a Reconfigurable Intelligent Surface (RIS). The RIS Is Interacting with the Incident Signal Through an Interaction Vector ψ. The Environment Is Represented by Various Scatterers, User Locations, Etc. The RIS Acts as a Reinforcement Learning Agent by Acquiring a State and a Reward from the Environment and Exerting an Action Back on the Environment.

2.8.1 Key Idea

From (2.8), the optimal interaction vector is a function of the channels between the two communication ends and the RIS. To avoid the prohibitive overhead of estimating the full RIS channels, the optimal interaction vector choice can be mapped to the surrounding *environment*, which the full RIS channels inherently describe. Modeling the various elements of the environment, mathematically, is notoriously complicated. In contrast, leveraging an awareness of the environment using a multipath signature [40] can be sufficient. In such case, deep reinforcement learning models can be adopted to learn the mapping function from multipath signatures to the optimal interaction vectors as illustrated in Fig. 2.5. The RIS active elements play a crucial role in capturing one form of multipath signatures: the *sampled* channels, $\overline{\mathbf{h}}_{\mathrm{T},k}, \overline{\mathbf{h}}_{\mathrm{R},k}$. Fortunately, estimating the sampled channel vectors can be accomplished with a few pilot signals; i.e., negligible training overhead. This solution also involves energy-efficient

low-complexity hardware architectures (few sparse active RIS elements) [28].

2.8.2 Proposed System Operation

The proposed deep reinforcement learning (DRL) based RIS interaction prediction solution operates in two modes: (I) the agent interaction and (II) the agent learning, as in Algorithm 2. The RIS interchanges between these two tasks continuously; this cycle repeats over time and across multiple users.

Task I: Agent Interaction: The RIS interaction with the *environment* can be outlined as follows: the RIS observes the current *state*, s, of the environment and takes an *action*, a, predicated upon the observed state. The RIS then receives a *reward*, r, for the action taken and a new *state* observation, s', from the environment. Once the experience is acquired, $\langle s, a, r, s' \rangle$, the RIS trains the DRL model using current and past experiences, in the second task.

Let the term "experience" indicates the information captured in one learning episode, and define the concatenated *sampled* channel vector as

$$\overline{\mathbf{h}} = \mathbf{vec}\left(\left[\overline{\mathbf{h}}_1, \overline{\mathbf{h}}_2, \ldots, \overline{\mathbf{h}}_K\right]\right). \tag{2.27}$$

Assume that the one learning episode occurs every coherence block and let T be the maximum number of episodes, $\overline{\mathbf{h}}(t)$ denotes the concatenated *sampled* channel vector at the t^{th} episode, where $t = 1, ..., T$. Task I steps are summarized as follows.

1. Sampled channel estimation (lines 3 and 13): The transmitter and receiver transmit two orthogonal uplink pilots. The RIS active elements will receive these pilots and estimate the *sampled* channel vectors to construct the multipath signature.

Algorithm 2 Deep Reinforcement Learning Based RIS Interaction Prediction

Inputs: Reflection beamforming codebook $\boldsymbol{\mathcal{P}}$.

Outputs: Trained network $Q\left(s, a|\boldsymbol{\theta}\right)$.

Initialize: Network $Q\left(s, a|\boldsymbol{\theta}\right)$, replay buffer $\mathcal{D}$.

repeat

  RIS receives two pilots to estimate $\widehat{\overline{\mathbf{h}}}(1)$. ▷ Current state

  for episode $t = 1$ **to** T **do** ▷ For every episode

    Task I: Agent Interaction

    Sample $\xi \sim \text{Uniform}\left(0, 1\right)$

    if $\xi \leq \epsilon$ **then** ▷ Select action

      Select interaction vector, $\boldsymbol{\psi}(t) \in \boldsymbol{\mathcal{P}}$ at random.

    else

      Select interaction vector, $\boldsymbol{\psi}(t) = \arg\max_{a'} Q\left(s, a'|\boldsymbol{\theta}\right)$.

    RIS reflects using $\boldsymbol{\psi}(t)$ beam. ▷ Carry out action

    RIS receives the feedback $R(t)$. ▷ Observe reward

    RIS quantizes the reward, $R_{\text{Q}}(t) \in \{\pm 1\}$.

    RIS receives two pilots to estimate $\widehat{\overline{\mathbf{h}}}(t+1)$ ▷ Next state

    Task II: Agent Learning

    $\langle s, a, r, s' \rangle \leftarrow \left\langle \widehat{\overline{\mathbf{h}}}(t), \boldsymbol{\psi}(t), R_{\text{Q}}(t), \widehat{\overline{\mathbf{h}}}(t+1) \right\rangle$.

    Store experience $\langle s, a, r, s' \rangle$ in $\mathcal{D}$ then minibatch the experiences from $\mathcal{D}$ for training.

    Feedforward s to calculate $\widehat{\mathbf{R}}(t) \leftarrow Q\left(s, a|\boldsymbol{\theta}\right) \forall a$.

    Feedforward s' to calculate $\Gamma \leftarrow \max_{a'} Q\left(s', a'|\boldsymbol{\theta}\right)$ and $a^{\star} \leftarrow \arg\max_{a'} Q\left(s', a'|\boldsymbol{\theta}\right)$.

    Construct the target vector, $\overline{\mathbf{R}}(t)$:

$$\left[\overline{\mathbf{R}}(t)\right]_{a^{\star}} \leftarrow R_{\text{Q}}(t) + \gamma\Gamma, \quad \left[\overline{\mathbf{R}}(t)\right]_{a' \neq a^{\star}} \leftarrow \left[\widehat{\mathbf{R}}(t)\right]_{a' \neq a^{\star}}, a' \in \{1, \ldots, |\boldsymbol{\mathcal{P}}|\}.$$

    Perform SGD on $\mathsf{MSE}\left(\overline{\mathbf{R}}(t), \widehat{\mathbf{R}}(t)\right)$ to find $\boldsymbol{\theta}^{\star}$.

    Update network weights $\boldsymbol{\theta}(t) \leftarrow \boldsymbol{\theta}^{\star}$ and decrease ϵ gradually.

    $s \leftarrow s'$. ▷ Assign next state to current state

until reaching a terminal goal

$$\widehat{\overline{\mathbf{h}}}_{\mathrm{T},k}(t) = \overline{\mathbf{h}}_{\mathrm{T},k}(t) + \mathbf{v}_k, \widehat{\overline{\mathbf{h}}}_{\mathrm{R},k}(t) = \overline{\mathbf{h}}_{\mathrm{R},k}(t) + \mathbf{w}_k, \tag{2.28}$$

$$\widehat{\overline{\mathbf{h}}}_{k}(t) = \widehat{\overline{\mathbf{h}}}_{\mathrm{T},k}(t) \odot \widehat{\overline{\mathbf{h}}}_{\mathrm{R},k}(t), \tag{2.29}$$

$$\widehat{\overline{\mathbf{h}}}(t) = \mathbf{vec}\left(\left[\widehat{\overline{\mathbf{h}}}_1(t), \widehat{\overline{\mathbf{h}}}_2(t), \ldots, \widehat{\overline{\mathbf{h}}}_K(t)\right]\right). \tag{2.30}$$

where $\mathbf{v}_k, \mathbf{w}_k \sim \mathcal{N}_{\mathbb{C}}\left(\mathbf{0}, \sigma_n^2 \mathbf{I}\right)$ are the receive noise vectors.

2. Interaction prediction (lines 5-10): The multipath signature is used to predict the interaction vector. To account for exploration (i.e., randomly sampling from the action space) besides exploitation (i.e., using prior learning experience), the factor ϵ is introduced such that an interaction vector can be randomly chosen out of the codebook $\boldsymbol{\mathcal{P}}$ with ϵ probability. Otherwise, the interaction vector is predicted from the current network. After that, the interaction vector chosen reflects the transmitted data from the transmitter.

3. Feedback reception (lines 11 and 12): The RIS receives feedback from the receiver indicating the achievable rate, $R(t)$, attained by using the interaction vector, which is defined as

$$R(t) = \frac{1}{K}\sum_{k=1}^{K} \log_2\left(1 + \mathsf{SNR}\left|\left(\mathbf{h}_{\mathrm{T},k}(t) \odot \mathbf{h}_{\mathrm{R},k}(t)\right)^T \boldsymbol{\psi}_a\right|^2\right). \tag{2.31}$$

After that, the rate is quantized based on a threshold level, such that $R_{\mathrm{Q}}(t) = 1$ if $R(t) > R^{\mathrm{TH}}$; otherwise, $R_{\mathrm{Q}}(t) = -1$. Reward clipping is substantial for learning convergence [64].

Task II: Agent Learning: The RIS leverages the acquired experiences to train the DRL model. Task II steps are summarized as follows.

1. Constructing a new experience (lines 14 and 15): The new experience acquired is now stored in the experience replay buffer $\mathcal{D}$ for the training of the deep Q-network [65].

2. Model training (lines 16-21): The deep Q-network is now trained to minimize the prediction loss. To do so, we use the stochastic gradient descent algorithm (SGD). The training operates sequentially using minibatchs from the replay buffer $\mathcal{D}$. It learns how to map an input state (*sampled* channel vector) to an output action (interaction vector).

2.8.3 Machine Learning Design

Input Representation: the concatenated *sampled* channel vector, $\widehat{\overline{\mathbf{h}}}$, is the input to the deep Q-network. The normalization method used is a simple per-dataset scaling [66]; all samples are normalized by the maximum absolute value over the whole input data. This method preserves distance information encoded in the multipath signatures. Each complex entry of the input data is split into real and imaginary values, doubling the dimensionality of each input vector to $2K\overline{M}$.

Q-Network Architecture: The Q-network is designed as a Multi-Layer Perceptron network of U layers. The first $U-1$ of them alternate between fully-connected and rectified linear unit layers and the last one (output layer) is a fully-connected layer. The u^{th} layer in the network has a stack of A_u neurons. Two deep Q-networks are used for training stability [67].

Training Loss Function: Given the objective of predicting the best interaction vector (with the highest achievable rate), the model is trained using a regression loss function. At the t^{th} episode, the training is guided through minimizing the loss function, $\mathsf{MSE}\left(\overline{\mathbf{R}}(t), \widehat{\mathbf{R}}(t)\right)$, which is the mean-squared-error between the desired and the predicted output, $\overline{\mathbf{R}}(t)$ and $\widehat{\mathbf{R}}(t)$.

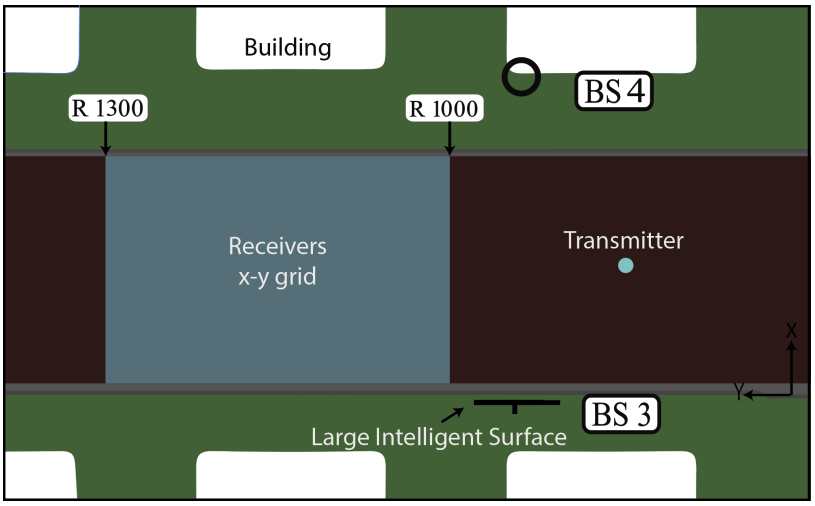


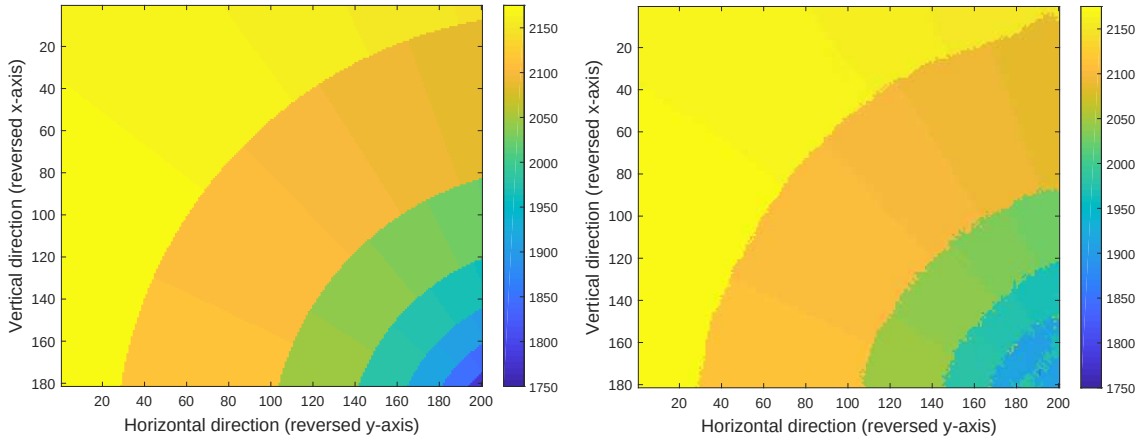
Figure 2.6: This Figure Illustrates the Adopted Ray-tracing Scenario Where an RIS Is Reflecting the Signal Received from One Fixed Transmitter to a Receiver. The Receiver Is Selected from an X-Y Grid of Candidate Locations. This Ray-tracing Scenario Is Generated Using Remcom Wireless InSite [1], And Is Publicly Available on the DeepMIMO Dataset [2].

2.9 Simulation Results

In this section, we evaluate the performance of the compressive sensing (CS), the supervised deep learning (SL), and the deep reinforcement learning (DRL) based reflection beamforming solutions. The flow of this section is as follows. First, we describe the adopted experimental setup and datasets. Then, we compare the performance of the supervised deep learning and compressive sensing solutions at both mmWave and sub-6 GHz bands. After that, we compare the performance of the supervised deep learning and deep reinforcement learning solutions. Lastly, we investigate the impact of different system and machine learning parameters on the performance of the deep learning solution.

2.9.1 Simulation Setup

Given the geometric channel model adopted in Section 2.3 and the nature of the reflection beamforming optimization problem, with its strong dependence on the



(a) Original Codebook Beams (b) Predicted Codebook Beams

Figure 2.7: This Figure Illustrates the *Optimal* and *Predicted* Index Map of the RIS Reflection Bemforming Codebook. Each Pixel Represents the Location of a Candidate Receiver on the X-Y User Grid Under-study (Shown In Fig. 2.6). The Pixel Color Represents the Index of the Optimal/Predicted Reflection Beamforming Vector for the User at This Location. In This Scenario with 64×64 RIS, the Optimum Achievable Rate, $R^{\star}$, Averaged Across All Candidate Locations, Is 5.06 bps/Hz, While the Achievable Rate of the Proposed Deep Learning Based Predicted Beams Is 4.74 bps/Hz.

environmental geometry, it is critical to evaluate the performance of the proposed solutions based on realistic channels. This motivates using channels generated by ray-tracing to capture the dependence on the key environmental factors such as the environment geometry and materials, the RIS and transmitter/receiver locations, the operating frequency among others. To do that, we adopted the DeepMIMO dataset, described in detail in [2], to generate the channels based on the outdoor ray-tracing scenario 'O1' [1], as will be discussed shortly. The DeepMIMO is a parameterized dataset published for deep learning applications in mmWave and massive MIMO systems. The machine learning simulations were executed using the Deep Learning Toolbox of MATLAB R2019a. The source code of this work is available on [68]. Next, we explain in detail the key components of the simulation setup.

System model: Following the system model in Section 2.3.1, we adopt an RIS-assisted communication system where one RIS aims to reflect the signal received from

Table 2.1: The Adopted DeepMIMO Dataset Parameters

DeepMIMO Dataset Parameter	**Value**
Frequency band	3.5GHz or 28GHz
Active BSs	3
Number of BS Antennas	$(M_x, M_y, M_z) \in \{(1, 16, 16)\,;$ $(1, 32, 32)\,; (1, 40, 10)\,; (1, 64, 64)\}$
Active users (receivers)	From row R1000 to row R1300
Active user (transmitter)	row R850 column 90
System bandwidth	100MHz
Number of OFDM subcarriers	512
OFDM sampling factor	1
OFDM limit	64
Number of channel paths	$\{1, 2, 5, 10, 15\}$
Antenna spacing	0.5λ

a transmitter to a receiver. The transmitter is assumed to be fixed in position while the receiver can take any random position in a specified x-y grid as illustrated in Fig. 2.6. We implemented this setup using the outdoor ray-tracing scenario 'O1' of the DeepMIMO dataset that is publicly available at [2]. As shown in Fig. 2.6, we select BS 3 in the 'O1' scenario to be the RIS and the user in row R850 and column 90 to be the fixed transmitter. The uniform x-y grid of candidate receiver locations includes 54300 points from row R1000 to R1300 in the 'O1' scenario where every row consists of 181 points. Unless otherwise stated, the adopted RIS employs a UPA with 64×64 ($M = 4096$) antennas at the mmWave 28GHz setup and a UPA with 16×16 ($M = 256$) antennas at the 3.5GHz setup. The active channel sensors described in

Section 2.5 are randomly selected from the M UPA antennas. The transmitter and receiver are assumed to have a single antenna each. The antenna elements have a gain of 3dBi and the transmit power is 35dBm. The antenna element spacing is set to half the wavelength, 0.5λ, where λ is the operating wavelength. The rest of the adopted DeepMIMO dataset parameters are summarized in Table 2.1.

Channel generation: The channels between the RIS and the transmitter/receiver, $\mathbf{h}_{\mathrm{T},k}, \mathbf{h}_{\mathrm{R},k}$, for all the candidate receiver locations in the x-y grid, are constructed using the DeepMIMO dataset generation code [2] with the parameters in Table 2.1. With these channels, and given the randomly selected active elements in the proposed RIS architecture, we construct the sampled channel vectors $\overline{\mathbf{h}}_{\mathrm{T},k}, \overline{\mathbf{h}}_{\mathrm{R},k}$. The noisy sampled channel vectors $\widehat{\overline{\mathbf{h}}}_{\mathrm{T},k}, \widehat{\overline{\mathbf{h}}}_{\mathrm{R},k}$ are then generated by adding noise vectors to $\overline{\mathbf{h}}_{\mathrm{T},k}, \overline{\mathbf{h}}_{\mathrm{R},k}$ according to (2.23), with the noise power calculated based on the bandwidth and other parameters in Table 2.1, and with receiver noise figure of 5dB. These noisy sampled channels are then used to design the RIS interaction (reflection beamforming) vectors following the proposed compressive sensing and deep learning solutions.

RIS interaction (reflection beamforming) codebook: We adopt a DFT codebook for the candidate RIS interaction vectors. More specifically, considering the UPA structure, we define the RIS interaction codebook as $\mathrm{DFT}_{M_\mathrm{H}} \otimes \mathrm{DFT}_{M_\mathrm{V}}$. The codebook $\mathrm{DFT}_{M_\mathrm{H}} \in \mathbb{C}^{M_\mathrm{H} \times M_\mathrm{H}}$ is a DFT codebook for the azimuth (horizontal) dimension where the m_Hth column, $m_\mathrm{H} = 1, 2, ..., M_\mathrm{H}$, is defined as $[1, e^{-j\frac{2\pi}{M_\mathrm{H}} m_\mathrm{H}}, ..., e^{-j(M_\mathrm{H}-1)\frac{2\pi}{M_\mathrm{H}} m_\mathrm{H}}]^T$. The codebook $\mathrm{DFT}_{M_\mathrm{V}}$ is similarly defined for the elevation (vertical) dimension.

As an example, Fig. 2.7(a) illustrates the *optimal* index map of the RIS reflection bemforming codebook at $f_\mathrm{c} = 28$ GHz, $M = 64 \times 64$ antennas, and $L = 1$ channel path. The map orientation and directions are set according to the adopted ray-tracing scenario, previously shown in Fig. 2.6. The pixel position represents the candidate location of the receiver on the x-y grid under-study. The pixel color represents the in-

dex number of the optimal reflection beamforming vector for each candidate location, calculated according to (2.8), under the assumption of perfect *full* channel knowledge, $\mathbf{h}_{\mathrm{T},k}$ and $\mathbf{h}_{\mathrm{R},k}$, at the RIS. By comparison, Fig. 2.7(b) depicts the *predicted* index map of the RIS reflection bemforming codebook using the proposed Deep Learning (SL) based reflection beamforming with only $\overline{M} = 8$ active channel sensors.

Compressive sensing parameters: We consider the developed compressive sensing solution in Section 2.6 to recover the full RIS-transmitter/receiver channels and design the RIS reflection beamforming vectors. For approximating the solution of (2.20), we use OMP with a grid of size $N_{\mathrm{D}}^{\mathrm{Az}} N_{\mathrm{D}}^{\mathrm{El}}$ points, where $N_{\mathrm{D}}^{\mathrm{Az}} = 2M_{\mathrm{H}}, N_{\mathrm{D}}^{\mathrm{El}} = 2M_{\mathrm{V}}$.

Supervised deep learning parameters: We adopt the deep learning model described in Section 2.7.3. To reduce the neural network complexity, however, we input the normalized sampled channels only at the first $K_{\mathrm{SL}} = 64$ subcarriers, $\widehat{\overline{\mathbf{h}}}_k$, $k = 1, \ldots, K_{\mathrm{SL}}$ and $K_{\mathrm{SL}} \leq K$, which sets the length of the SL input vector to be $2\overline{M}K_{\mathrm{SL}}$. This is motivated by the fact that the channel is highly correlated in the frequency domain, a consequence of channel sparsity, especially in the mmWave range. The length of the SL output vector is $M = |\boldsymbol{\mathcal{P}}|$, as described in Section 2.7.3. The neural network architecture consists of four fully connected layers. Unless otherwise mentioned, the number of hidden nodes of the four layers is $(M, 4M, 4M, M)$, where M is the number of RIS antennas. Given the size of the x-y grid of the candidate receiver locations in Fig. 2.6, the deep learning dataset has 54300 data points. We split this dataset into two sets, namely a training set and a testing set with 85% and 15% of the points, respectively. A dropout layer is added after every ReLU layer. Unless otherwise mentioned, we consider a batch size of 500 samples, a 50% dropout rate, an L_2 regularization factor of 10^{-4}, and 20 epochs of training. The learning rate starts from 0.1 and drops by 50% every 3 epochs.

Deep reinforcement learning parameters: We adopt the DRL model described in

Section 2.8.3. States are represented by the normalized concatenated *sampled* channel of each user pair, and actions are represented by each candidate interaction vector, $\boldsymbol{\psi} \in \boldsymbol{\mathcal{P}}$. To reduce the Q-network complexity, we input the normalized *sampled* channels only at the first 64 subcarriers. The neural network architecture of the Q-network consists of four fully-connected layers of $4096, 16384, 16384, 4096$ nodes, respectively. We consider a replay buffer of 8192 samples and a batch size of 512 samples. ϵ starts from 0.99 and decreases gradually by a factor of 0.5% every 40 training iterations till it reaches 0.1. $\gamma = 0$. $R^{\mathrm{TH}} = 8.9$ bps/Hz is set to the min-max rate of the dataset. Only when evaluating the DRL based solution, and when comparing the SL based solution with the DRL based solution, we follow the following dataset settings. we select the size of the receiver x-y grid From row R1000 to row R1200, the DRL dataset has 36200 data points. We split this dataset into two sets: training and testing sets, with 70% and 30% of the points, respectively.

Next, given this described setup, and adopting the novel RIS architecture in Fig. 2.1 with only $\overline{M}$ active channel sensors, we evaluate the performance of the developed compressive sensing and deep learning solutions.

2.9.2 Achievable Rates with Compressive Sensing and Deep Learning Based RIS Systems

In this subsection, we evaluate the achievable rates of the proposed compressive sensing (CS), supervised deep learning (SL), and deep reinforcement learning (DRL) based reflection beamforming solutions for RIS systems, as previously described in Section 2.6.1, Section 2.7.2, and Section 2.8.2, respectively. These rates are compared to the genie-aided upper bound, $R^{\star}$, in (2.9) which assumes perfect knowledge of the *full* channel vectors, $\mathbf{h}_{\mathrm{T},k}$ and $\mathbf{h}_{\mathrm{R},k}$, at the RIS. The average achievable rate used for

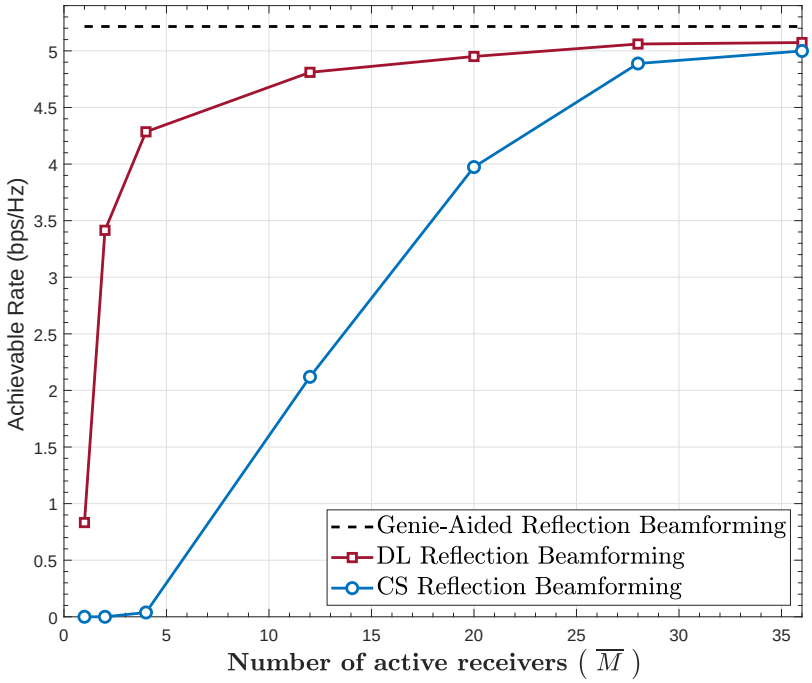


Figure 2.8: The Achievable Rate of Both Proposed CS and SL Based Reflection Beamforming Solutions Are Compared to the Upper Bound $R^{\star}$, for Different Numbers of Active Receivers, $\overline{M}$. The Figure Is Generated At $f_{\mathrm{c}} = 28$GHz, $M = 64 \times 64$ Antennas, and $L = 10$ Paths.

assessing the performance of these proposed solutions can be formulated as

$$R = \frac{1}{K} \sum_{k=1}^{K} \log_2 \left(1 + \mathsf{SNR} \left| \left(\mathbf{h}_{\mathrm{T},k} \odot \mathbf{h}_{\mathrm{R},k} \right)^T \boldsymbol{\psi} \right|^2 \right), \tag{2.32}$$

where $\boldsymbol{\psi} \in \{ \boldsymbol{\psi}_{n^{\mathrm{CS}}}, \boldsymbol{\psi}_{n^{\mathrm{SL}}}, \boldsymbol{\psi}_{n^{\mathrm{DRL}}} \}$ is the reflection beamforming vector chosen by the CS, SL or DRL based reflection beamforming solutions, respectively. To reduce the computational complexity of the performance evaluation, we compute the achievable rate summation over the first subcarrier instead of computing over all the $K = 512$ subcarriers.

In Fig. 2.8, we consider the simulation setup in Section 2.9.1 at the mmWave 28GHz band with RIS employing a UPA of 64×64 antennas. The channels are constructed to include the strongest $L = 10$ channel paths. Fig. 2.8 shows that the proposed supervised deep learning (SL) solution approaches the optimal upper bound with a very small number of active antennas. For example, with only $\overline{M} = 4$ active

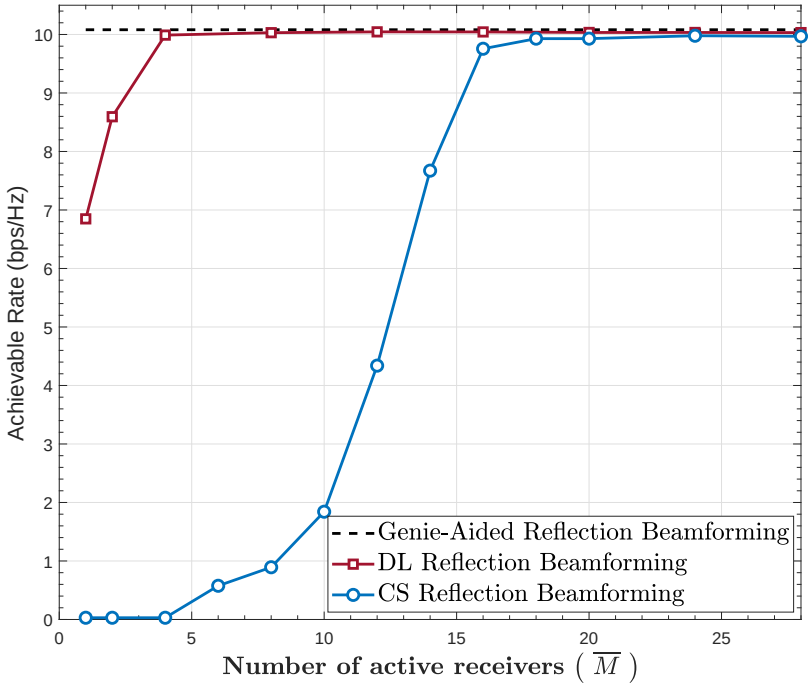


Figure 2.9: The Achievable Rate of Both Proposed CS and SL Based Reflection Beamforming Solutions Are Compared to the Upper Bound $R^\star$, for Different Numbers of Active Receivers, $\overline{M}$. The Figure Is Generated At $f_c = 3.5$GHz, $M = 16 \times 16$ Antennas, and $L = 15$ Paths.

antennas (out of $M = 4096$ total antennas), the supervised deep learning solution achieves almost 85% of the optimal achievable rate. This figure also illustrates the performance gain of the SL solution compared to the compressive sensing solution, especially when the number of active antennas is very small. Note that the two CS and SL solutions approach the upper bound with $28 - 36$ active antennas, which represent less than 1% of the total number of antennas ($M = 4096$) in the RIS. This illustrates the high energy efficiency of the proposed RIS architecture and reflection beamforming solutions, as will be demonstrated in the upcoming subsection.

Additionally, to evaluate the performance at sub-6 GHz systems, we plot the achievable rates of the proposed supervised deep learning and compressive sensing solutions compared to the optimal rate $R^\star$ as illustrated in Fig. 2.9. This figure adopts the simulation setup in Section 2.9.1 at a 3.5GHz band. The RIS is assumed

to employ a UPA with 16×16 antennas, compared to 64×64 in the 28GHz band, given the path loss difference between the 3.5GHz and 28GHz bands. Each channel incorporates the strongest $L = 15$ paths, compared to $L = 10$ in the 28GHz band, motivated by the fact that the channels are less sparse in the sub-6 GHz systems compared to the mmWave systems.

Fig. 2.9 shows that the proposed supervised deep learning and compressive sensing solutions are also promising for sub-6 GHz RIS systems. This is captured by the convergence to the upper bound with only 4 active elements in the deep learning case and around 18 elements in the compressive sensing case. This figure also illustrates the gain from employing the supervised deep learning approach over the compressive sensing approach in the sub-6 GHz systems, where the channels are less sparse than mmWave systems. This gain, however, has the cost of collecting a dataset to train the deep learning model, which is not required in the compressive sensing approach.

In Fig. 2.8 and Fig. 2.9, the number of active sensors ($\overline{M}$) is a design parameter that controls the size of the input of the neural network. As that number varies, the relation between the input vector and the output target vector also varies. This suggests that the neural network architecture needs to be designed carefully to capture that relation. In Fig. 2.8, the neural network architecture used for $\overline{M} = \{1, 2, 4, 12, 20\}$ has the following number of nodes $(M, 4M, 4M, M)$. This architecture changes to $(3M, 4M, 4M, M)$ to account for the change in the input-output relation as the number of sensors increases to $\overline{M} = \{28, 36\}$. For the results in Fig. 2.9, we have found that the architecture with $\left(4\overline{M}K_{\text{SL}}, 16384, 16384, M\right)$ performs consistently well across all choices of $\overline{M}$.

Fig. 2.10 illustrates the achievable rate of both the proposed deep reinforcment learning (DRL) based solution and the supervised deep learning (SL) based solution, using 4 active elements with $L \in \{1, 15\}$ channel paths. Their performances are com-

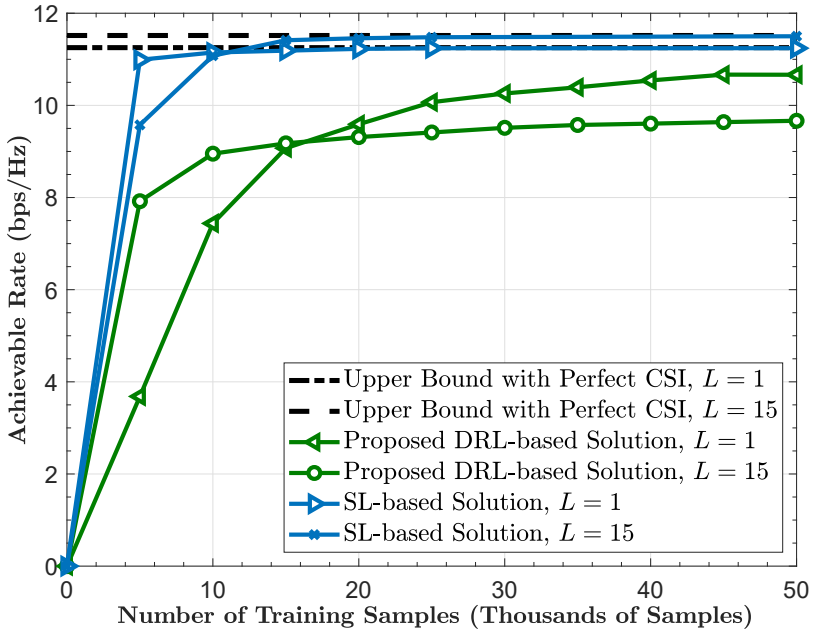


Figure 2.10: The Achievable Rates of Both the Proposed Deep Reinforcement Learning (Drl) Solution and the Supervised Deep Learning (Sl) Solution Are Compared to the Upper Bound, Using $\overline{M} = 4$ Active Elements for A 3.5GHz Scenario with $L \in \{1, 15\}$ Channel Path(/s). The Simulation Considers An RIS with A 40×10 UPA Architecture. The Upper Bound, $R^{\star}$ in (2.9), Assumes Perfect Channel Knowledge. The Figure Shows the Potential of the Proposed DRL Solution in Approaching the Optimal Rate with Almost No Beam Training Overhead and a Small Fraction of the RIS Elements to Be Active.

pared to the upper bound with perfect full channel knowledge, calculated according to (2.9). As shown, the proposed DRL solution is capable of approaching the optimal rate with more training samples than the one needed by the SL solution. **In contrast, the proposed DRL solution uses only one beam for each training episode, which constitute almost 0.3% of the beams used by the SL solution in the training phase (400 beams).** This emphasizes the efficiency of the DRL solution in operating with almost no beam training overhead.

2.9.3 Energy Efficiency

In this subsection, we evaluate the energy efficiency of both proposed CS and SL based reflection beamforming solutions, compared to the upper bound on spectral

energy efficiency, which assumes perfect *full* channel knowledge at the RIS. Starting with a formulation of a generic power consumption model for the proposed RIS architecture, we can then evaluate the energy efficiency, as formulated in [13, 69]. Consider the proposed RIS architecture shown in Fig. 2.1 and described in Section 2.5, with $\overline{M}$ active elements connected to the baseband through fully-digital architecture of b-bit ADCs. Let $P_{\mathrm{BB}}, P_{\mathrm{RFchain}}, P_{\mathrm{ADC}}, P_{\mathrm{PS}}, P_{\mathrm{LNA}}$ denote the power consumption in the baseband processor, RF chains, ADC, phase shifter (passive reflector), and LNA, respectively. The RIS power consumption model, P_{c}, can be generally formulated as [69]

$$P_{\mathrm{c}} = MP_{\mathrm{PS}} + \overline{M}(P_{\mathrm{LNA}} + P_{\mathrm{RFchain}} + 2P_{\mathrm{ADC}}) + P_{\mathrm{BB}}. \tag{2.33}$$

The power consumption of the ADC, P_{ADC}, can be further calculated as

$$P_{\mathrm{ADC}} = FOM_{\mathrm{W}} \times f_{\mathrm{S}} \times 2^{b}, \tag{2.34}$$

where b is the number of bits, f_{S} is the Nyquist sampling frequency, and FOM_{W} is the Walden's figure-of-merit for power efficiency ranking of the ADCs [70, 71]. Finally, the energy efficiency can be formulated as

$$\eta_{\mathrm{EE}} = \frac{R \times W}{P_{\mathrm{c}}} \text{bits/Joule}, \tag{2.35}$$

where W is the transmission bandwidth and R is the achievable rate.

Next, using (2.33)-(2.35), we evaluate the energy efficiency of both proposed CS and SL based reflection beamforming solutions compared to the upper bound, as depicted in Fig. 2.11. The various power consumption variables are assumed to be $P_{\mathrm{BB}} = 200$ mW, $P_{\mathrm{RF}} = 40$ mW, $P_{\mathrm{PS}} = 10$ mW, $P_{\mathrm{LNA}} = 20$ mW, and $W = 100$MHz [69]. Assume $b = 4$ bits according to the trade-off figure between the achievable rate and power consumption for fully-digital architecture, illustrated in [69]. Also, assume $FOM_{\mathrm{W}} = 46.1$ fJ/conversion-step at 100 MHz bandwidth according to the

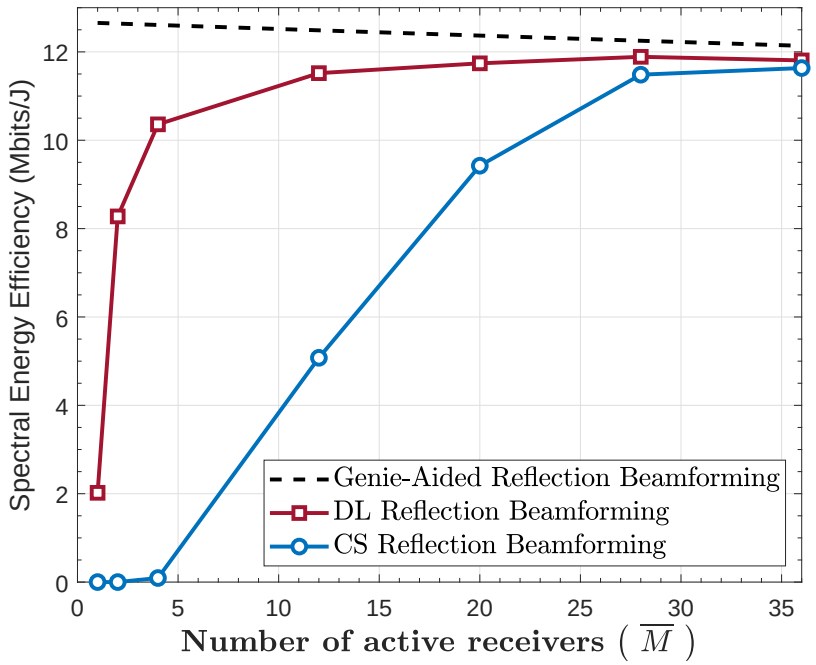


Figure 2.11: The Spectral Energy Efficiency of Both Proposed CS and SL Based Reflection Beamforming Solutions Are Compared to the Upper Bound $R^{\star}$, for Different Numbers of Active Receivers, $\overline{M}$. The Figure Is Generated at $f_{\mathrm{c}} = 28$GHz, $M = 64 \times 64$ Antennas, and $L = 10$ Paths.

architecture in [71, 72]. In Fig. 2.11, The energy efficiency values across different numbers of active channel sensors are calculated from the achievable rate values of Fig. 2.8.

Fig. 2.11 shows the high energy efficiency gained from employing the proposed RIS architecture with few active channel sensors. This figure also illustrates that both proposed CS and SL based beamforming solutions can approach the upper bound with only $28 - 36$ active antennas. The SL solution achieves more energy efficiency gains when compared to the CS solution. Also, according to (2.33)-(2.35), since the upper bound is a monotonically decreasing bound when the number of active elements increases, it's safe to state that the optimal operating point for the SL based reflection beamforming solution is at $\overline{M} = 28$ active antenna elements, with an optimal energy efficiency of ~ 12Mbits/J, for the described scenario only.

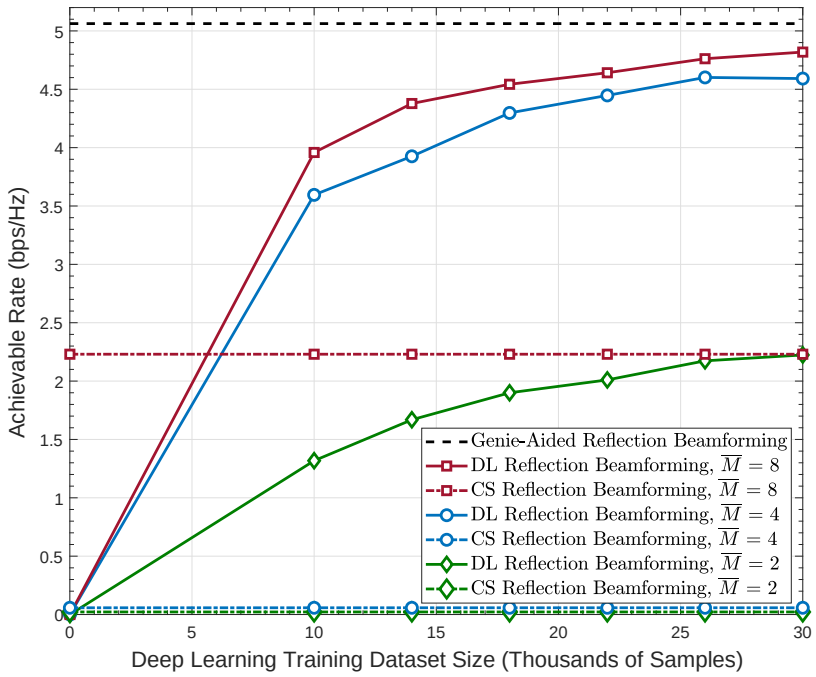


Figure 2.12: The Achievable Rate of the Proposed SL Based Reflection Beamforming Solution Is Compared to the Upper Bound $R^{\star}$ and the CS Beamforming Solution, for Different Numbers of Active Receivers, $\overline{M}$. The Adopted Setup Considers an RIS with 64×64 UPA, at 28GHz with $L = 1$ Channel Path. This Figure Highlights the Promising Gain of the Proposed Supervised Deep Learning Solution That Approaches the Upper Bound Using Only 8 Active Elements (Less than 1% of the Total Number of Antennas). This Performance Requires Collecting a Dataset of Around 20-25 Thousand Data Points (User Locations).

2.9.4 How Much Training is Needed for the Deep Learning Models?

The data samples in the deep learning dataset are captured when the receiver is randomly sampling the x-y grid. In Fig. 2.12, we study the performance of the developed deep learning approaches for designing the RIS interaction vectors for different dataset sizes. This illustrates the improvement in the machine learning prediction quality as it sees more data samples. For Fig. 2.12, we adopt the simulation setup in Section 2.9.1 with an RIS of 64×64 UPA and a number of active channel sensors $\overline{M} = 2, 4$, and 8. The setup considers a mmWave 28GHz scenario and the channels are constructed with only the strongest path, i.e., $L = 1$. Fig. 2.12 shows that with only 8 active antennas, the proposed supervised deep learning solution can achieve

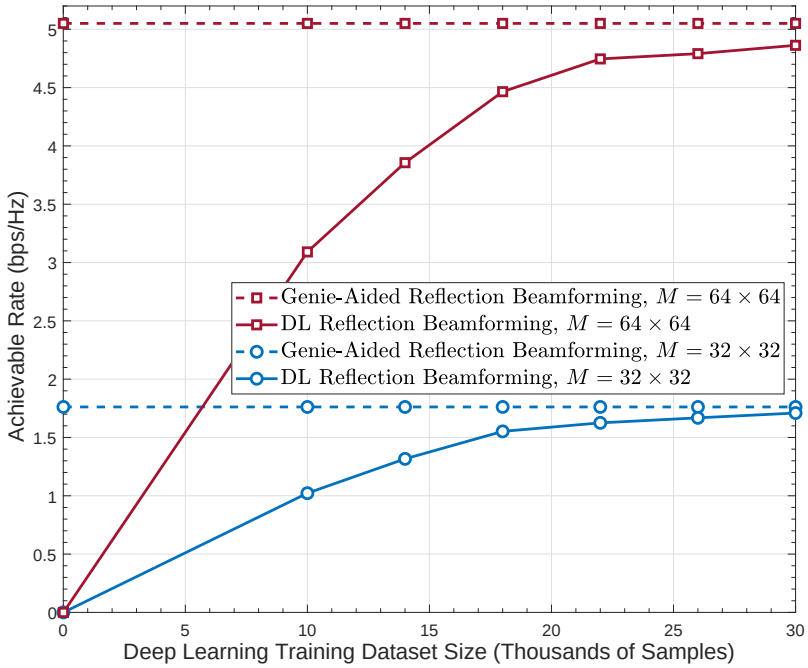


Figure 2.13: The Achievable Rate of the Proposed SL Based Reflection Beamforming Solution Is Compared to the Upper Bound $R^{\star}$ for Different Sizes of Intelligent Surfaces, Namely with RIS Of 32×32 and 64×64 UPAs. The Number of Active Elements (Channel Sensors) Equals $\overline{M} = 8$. This Figure is Generated at 28GHz with $L = 1$ Channel Path.

almost 90% of the optimal rate in (2.9) when the model is trained on 14 thousand data points (out of the 54300 points) in the x-y grid. Further, this figure highlights the performance gain of the supervised deep learning solution compared to the compressive sensing solution. This gain increases with larger dataset sizes as the compressive sensing solution does not leverage the prior channel estimation/RIS interaction observations and its performance does not depend on the size of the dataset.

2.9.5 Impact of Important System and Channel Parameters

In this subsection, we evaluate the impact of the key system and channel parameters on the performance of the supervised deep learning solution.

Number of RIS antennas: Fig. 2.13 examines the achievable rate performance of the developed solutions for designing the RIS interaction vectors when the RIS

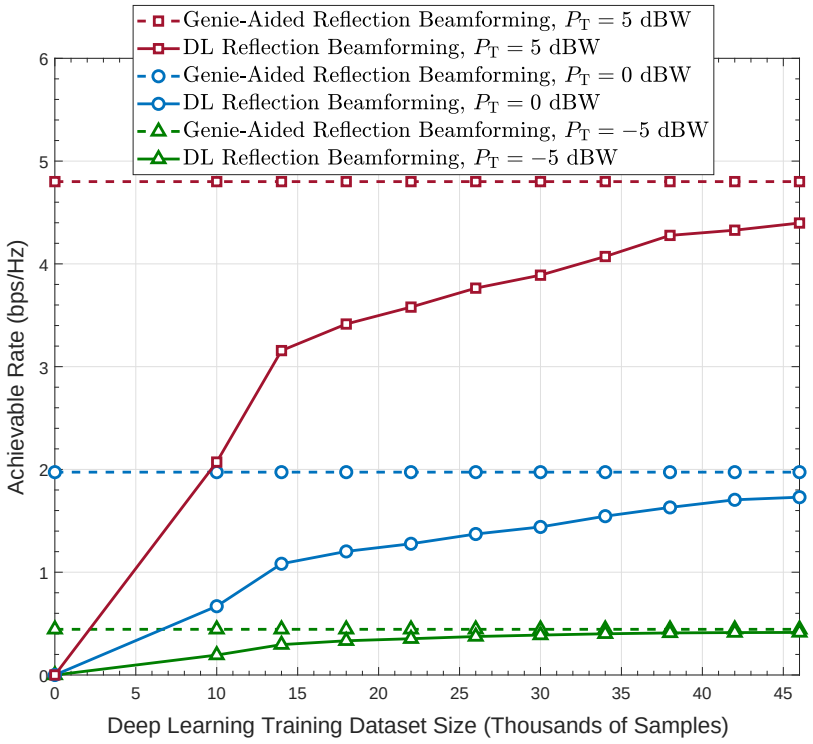


Figure 2.14: The Achievable Rate of the Proposed Supervised Deep Learning Based Reflection Beamforming Solution Is Compared to the Upper Bound $R^{\star}$, for Different Values of User Transmit Power, P_{T}. The Figure is Generated for an RIS with $M = 64 \times 64$ UPA and $\overline{M} = 8$ Active Elements, at 28GHz with $L = 1$ Channel Path. This Figure Shows That the Proposed SL Solution Is Capable of Learning and Approaching the Optimal Achievable Rate Even with a Relatively Small Transmit Power.

employs either a 32×32 or a 64×64 UPA. This figure adopts the same mmWave scenario considered in Fig. 2.12. As illustrated, with only $\overline{M} = 8$ active receivers, the proposed supervised deep learning solution approaches the optimal rate in (2.9) that assumes perfect channel knowledge for different RIS sizes. This shows the potential of the proposed RIS architecture and supervised deep learning solution in enabling reconfigurable intelligent surfaces with large numbers of antennas. **Note that the proposed solution does not require any beam training overhead (as it relies on the deep learning prediction of the best beam) and needs only 8 active receivers to realize this near-optimal performance in Fig. 2.13**.

Transmit power: In Fig. 2.14, we study the impact of the transmit power (and

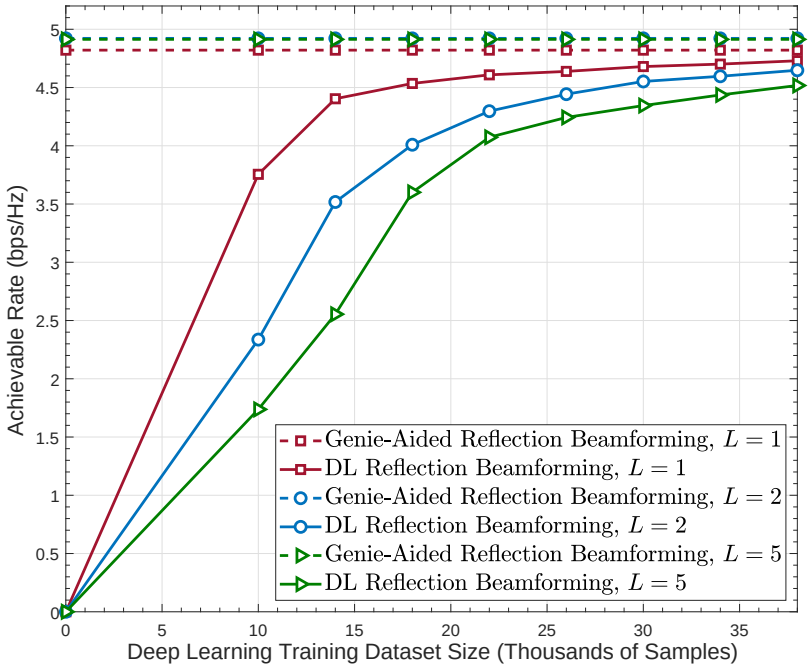


Figure 2.15: The Achievable Rate of the Proposed Sl Based Reflection Beamforming Solution Is Compared to the Upper Bound $R^{\star}$, for Different Numbers of Channel Paths, L. The Figure Is Generated for an RIS with 64×64 UPA and $\overline{M} = 4$ Active Elements, at 28GHz. As the Number of Channel Paths Increases, the Achievable Rate Achieved by the Proposed SL Solution Converges Slower to the Upper Bound. Hence, Using More Training Data Can Help Learn Multi-path Signatures.

receive SNR) on the achievable rate performance of the supervised deep learning (SL) solution. This is important in order to evaluate the robustness of the learning and prediction quality, as we input the noisy sampled channel vectors to the deep learning model. In Fig. 2.14, we plot the achievable rates of the proposed SL solution as well as the upper bound in (2.9) for three values of the transmit power, $P_{\mathrm{T}} = -5, 0, 5$ dBW. These transmit powers map to *receive* SNR values of $-3.8, 6.2, 16.2$ dB, respectively, including the RIS beamforming gain of the 4096 antennas. The rest of the setup parameters are the same as those adopted in Fig. 2.12. Fig. 2.14 illustrates that the proposed SL solution can perform well even with relatively small transmit powers and low SNR regimes.

Number of channel paths: In Fig. 2.15, we investigate the impact of the number

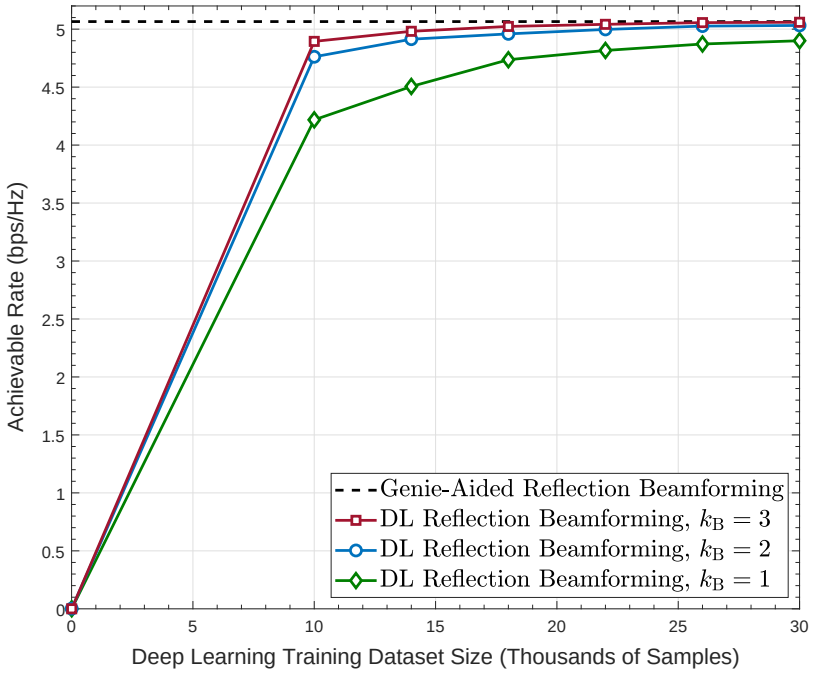


Figure 2.16: The Achievable Rate of the Proposed SL Based Reflection Beamforming Solution Is Compared to the Upper Bound $R^{\star}$. The Simulation Considers an RIS with 64×64 UPA and $\overline{M} = 4$ Active Channel Sensors, at 28GHz with $L = 1$ Channel Path. The Figure Illustrates the Achievable Rate Gain When the Beams Selected by the Deep Learning Model Are Further Refined Through Beam Training Over k_{B} Beams.

of channel paths on the performance of the developed SL solution. In other words, we examine the robustness of the proposed deep learning model with multi-path channels. For this figure, we adopt the same simulation setup of Fig. 2.12 with an RIS employing 64×64 UPA. The channels are constructed considering the strongest $L = 1, 2,$ or 5 channel paths. As shown in Fig. 2.15, with the increase in the number of channel paths, the achievable rate by the proposed SL solution converges slower to the upper bound. This shows that the proposed deep learning model can learn from multi-path channels if a large enough dataset is available.

2.9.6 Refining the Deep Learning Prediction

In Fig. 2.8-Fig. 2.15, we considered the proposed SL solution where the deep learning model uses the sampled channel vectors to predict the best beam and this

beam is directly used to reflect the transmitted data. Relying completely on the deep learning model to determine the reflection beamforming vector has the clear advantage of eliminating the beam training overhead and enabling highly mobile applications. The achievable rates using this approach, however, may be sensitive to small changes in the environment. A candidate approach for enhancing the reliability of the system is to use the machine learning model to predict the most promising k_{B} beams. These beams are then refined through beam training with the receiver to select the final beam reflection vector. Note that the most promising k_{B} beams refer to the k_{B} beams with the highest predicted rates from the deep learning model. To study the performance using this approach, we plot the achievable rate of the SL solution in Fig. 2.16, for different values of k_{B}. As this figure shows, refining the most promising k_{B} yields higher achievable rates compared to the case when the RIS relies completely on the deep learning model to predict the best beam, i.e., with k_{B}. The gain in Fig. 2.16 is expected to increase with a more time-varying and dynamic environment, which is an interesting extension in future work.

Similarly, for the DRL solution, another candidate approach for refining the DRL prediction is to use the trained DRL model in predicting the most promising k_{B} beams. Then, these beams are used for beam training to identify the best beam that will be utilized for the rest of the coherence block. Fig. 2.17 illustrates the achievable rate of the proposed DRL based solution compared to the upper bound, at different values of k_{B} $(1, 3)$, using 4 active elements. As demonstrated, the beam training of the promising k_{B} beams achieves better performance than just relying on the best network-predicted beam to reflect the incident signals. To test the effectiveness of the proposed framework, we examined another variant of the algorithm by updating its reward policy such that $R_{\mathrm{Q}}(t) = 1$ if $R(t) = R^{\star}(t)$; otherwise, $R_{\mathrm{Q}}(t) = -1$, as illustrated in Fig. 2.17. The proposed DRL solution under this ideal rewarding as-

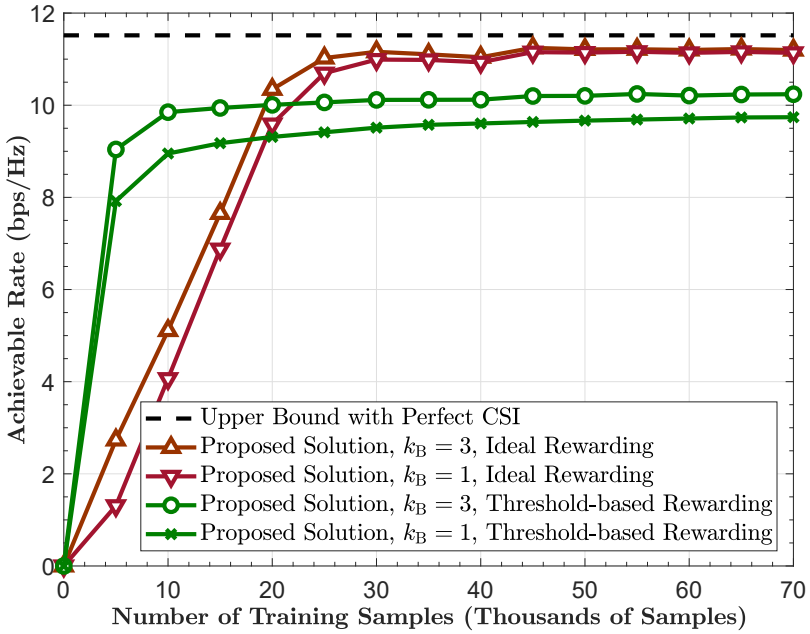


Figure 2.17: The Achievable Rate of the Proposed DRL Based Approach Is Compared to the Upper Bound $R^{\star}$. The Simulation Considers an RIS with 40×10 UPA, $\overline{M} = 4$ Active Elements, and $L = 15$ Channel Paths, at 3.5GHz. The Figure Illustrates the Achievable Rate Gain When the Beams Selected by the Deep Reinforcement Learning Model Are Further Refined Through Beam Training Over k_{B} Beams.

sumption can converge to the optimal rate. This indicates that the small gap between the performance of the proposed solution and the upper bound can be explained by the practical assumptions of using threshold-based rewards and operating in an environment with 15 channel paths. These results show the gains from exploring deep reinforcement learning frameworks to develop *standalone* RIS architectures.

2.10 Conclusion

In this chapter, we considered RIS-assisted wireless communication systems and developed efficient solutions that design the RIS interaction (reflection) matrices with negligible training overhead. We first introduced a novel RIS architecture where only a small number of the RIS elements are active (connected to the baseband). Then, we developed three solutions that design the RIS reflection matrices for this new ar-

chitecture with almost no training overhead. The first solution leverages compressive sensing tools to construct the channels at all the antenna elements from the *sampled* channels seen only at the active elements. The second approach exploits deep learning tools to learn how to predict the optimal RIS reflection matrices directly from the sampled channel knowledge, which represents what we call *environment descriptors*. Given the objective of developing standalone RIS architectures, the third approach exploits deep reinforcement learning frameworks for the RIS to learn how to predict, on its own, the optimal interaction matrices directly from the sampled channel knowledge. This solution does not require an initial dataset collection phase as opposed to the supervised learning based solutions.

Extensive simulation results based on accurate ray-tracing showed that the three proposed solutions can achieve near-optimal data rates with negligible training overhead and with a few active elements. Compared to the compressive sensing solution, the deep learning solutions requires a smaller number of active elements to approach the optimal rate, thanks to leveraging its prior observations. Further, the deep learning solutions does not require any knowledge of the RIS array geometry and does not assume sparse channels. To achieve these gains, however, the deep learning model needs to collect enough dataset, which is not needed in the compressive sensing solution. The proposed deep reinforcement learning solution, however, can converge near the optimal data rates with no dataset collection overhead and with few active elements.

For the compressive sensing solution, there are several interesting extensions, including the optimization of the sparse distribution of the active sensors leveraging tools from nested and co-prime arrays. Further, based on the expected large physical dimensions of the RIS, different parts of the RIS may observe different views of the propagation environment [36]. This motivates the need to account for non-stationary

properties in RIS performance evaluation as well as reflection beamforming design. Finally, in order to quantify the real-world performance of the RIS systems, it is neither realistic to assume a nearly stationary outdoor environment nor to consider this wide range of reflection phase shift values away from the specular reflection angle. For this reason, it is interesting to investigate the robustness of RIS systems in highly dynamic environments, under more practical reflection phase shift hardware constraints.

Chapter 3

MILLIMETER WAVE MIMO BASED SCENE DEPTH ESTIMATION

3.1 Abstract

Augmented and virtual reality systems (AR/VR) are rapidly becoming key components of the wireless landscape. For immersive AR/VR experience, these devices should be able to construct accurate depth perception of the surrounding environment. Current AR/VR devices rely heavily on using RGB-D depth cameras to achieve this goal. The performance of these depth cameras, however, has clear limitations in several scenarios, such as the cases with shiny objects, dark surfaces, and abrupt color transition among other limitations. In this chapter [1] , we propose a novel solution for AR/VR depth map construction using mmWave MIMO communication transceivers. This is motivated by the deployment of advanced mmWave communication systems in future AR/VR devices for meeting the high data rate demands and by the interesting propagation characteristics of mmWave signals. Accounting for the constraints on these systems, we develop a comprehensive framework for constructing accurate and high-resolution depth maps using mmWave systems. In this framework, we developed new sensing beamforming codebook approaches that are specific for the depth map construction objective. Using these codebooks, and leveraging tools from successive interference cancellation, we develop a joint beam processing approach that can construct high-resolution depth maps using practical mmWave antenna arrays. Extensive

[1] This chapter is based on the work published in the journal paper: A. Taha, Q. Qu, S. Alex, P. Wang, W. L. Abbott and A. Alkhateeb, "Millimeter Wave MIMO-Based Depth Maps for Wireless Virtual and Augmented Reality," in IEEE Access, vol. 9, pp. 48341-48363, 2021. This work was supervised by Prof. Ahmed Alkhateeb. Dr. Qi Qu, Dr. Sam Alex, Dr. Ping Wang, and Dr. William L. Abbott provided important ideas for the millimeter wave MIMO based system design that greatly improved the work.

simulation results highlight the potential of the proposed solution in building accurate depth maps. Further, these simulations show the promising gains of mmWave based depth perception compared to RGB-based approaches in several important use cases.

3.2 Introduction

Wireless augmented and virtual reality (AR/VR) applications are recently attracting increasing interest. Realizing wireless AR/VR in practice can open the door for a wide range of interesting applications and use cases. Enabling immersive AR/VR experience, however, requires high resolution and accurate depth perception. This can potentially allow the wireless AR/VR users to move freely within their indoor or outdoor environment. Current depth perception approaches for AR/VR systems rely mainly on RGB-D (depth) cameras for constructing the depth maps. While RGB-D based depth map construction approaches can generally provide good accuracy, they suffer from critical limitations in scenarios with bright shiny or transparent surfaces, dark objects, and large rooms among others. These limitations stem from the fundamental properties of the way visible light propagate and interact with the different surfaces.

In order to overcome these limitations, **we propose to leverage mmWave systems and signals for improving the depth map estimation accuracy**. This is motivated by the interesting characteristics of mmWave signals and by the note that mmWave systems will be deployed in future AR/VR devices anyway for meeting the wireless communication requirements [73]. In terms of the mmWave signal characteristics, the propagation of these signals is not affected by the interference from the light sources which makes mmWave systems capable of detecting bright and dark objects. Further, the mmWave diffuse scattering and specular reflection properties could help in detecting transparent objects as well as rough surfaces. These aspects among others

motivate exploring the potential of leveraging mmWave transceivers for complementing the RGB-D depth-maps in AR/VR systems, which is the focus of this chapter.

3.2.1 Prior Work

Previous depth map construction approaches focused on leveraging: (i) monocular images using RGB cameras [8], (ii) passive/active stereo images using either RGB-D depth cameras [15, 16] or infrared (IR) stereo cameras [17, 18], and (iii) gated images using active gated imaging cameras [19, 20]. In [8], a monocular depth estimation approach capable of capturing the object boundaries is proposed. In [15], RGB images along with sparse depth samples, acquired from depth cameras or computed via Simultaneous Localization and Mapping (SLAM) algorithms, are used jointly to reconstruct the depth maps. An alternative approach for depth estimation was proposed in [16], where a monocular structured-light camera — a calibrated stereo set-up with one camera and one laser projector— is leveraged for estimating the disparity. As for the active stereo systems, in [17], IR projected pattern from stereo IR cameras is adopted for depth estimation through active stereo matching. The IR images are acquired from the Intel Realsense camera [74]. Also, the IR pattern characteristics needed for active stereo matching are described in [18]. In addition, high-resolution depth images can be achieved for far objects using active gated imaging systems, as in [19, 20].

These depth map construction approaches [8, 15–20, 26], however, have several important complications as follows. (i) First, these depth map construction approaches normally fail to sense the depth for shiny, dark, transparent, and distant surfaces. While there are some attempts in solving these challenges using IR stereo cameras [17] or excessive processing of the RGB-D images [27], there is no complete and general solution yet to this problem. (ii) Further, these IR and RGB-D based depth map

construction algorithms suffer from a critical limitation, which is the depth ambiguity for faraway objects/surfaces. The depths for distant surfaces can not be resolved by the algorithms in [17, 27]. (iii) Another key challenge is the additional bill of materials (BOM) cost incurred from integrating the IR stereo camera systems into the wireless AR/VR device architectures. On the contrary, the existing mmWave systems in the wireless AR/VR device architectures incurs no additional BOM cost when leveraged for depth map estimation purposes jointly with the primary purpose of wireless communications. (iv) The field of view coverage is also a main challenge. The depth map coverage is limited by the camera field of view. The camera field of view is constrained by the camera lens and by the light sensor. The field of view in mmWave MIMO systems, however, is constrained by the array radiation pattern, as will be explained in Section 3.7. By contrast, the typical field of view in mmWave MIMO systems can be larger than the typical camera field of view.

These challenges motivate the research for other technologies to complement the RGB-D cameras in accurately sensing the VR/AR environment. One promising technology for this goal is employing wireless millimeter wave (mmWave) systems. Since mmWave antenna arrays will be used to satisfy the communication high data rate demands of wireless VR/AR, it is interesting to investigate if they could also be useful for VR/AR-relevant sensing functions, such as depth estimation. Initial studies for using mmWave communication arrays for radar and sensing were presented in [75, 76]. These studies, however, focused only on the ranging problem (of one or multiple targets), not on the depth map construction problem. Other mmWave sensing and tracking work that was not restricted to communications hardware was presented in [77, 78]. The research in [77, 78], though, targeted tracking a single object in a small distance, and cannot be directly applied to depth estimation of surrounding surfaces in VR/AR. Further, the work in [75–78], did not study the trade-

offs between estimation accuracy and different system parameters, such as number of antennas and adopted bandwidth, and did not compare between the system performance under transceiver architectures constraints, such as those imposed on the analog phased-array transceiver architectures. By contrast, interesting research challenges are accompanying the mmWave MIMO based scene depth map construction framework ranging from beam codebook design challenges to scene depth estimation challenges. These challenges will be addressed in this work and will be explained in detail in Section 3.6.

3.2.2 Contribution

In this chapter, we consider the mmWave MIMO based depth map construction problem for AR/VR systems, adopting mmWave communication hardware and frame structure. The contributions of this chapter can be summarized as follows.

- *mmWave MIMO depth map construction framework:* We formulate the mmWave MIMO depth map construction problem and propose a general framework for building depth maps under the constraints imposed by mmWave communication hardware and frame structure.

- *A design for depth-map suitable sensing beamforming codebook:* We define the characteristics of the desirable mmWave sensing beamforming codebook for efficient depth map construction and develop a codebook construction approach that meets these characteristics.

- *High-resolution depth map construction approach:* Given the designed beamforming codebook, we develop a novel signal processing approach for jointly processing the signals received by the sensing beams and building high-resolution depth maps.

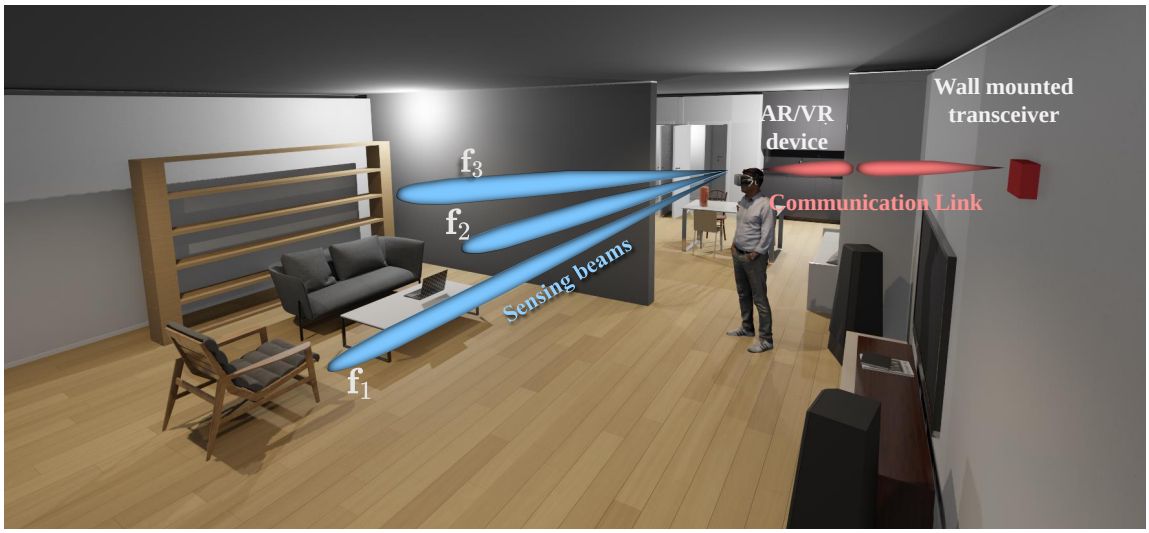


Figure 3.1: The Considered Setup Where the mmWave Communication System, Deployed at the AR/VR Device, Is Jointly Leveraged for Sensing and Depth Map Construction. This Figure Is Generated Using Blender [3] with 3D Models Downloaded from [4–7].

The proposed solution is extensively evaluated using accurate ray-tracing channels generated from Wireless InSite [1], and ground truth depth images generated from Blender [3]. The simulation results show the promise of mmWave MIMO sensing in becoming a viable depth estimation solution for communication-constrained sensing systems, either as a standalone approach or as an integrated approach with RGB-D depth cameras. These simulation results can be of great usefulness for various applications; they can be generally applied to AR/VR devices, smart home devices, or auto drive devices.

3.3 System and Channel Models

In this section, the system model for the adopted communication-constrained sensing framework is first formulated, followed by the characterization of the adopted channel model.

3.3.1 System Model

In this work, we propose to reuse the same AR/VR mmWave communication system/circuits to do the sensing and depth map construction, as shown in Fig. 3.1. Hence, we adopt a sensing model that accounts for the mmWave communication system/circuit constraints. This communication-constrained sensing model consists of a transmitter and a receiver; both are connected through a self-isolation circuitry to a shared N antenna array, as depicted in Fig. 3.2. This type of operation is commonly referred to as MIMO in-band full-duplex operation [79]. We assume that the transmitter and receiver chains are well-isolated by an isolation circuitry to avoid any self-interference. This assumption is reasonable with the recent developments of self-interference systems. One example of these systems is the magnetic-free non-reciprocal circulators (i) based on coupled-resonator loops [80] or (ii) based on CMOS circulators operating in the 28GHz mmWave band [81]. Another example is the receiver with integrated magnetic-free non-reciprocal circulator and baseband self-interference cancellation operating in the Sub-6 GHz band [82]. A third example is the magnetic-free SOI CMOS circulator operating in the 60GHz mmWave band [83]. Accounting for this self-interference, however, is an important direction for future extensions.

Further, and for the sake of having low-cost and power consumption mmWave transceivers, we adopt an analog-only architecture for the N-antenna array used for transmission and reception, [24, 25], where the beamforming/combining is done in the analog domain using a network of phase shifters. Next, we summarize the transmit and receive signal models.

Transmit Signal Model: We consider a wideband single-carrier waveform comprising multiple time frames. These frames are transmitted over an aggregated time

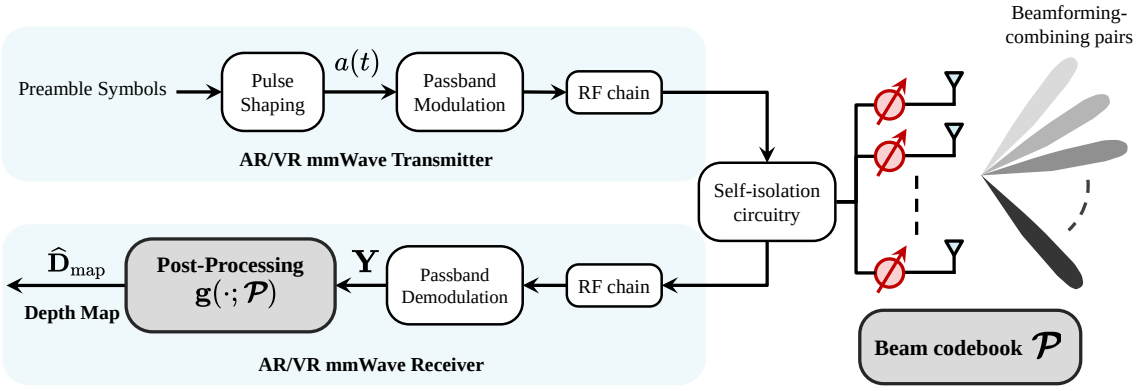


Figure 3.2: A Block Diagram of the Communication-Constrained Sensing Model Is Illustrated. The Sensing Framework, $\boldsymbol{\Pi}$, Consists of (a) the Beam Codebook Design $\boldsymbol{\mathcal{P}}$ and (b) the Post-Processing Design $\mathbf{g}\,(.,\boldsymbol{\mathcal{P}})$, to Estimate the Scene Depth Map $\widehat{\mathcal{D}}$. The Upper Path Represents the Transmitter Path, While the Lower Path Represents the Receiver Path.

interval of T seconds during which the environment is assumed to be relatively static. This time interval is commonly referred to as a coherent processing interval (CPI) [84]. Each frame consists of both data and preamble sequences designed for the wireless communication function. The co-existing sensing model also uses these preamble sequences to sense the environment and build the depth maps, as will be explained in detail in the following sections. This can be achieved by either splitting the frames between sensing and communication or by designing the sensing and communication beam training operations to share the same preamble sequences. Next, for ease of exposition, we assume that M frames/preamble sequences are dedicated for sensing. If $s_m[n]$ denotes the n^{th} transmitted symbol at the m^{th} frame, with $\mathbb{E}\big[\,|s_m[n]|^2\,\big] = 1$, then the complex-baseband representation of the transmit waveform can be written as [85]

$$a(t) = \sqrt{\mathcal{E}_{\text{s}}} \sum_{m=0}^{M-1} \sum_{n=0}^{N_m-1} s_m[n]\ \delta(t - nT_{\text{S}} - mT_{\text{F}}), \tag{3.1}$$

where $\mathcal{E}_{\text{s}}$ represents the average energy per symbol, T_{S} is the symbol time, and T_{F} is the frame duration. N_m is the number of symbols in the m^{th} frame, which is divided into a preamble sequence of length N^{p} and a set of data symbols of length

N_m^{d}. Further, we assume that the same preamble sequence $s^{\mathrm{p}}[n], n \in \{1, \ldots, N^{\mathrm{p}}\}$, is transmitted in the first N^{p} symbols of each frame. Note that for the sake of simplifying the transmit and receive signal representation, we incorporated the transmit pulse shaping and receive filtering functions into the channel model. Finally, if a beamforming vector $\mathbf{f} \in \mathbb{C}^{N \times 1}$ is used to transmit the signal at the AR/VR device, the complex-baseband representation of the transmitted signal can expressed as

$$\mathbf{x}(t) = \mathbf{f} a(t). \tag{3.2}$$

This transmitted signal will interact with the environment (through reflection, scattering, etc.) and will be received back by the AR/VR device. Next, we describe the receive signal model.

Receive Signal Model: Let $\mathsf{G}_{\mathrm{tar}}$ denote the number of targets/scatterers in the environment. Then, focusing on the preamble sequence transmission/reception (i.e., the first N^{p} symbols of each frame), the receive *sensing* signal of the m^{th} frame can be written as

$$y_m[n] = \sum_{g=1}^{\mathsf{G}_{\mathrm{tar}}} \sum_{d=0}^{L_{\mathrm{d}}-1} \sqrt{\mathcal{E}_{\mathrm{s}}} \mathbf{w}^H \mathbf{H}_{d,g} \mathbf{f} s^{\mathrm{p}}[n-d] + \mathbf{w}^H \mathbf{v}_m[n], \tag{3.3}$$

where $\mathbf{w} \in \mathbb{C}^{N \times 1}$ is the combining vector at the AR/VR, and $\mathbf{v}_m[n] \sim \mathcal{N}_{\mathbb{C}}\left(\underline{0}, \sigma_{\mathrm{n}}^2 \mathbf{I}\right)$ is the receive noise with variance σ_{n}^2. $\mathbf{H}_{d,g} \in \mathbb{C}^{N \times N}, d \in \{1, \ldots, L_{\mathsf{d}} - 1\}$, is the delay-$d$ channel matrix between the transmission from and the reception by the AR/VR antenna array, which is described in the following subsection.

3.3.2 Channel Model

Given that the depth sensing problem highly relies on the accurate modeling of the surrounding environment and its geometry, we adopt a geometric channel model in this work. More specifically, we consider the extended Saleh-Valenzuela wideband

geometric channel model [86–89]. Based on that, the g^{th} target contribution in the delay-d channel, $\mathbf{H}_{d,g}$, can be modeled as

$$\mathbf{H}_{d,g}=\sqrt{G_g}\sum_{\ell=1}^{L_{\text{ray}}}\Big[\alpha_\ell \mathrm{e}^{-\jmath 2\pi f_{\text{c}}\tau_\ell} p\left(dT_{\text{s}}-\tau_\ell\right)\times \mathbf{a}_{\text{R}}\left(\phi_{\ell,g}^{\text{R}},\theta_{\ell,g}^{\text{R}}\right)\mathbf{a}_{\text{T}}^{H}\left(\phi_{\ell,g}^{\text{T}},\theta_{\ell,g}^{\text{T}}\right)\Big], \tag{3.4}$$

where L_{ray} is the number of channel clusters; each cluster is contributing with one ray of complex channel coefficient α_ℓ, time delay τ_ℓ, and azimuth/elevation angles of departure and arrival, $\phi_{\ell,g}^{\text{T}},\theta_{\ell,g}^{\text{T}}$ and $\phi_{\ell,g}^{\text{R}},\theta_{\ell,g}^{\text{R}}$, respectively. $\mathbf{a}_{\text{T}}(\cdot,\cdot)$ and $\mathbf{a}_{\text{R}}(\cdot,\cdot)$ represent the transmit and receive array response vectors associated with the angles of departure and arrival respectively. The transmit and receive pulse shaping signals are included within $p(t)$ such that $p(t) = p_{\text{T}}(t) * p_{\text{R}}(t)$. The path gain associated with the g^{th} target is denoted by G_g and can be expressed as

$$G_g = \frac{G_{\text{T}}G_{\text{R}}\lambda^2\sigma_g^{\text{RCS}}}{(4\pi)^3(\rho_g)^{2\text{PL}}}, \tag{3.5}$$

where G_{T} and G_{R} are the transmitter and receiver gains, λ is the operating wavelength, PL is the path loss exponent. Finally, ρ_g denotes the distance (range) between the AR/VR device and the g^{th} target/scatterer and σ_g^{RCS} denotes the radar cross section of this target.

3.4 Problem Definition

Our objective in this work is to efficiently estimate the depth/range map of the surrounding environment using the communication-constrained mmWave MIMO sensing model in Section 3.3. Before delving into the formal problem definition, it is important to distinguish between the range and the depth of a certain target. As depicted in Fig. 3.3, the range of a target with respect to the AR/VR camera (which is aligned with the AR/VR antenna array) is the linear radial distance from the camera center

(focal point) to the target. For the depth, it is measured by the y-coordinate of the camera center (focal point) with respect to the x-z plane of the target. Given that the range and depth can be calculated from one another, we focus our formulation on the depth estimation problem. Next, we define the depth map of the surrounding environment with respect to an AR/VR device.

Definition (Depth Map): *We define the depth map* $\mathbf{D}_{\text{map}}$ *of resolution* $M_h \times M_w$ *as an image of* M_h *pixels high and* M_w *pixels wide, where the value of each pixel represents the smallest depth between the AR/VR device and the targets/objects in this pixel.*

In this work, we express this depth map as an $M_h \times M_w$ matrix $\mathbf{D}_{\text{map}} \in \mathbb{R}^{M_h \times M_w}$. Further, we use $M_{\text{res}} = M_h M_w$ to denote the total number of pixels in the depth map. The range map $\mathbf{R}_{\text{map}} \in \mathbb{R}^{M_h \times M_w}$ is similarly defined. Now, given the system and channel models in Section 3.3, the AR/VR device constructs the estimated mmWave-based depth map through two main steps: (i) sensing the environment using several beamforming and combining sensing vectors and (ii) post-processing the receive sensing signal to construct the estimated depth map. More formally, if a beamforming-combining pair $(\mathbf{f}_m, \mathbf{w}_m)$ is used to transmit and receive the N^{p} symbols of the m^{th} preamble sequence, then the receive sensing signal can be expressed as

$$y_m[n] = \sum_{d=0}^{L_{\text{d}}-1} \sqrt{\mathcal{E}_{\text{s}}} \mathbf{w}_m^H \mathbf{H}_d \mathbf{f}_m s^{\text{p}}[n-d] + \mathbf{w}^H \mathbf{v}_m[n], \tag{3.6}$$

where $n \in \{0, 1, \ldots, N^{\text{p}} + L_{\text{d}} - 1\}$, $\mathbf{H}_d = \sum_{g=1}^{\mathsf{G}_{\text{tar}}} \mathbf{H}_{d,g}$. By stacking the $N^{\text{p}} + L_{\text{d}}$ receive symbols (from transmitting the preamble sequence), we get $\mathbf{y}_m = \big[y_m[0], \ldots, y_m[N^{\text{p}} + L_{\text{d}} - 1]\big]^T$, which represents the receive sensing vector of one preamble sequence using one beamforming-combining pair. Next, if M preamble sequences are used to sense the environment via M beamforming-combining pairs $(\mathbf{f}_m, \mathbf{w}_m), m \in \{1, \ldots, M\}$,

then the aggregated receive sensing signal can be written as

$$\mathbf{Y} = [\mathbf{y}_1, \ldots, \mathbf{y}_M]. \tag{3.7}$$

For ease of exposition, we define the sensing beamforming codebook $\boldsymbol{\mathcal{P}}$ as the codebook that includes the M beamforming-combining pairs, i.e., $\boldsymbol{\mathcal{P}} = \{(\mathbf{f}_m, \mathbf{w}_m) : m \in \{1, \ldots, M\}\}$. Finally, given the receive sensing matrix $\mathbf{Y}$, a post-processing is applied for estimating the depth map. If $\mathbf{g}(.)$ denotes the post-processing function, the estimated depth map $\widehat{\mathbf{D}}_{\text{map}} \in \mathbb{R}^{M_h \times M_w}$ can be written as

$$\widehat{\mathbf{D}}_{\text{map}} = \mathbf{g}(\mathbf{Y}; \boldsymbol{\mathcal{P}}). \tag{3.8}$$

Our objective in this work then is to design the sensing beamforming codebook $\boldsymbol{\mathcal{P}}$ and the post-processing $\mathbf{g}(\cdot)$ to efficiently estimate the depth map $\widehat{\mathbf{D}}_{\text{map}}$ to be as close as possible to the actual depth map $\mathbf{D}_{\text{map}}$. To evaluate the performance of the proposed approaches, we will adopt the root-mean squared estimation error (RMSE) and the mean absolute error (MAE) between the depth maps, which are defined as

$$\Delta_{\text{RMSE}} = \sqrt{\frac{1}{M} \left\| \mathbf{D}_{\text{map}} - \mathbf{g}\left(\mathbf{Y}; \boldsymbol{\mathcal{P}}\right) \right\|_2^2}, \tag{3.9}$$

$$\Delta_{\text{MAE}} = \frac{1}{M} \left\| \mathbf{D}_{\text{map}} - \mathbf{g}\left(\mathbf{Y}; \boldsymbol{\mathcal{P}}\right) \right\|_1^2. \tag{3.10}$$

In Section 3.6, we will present the general framework of our proposed depth map estimation approach. This will be followed by a detailed description of the two main components in this framework, namely the beamforming codebook design $\boldsymbol{\mathcal{P}}$ and the post-processing solution $\mathbf{g}(.)$, in Sections 3.7 and 3.8.

3.5 Background

Before going into the proposed framework for estimating depth/range maps using mmWave MIMO, we provide a brief background on the basis of the single-target

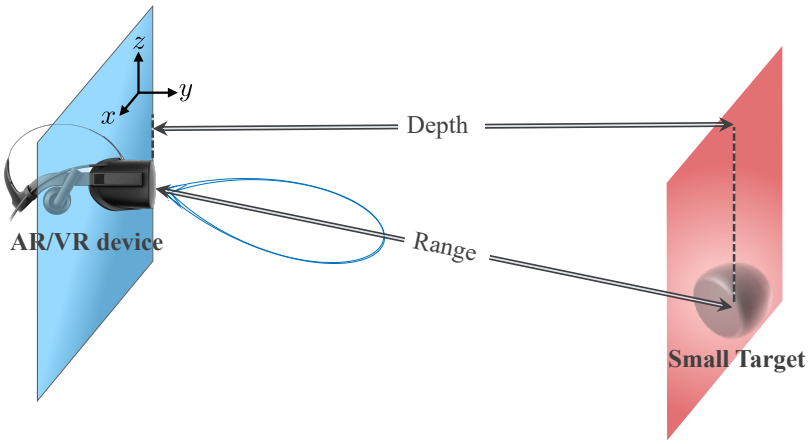


Figure 3.3: This Figure Shows the Conventional Single Target Range Estimation Problem, Where One Target Exists in Free Space in Line-of-sight (LoS) with the AR/VR Device. This Device Steers Perfectly One Beam Towards That Target to Estimate the Range.

range estimation problem. For a preliminary model, consider one target in the free space with a Line-of-sight (LoS) path to the AR/VR device. Further, consider the case when one mmWave beam is perfectly steered towards that target, as depicted in Fig. 3.3. Adopting this preliminary model, the target range estimation accuracy bound will be first examined. Then, a description of the main algorithms used in the literature to approach this problem is provided.

3.5.1 Target Range Estimation Accuracy

Our main objective is to find the fundamental limit for mmWave MIMO based depth estimation, which can be considered as range estimation at every possible eyesight direction, i.e. at every azimuth angle $\phi \in [0, 2\pi[$ and every elevation angle $\theta \in [0, \pi[$. For the range estimation accuracy, one useful metric is the Cramer-Rao lower bound (CRLB) on the range estimation. For white Gaussian noise, the CRLB provides a lower bound on the mean-squared-error of any unbiased estimator, hence it is used as a benchmark for the performance analysis of parameter estimation[90]. Considering the case of range estimation for a single target, the CRLB of this single

target range is formulated as [76, 84, 90]

$$\sigma_{\hat{\rho}}^2 \geq \frac{\varsigma^2}{8P_{\text{int}}\ \eta^2 B^2\ \text{SNR}_{\text{rad}}}, \tag{3.11}$$

where ς is the speed of light, B is the transmission bandwidth, P_{int} is the integration gain and is equal to the number of symbols used for preamble estimation, and η depends on the power spectral density shape of $a(t)$ over the preamble duration. Under the assumption of a flat spectral density for $a(t)$, $\eta^2 = (2\pi)^2/12$. The radar signal-to-noise ratio for this target can then be expressed as $\text{SNR}_{\text{rad}} = \mathcal{E}_{\text{s}} G_{\text{rad}}/\sigma_{\text{n}}^2$, where G_{rad} denotes the path gain associated with the target.

3.5.2 Target Range Estimation Algorithms

Estimating the round trip delay $\hat{\tau}$ is equivalent to finding the range estimate $\hat{\rho}$, since they are directly related through $\hat{\tau} = 2\hat{\rho}/\varsigma$. Given the extensive research on delay estimators in the literature [84, 91], we will restrict the scope of this work on the magnitude based delay estimators in [92] for simplicity. In a general sense, given a known transmit preamble sequence $x_0[n]$ and the received baseband sequence $z[n]$, the receiver can estimate the round-trip delay by maximizing an objective function, the cross-correlation function between the two time-sequences, over a range of possible delays. Based on this notion, two delay estimators are formulated as follows [92].

Basic Correlator

The basic correlator is a coarse delay estimator that performs the maximization at the same sampling frequency, f_{S}, tuned by the AR/VR communication system. Assume that the length of the received baseband sequence, $z[n]$, is L_{z} samples, where the last N_{z} samples are non-zeros. The range estimate can then be formulated as

$$\hat{\rho}^{\text{BC}} = \frac{\varsigma T_{\text{S}}}{2} \underset{q:q\in\mathcal{Q}}{\arg\max} \left| \sum_{n=L_{\text{z}}-N_{\text{z}}}^{L_{\text{z}}-1} x_0[n] \times z^*[n-q] \right|^2. \tag{3.12}$$

where $T_{\mathrm{S}} = \frac{1}{f_{\mathrm{S}}} = \frac{1}{B}$ denotes the sampling time, $\mathcal{Q}$ represents the set of possible discrete sample delays, and the optimal q solution is denoted by q^{BC}. Unfortunately, the accuracy of this range estimate is limited by the sampling frequency f_{S}. One attempt of improving the estimation accuracy is by performing the maximization at a higher sampling frequency. This attempt, however, increases the computational complexity dramatically, which motivates the role of the upcoming delay estimator, the massive correlator [92].

Massive Correlator

The primary function of the massive correlator is to perform the maximization of the objective function at a higher sampling frequency without the computational burden of computing the shift in real time. For this reason, [92, 93] introduced the solution of pre-designing a specific correlator bank that contains shifted versions of the reference sequence, $x_0[n]$. The receiver will then multiply the received sequence by the correlator bank to compute the objective function.

We describe the steps of the massive correlator algorithm as follows [92, 93]. (a) Upsample $x_0[n]$ with a sampling frequency higher than f_{S}, denoted as f_{est}. (b) Define the correlator bank matrix, $\mathbf{X}_0$, where each row of this matrix is a shifted version of the upsampled $x_0[n]$. Let the number of rows in $\mathbf{X}_0$ be equal to $(2\delta + 1)$, where δ is the largest lag/advance discrete fractional delay in the receive sequence, such that $\delta = \frac{f_{\mathrm{est}}}{2f_{\mathrm{S}}}$. (c) Downsample independently each row of the correlator matrix to the lower sampling frequency f_{S}; let the resulting matrix be named as $\mathbf{B}_0$. The reason for this step is to test delays at the higher sampling frequency, f_{est}, but only apply multiplications at the lower sampling frequency, f_{S}. (d) Shift back the receive sequence, $z[n]$, with the coarse discrete delay estimate, q^{BC}, such that $\bar{z}[n] = z[n+q^{\mathrm{BC}}]$, and then concatenate the sequence into one row vector, $\bar{\mathbf{z}}$. (e) Calculate the fractional range estimate, $\hat{\rho}'$,

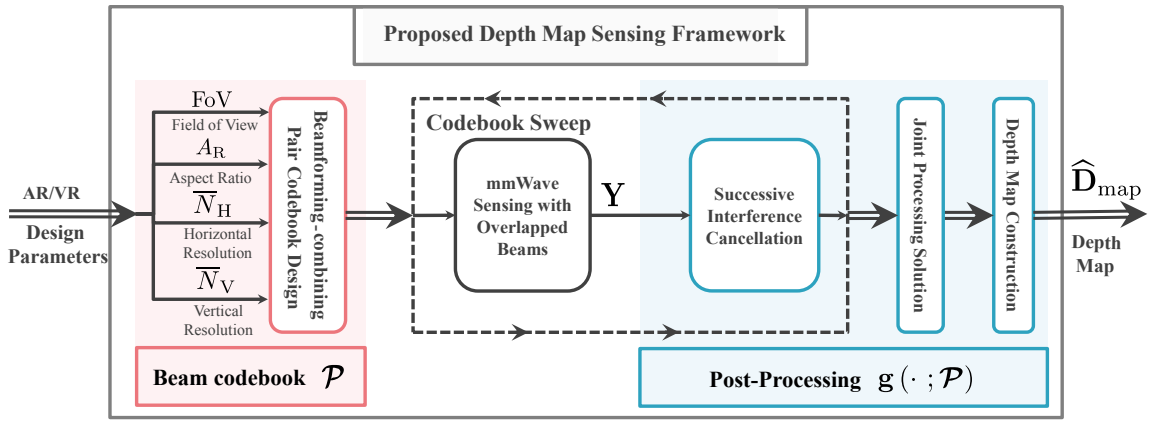


Figure 3.4: The Figure Summarizes the Proposed Sensing Framework for mmWave MIMO Based Depth Estimation, Which Involves Sensing the Scene Using the Designed Beamforming Codebook $\boldsymbol{\mathcal{P}}$ and Applying the Proposed Post-Processing Operations $\mathbf{g}(.;\boldsymbol{\mathcal{P}})$ to The Receive Signal to Construct the Estimated Depth Map $\widehat{\mathbf{D}}_{\text{map}}$.

such that $\hat{\rho}' = \frac{c}{2f_{\text{est}}}\left(-(\delta+1) + \arg\max_{q'} [\mathbf{g}]_{q'}\right)$, where $\mathbf{g} = \bar{\mathbf{z}} \times \mathbf{B}_0^H$. (f) Calculate the fine range estimate, $\hat{\rho}^{\text{MC}}$, such that $\hat{\rho}^{\text{MC}} = \hat{\rho}^{\text{BC}} + \hat{\rho}'$.

With an end goal of constructing depth maps, these range estimation algorithms will then be leveraged by the mmWave MIMO based depth estimation as explained in the upcoming sections. In the next section, we formulate a general framework for scene depth estimation.

3.6 General Framework for Scene Depth Estimation

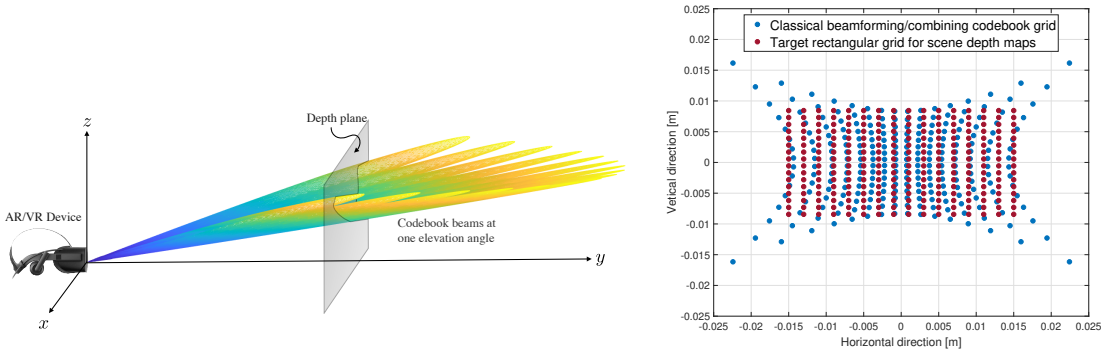
In this section, we highlight the key elements of the proposed depth map estimation approach, namely the sensing beamforming codebook $\boldsymbol{\mathcal{P}}$ and post-processing $\mathbf{g}(.)$, and discuss the challenges associated with designing these elements. As depicted in Fig. 3.4, we first design the sensing beamforming codebook $\boldsymbol{\mathcal{P}}$ offline based on the desired AR/VR properties such as the field of view, the scene aspect ratio, and the number of horizontal and vertical beams covering the scene view. To build the depth map of a certain scene, the beam pairs of the designed codebook are used to sense the environment and acquire the receive sensing matrix $\mathbf{Y}$ in (3.6). This receive signals

are then jointly processed using the post-proposed approach to build the depth map. In the remaining of this section, we explain the challenges associated with designing the codebook and the post-processing operations. Then, we will present how our proposed solutions overcome these challenges in Sections 3.7 and 3.8.

3.6.1 Codebook Design Challenges

To effectively sense the surrounding environment and build efficient depth maps, the beams of the sensing codebooks should be designed to scan the full scene. Since the mmWave MIMO based depth maps may potentially complement the RGB-D based maps, our objective is to build a beamforming codebook that scans the full rectangular grid of the typical depth sensors of the AR/VR cameras. However, the classical beam steering codebooks such as the DFT codebooks [94], that independently sample the azimuth and elevation directions, do not normally fit a rectangular grid, as shown in Fig. 3.5(b). They instead form a parabolic grid, i.e., for a fixed elevation angle, the grid line of these codebook beams are parabolic curves as shown in Fig. 3.5(a). This mismatch between the mmWave MIMO-based and camera-based depth grids could lead to clear distortion in the joint mmWave/RGB-D depth map construction and make it hard to complement the RGB-D depth map using mmWave MIMO sensing.

One possible solution is to estimate the depths on the parabolic grid using the classical beamforming codebook and then interpolate/extrapolate to calculate the rectangular depth map. The main disadvantage of this solution, however, is that the interpolation can potentially lead to considerable loss in the depth map accuracy as the changes of the depth are not normally smooth in nature. Hence, in order to avoid the interpolation loss, the more persuasive solution is to develop a depth map compatible beamforming codebook that fits exactly the desirable rectangular sensor grid. With this motivation, we propose a beamforming/combining design approach



(a) Parabolic Shape of the Classical Codebook Grid (b) Classical Codebook Grid Mismatch

Figure 3.5: (a) The Intersections Between the Classical Codebook Beam Directions and the x-z Depth Plane Form the Parabolic Shape of the Classical Codebook Grid. (b) The Mismatch Between the Classical Codebook Grid of a 16×16 UPA and the Desirable Rectangular Grid for a Depth Map Is Illustrated at a $y = 13.32$mm Depth Plane, for a Scene of 100° Field of View and $16/9$ Aspect Ratio.

in Section 3.7 to overcome the codebook mismatch challenge.

3.6.2 Scene Depth Estimation Challenges

The sensing beamforming codebook is used to sense the surrounding environment. Now, given the receive sensing matrix $\mathbf{Y}$, the objective of the post-processing is to construct an accurate depth map of the facing scene. This process, however, has several challenges. In order to explain these challenges, let's first consider the case when the environment has only a single target. In this case, the sensing/scanning beam that is directed towards the region that includes this target will result in some backscattering signal. This signal can be used for calculating the round-trip time of flight and consequently the range of this target, leveraging the MIMO radar concepts [84, 95] and the algorithms detailed in Section 3.5.2. In terms of the range/depth map, the pixel that includes the region of this target will simply have the value of the estimated range/depth. In practice, however, the environment has several targets/surfaces and the mmWave arrays have strict constraints on their hardware: power budget, compu-

tational complexity, etc. These limitations lead to critical challenges for our objective of building accurate depth and range maps of the environment. More specifically, if we adopt the approach that scans the surrounding scene using a beamforming codebook and processed the receive sensing signal of each beam independently to estimate the depth of the region defined by this beam, then this approach will have the following key drawbacks.

- **Low-resolution depth-maps:** The low resolution drawback is mainly due to (i) the limitation on the number of AR/VR antennas, which is controlled by many factors in the AR/VR device such as the device dimensions, computational complexity, circuit routing, power consumption, etc., and (ii) the number of beams in the sensing codebook $\boldsymbol{\mathcal{P}}$, which is limited by the time allocated for the depth estimation process.

- **Inter-target interference:** The constraints on the number of antennas at the AR/VR device limit the system spatial resolution. This makes it hard to differentiate between the ranges/depths of the different targets/surfaces that are close to each other. In other words, when measuring the depth of the object in a particular region/ direction, multiple objects/surfaces may reflect the incident signal at the same time. The interference between these reflected/scattered signals may highly affect the accuracy of the range/ depth estimation. Hence, if a certain pixel has multiple objects/surfaces, it will be difficult to estimate the shortest depth of the objects in this pixel (to follow the depth map definition in Section 3.4).

- **Inter-path interference:** When sensing the range/depth of a certain target, the optimal situation (in terms of depth estimation accuracy) is when the target backscatters a single ray to the receiving array. In practice, however, the signal

incident on a certain target may experience more than one phenomenon, such as scattering, reflection, diffraction, etc., which results in multiple rays. More than one of these rays could traverse the environment in different ways/directions, especially in indoor environments, before reaching the receiver. This means that they may reach the receiving array from multiple angles and with different time of flights. This causes an inter-path interference which makes it hard to accurately estimate the range/depth of the target of interest. For example, if the receiver estimates the range/depth based on a wrong path, this may noticeably degrade the accuracy of the depth map estimation. This challenge is depicted in Fig. 3.6. As illustrated, the challenge is how to design the sensing framework (the codebook and post-processing) to detect the desired channel path (the path in blue) while filtering out all the undesired channel paths. Examples of undesired paths are the paths 2-4. Path 2 is transmitted and received within the main lobe. Path 3 is transmitted and received within the side lobe. Path 4 experiences multiple reflections instead of back-scattering, before reaching back the receiver. It has to be noted that the diffuse scattering and specular reflection properties of the mmWave signals are still crucial for constructing depth maps despite their contribution to the inter-path interference challenge. Without these properties, the sensing framework may not be able to construct a meaningful depth scene of the surrounding environment.

In the next two sections, we efficiently design the two elements of our proposed depth map sensing framework, namely the sensing codebook and the post-processing, to address these challenges.

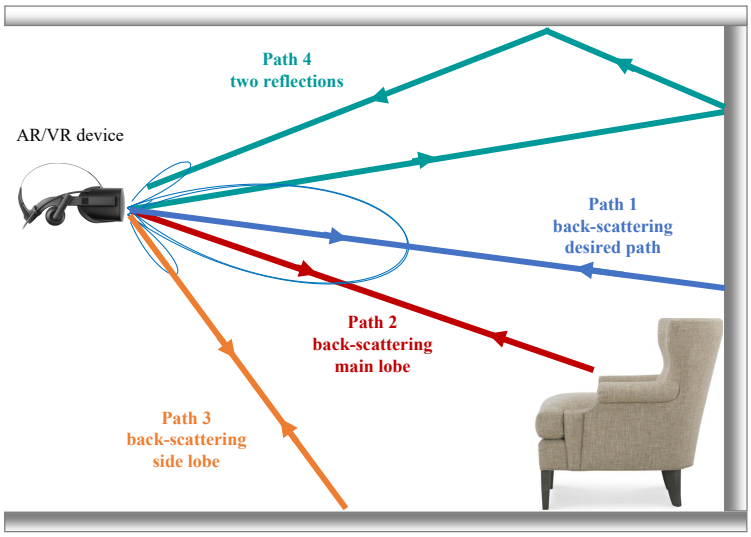


Figure 3.6: The Multipath Estimation Challenge for Scene Range Estimation Is Illustrated. The Design Challenge Is How the Sensing Framework Can Detect and Estimate the Range Through the Desired Channel Path (Path 1 in Blue) and Avoid Making Faulty Estimation Because of the Other Undesired Paths (Paths 2-4) in the Environment.

3.7 Depth Map Based Design for Sensing Codebooks

As discussed in Section 3.6.1, our objective is to design a sensing beamforming codebook that fits the rectangular grid of the depth camera. In this section, we first present our codebook design that achieves this objective. Then, we incorporate a new side-lobe reduction approach to ameliorate the inter-path interference problem.

3.7.1 Proposed Codebook Design

Since the objective from the beamforming-combining pair codebook design is for the codebook grid to match the desired rectangular grid of a range/depth scene, we start with the relevant camera geometry equations. The scene definition starts by defining the key quantities of the field of view, FoV, and the scene aspect ratio, A_{R}. Let the field of view be centered around the boresight antenna array direction. It is worth noting that the separation distance of the camera plane away from the antenna array reference point, aka the focal length, is irrelevant in our codebook design. This

is based on the notice that the beamforming/combining codebook design normally depends on angles rather than distances.

In a general sense, for any chosen value of focal length, the sensor grid points' coordinates are first calculated to determine the codebook angles accordingly. More specifically, assume that the focal length is set to a certain value, F_{L}. The camera plane width, aka the sensor grid width in the horizontal dimension, S_{H}, and camera plane height, aka the sensor grid height in the vertical dimension, S_{V}, can be calculated as

$$S_{\mathrm{H}} = 2F_{\mathrm{L}} \tan\left(\mathrm{FoV}/2\right), \text{ and } S_{\mathrm{V}} = S_{\mathrm{H}}/A_{\mathrm{R}}. \tag{3.13}$$

For designing a beamforming-combining pair codebook, let $N = N_{\mathrm{V}} \times N_{\mathrm{H}}$, where N_{V} and N_{H} denote the number of UPA antennas on the elevation (vertical) and azimuth (horizontal) dimensions, respectively. Consider an oversampled beamforming codebook of $M = \overline{N}_{\mathrm{V}}\overline{N}_{\mathrm{H}}$ beams, where $\overline{N}_{\mathrm{V}} = N_{\mathrm{V}}F_{\mathrm{V}}^{\mathrm{OS}}$ and $\overline{N}_{\mathrm{H}} = N_{\mathrm{H}}F_{\mathrm{H}}^{\mathrm{OS}}$ with $F_{\mathrm{V}}^{\mathrm{OS}}$ and $F_{\mathrm{H}}^{\mathrm{OS}}$ denoting the oversampling factors in the elevation and azimuth dimensions, respectively. The grid spacing in the vertical and horizontal directions are expressed as $Q_{\mathrm{V}} = S_{\mathrm{V}}/\overline{N}_{\mathrm{V}}$ and $Q_{\mathrm{H}} = S_{\mathrm{H}}/\overline{N}_{\mathrm{H}}$. Notice that the codebook resolution of $\overline{N}_{\mathrm{V}}\overline{N}_{\mathrm{H}}$ beams will be mapped at the end to the desired up-scaled depth image resolution of $M_{\mathrm{res}} = M_h \times M_w$ pixels.

Let the x- and z-axes be aligned in the direction of the sensor grid width and height respectively, and let the y-axis be the direction of the depth. The (x, y, z) rectangular coordinates of the sensor grid points on the camera plane can then be

defined as

$$
\begin{aligned}
&(x,y,z)\in\mathcal{C},\ \ \mathcal{C}=\mathcal{X}\times\mathcal{Y}\times\mathcal{Z}, \qquad (3.14)\\
&\mathcal{X}=\left\{x: x\in\left\{\tfrac{-S_{\mathrm{H}}}{2}+\tfrac{Q_{\mathrm{H}}}{2},\tfrac{-S_{\mathrm{H}}}{2}+\tfrac{3Q_{\mathrm{H}}}{2},\ldots,\tfrac{S_{\mathrm{H}}}{2}-\tfrac{Q_{\mathrm{H}}}{2}\right\}\right\},\\
&\mathcal{Y}=\left\{y: y=F_{\mathrm{L}}\right\},\\
&\mathcal{Z}=\left\{z: z\in\left\{\tfrac{-S_{\mathrm{V}}}{2}+\tfrac{Q_{\mathrm{V}}}{2},\tfrac{-S_{\mathrm{V}}}{2}+\tfrac{3Q_{\mathrm{V}}}{2},\ldots,\tfrac{S_{\mathrm{V}}}{2}-\tfrac{Q_{\mathrm{V}}}{2}\right\}\right\},
\end{aligned}
$$

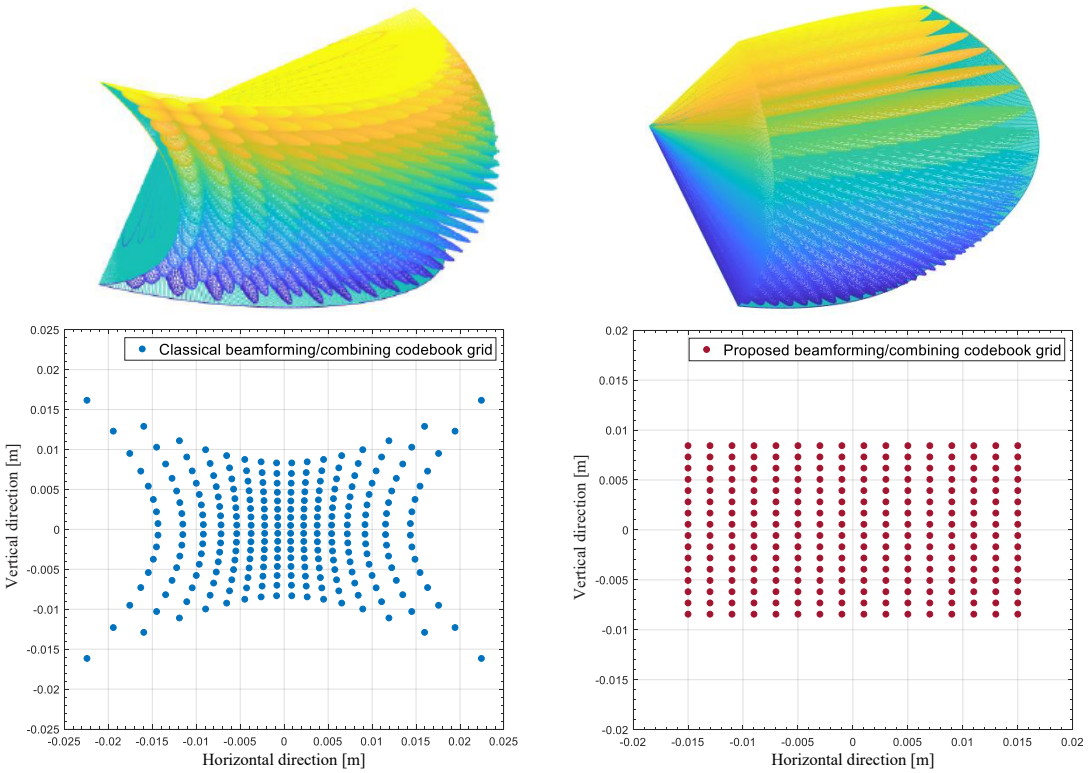
where we note that $|\mathcal{C}|=\overline{N}_{\mathrm{V}}\overline{N}_{\mathrm{H}}=M$. After defining the (x,y,z) coordinates of every grid point on the camera plane, their M corresponding $(\theta_{\mathrm{z}},\theta_{\mathrm{x}})$ angles with respect to the z- and x-axes can now be calculated using the mapping from rectangular to spherical coordinates, such that

$$
\begin{aligned}
\mathcal{O}=\Bigg\{(\theta_{\mathrm{z}},\theta_{\mathrm{x}}):\ &\theta_{\mathrm{z}}=\left[\tfrac{\pi}{2}-\arctan\left(\tfrac{z}{\sqrt{x^2+y^2}}\right)\right], \qquad (3.15)\\
&\theta_{\mathrm{x}}=\left[\tfrac{\pi}{2}-\arctan\left(\tfrac{x}{\sqrt{y^2+z^2}}\right)\right],(x,y,z)\in\mathcal{C}\Bigg\}.
\end{aligned}
$$

Finally, after calculating the $(\theta_{\mathrm{z}},\theta_{\mathrm{x}})$ angles for each and every grid point, the beamforming codebook, $\boldsymbol{\mathcal{F}}$, for an $N_{\mathrm{H}}\times N_{\mathrm{V}}$ transmit UPA, is then expressed as

$$
\begin{aligned}
&\boldsymbol{\mathcal{F}}=\left\{\mathbf{f}\in\mathbb{C}^{N\times 1}:\mathbf{f}=\widetilde{\mathbf{b}}_{\mathrm{V}}\left(\theta_{\mathrm{z}}\right)\circ\widetilde{\mathbf{b}}_{\mathrm{H}}\left(\theta_{\mathrm{x}}\right),(\theta_{\mathrm{z}},\theta_{\mathrm{x}})\in\mathcal{O}\right\}, \qquad (3.16)\\
&\widetilde{\mathbf{b}}_{\mathrm{V}}\left(\theta_{\mathrm{z}}\right)=\left[1,\mathrm{e}^{-\jmath\kappa d_s\cos(\theta_{\mathrm{z}})},\ldots,\mathrm{e}^{-\jmath(N_{\mathrm{V}}-1)\kappa d_s\cos(\theta_{\mathrm{z}})}\right]^T,\\
&\widetilde{\mathbf{b}}_{\mathrm{H}}\left(\theta_{\mathrm{x}}\right)=\left[1,\mathrm{e}^{-\jmath\kappa d_s\cos(\theta_{\mathrm{x}})},\ldots,\mathrm{e}^{-\jmath(N_{\mathrm{H}}-1)\kappa d_s\cos(\theta_{\mathrm{x}})}\right]^T,
\end{aligned}
$$

where $\kappa=\frac{2\pi}{\lambda}$ is the wave number, λ is the operating wavelength, and d_s is the antenna element spacing between adjacent UPA elements in meters. $\widetilde{\mathbf{b}}_{\mathrm{H}}\in\mathbb{C}^{N_{\mathrm{H}}\times 1}$ and $\widetilde{\mathbf{b}}_{\mathrm{V}}\in\mathbb{C}^{N_{\mathrm{V}}\times 1}$ are the horizontal and vertical basic vectors used for constructing the beamforming codebook. We will call these vectors, $\widetilde{\mathbf{b}}_{\mathrm{H}}$ and $\widetilde{\mathbf{b}}_{\mathrm{V}}$, the *constituent* horizontal and vertical beamforming vectors, respectively. In our depth estimation problem, the receive combining codebook, $\boldsymbol{\mathcal{W}}$, can be similarly defined for the $N_{\mathrm{H}}\times N_{\mathrm{V}}$



(a) The classical beamforming/combining codebook (b) The proposed beamforming/combining codebook

Figure 3.7: The Comparison Between (a) the Classical (on the Left Side) and the Proposed (on the Right Side) Beam Codebook Design Is Demonstrated for a Scene of 100° Field of View and $16/9$ Aspect Ratio, Using 16×16 UPAs. The Proposed Codebook Eliminates Any Grid Mismatch Distortion. The Top Figures Are the 3D Codebook Radiation Patterns, While the Bottom Figures Are the 2D Codebook Grids at a Plane Within 13.32mm Depth.

receive UPA. For such case, the cardinalities of the sets are equal, $|\boldsymbol{\mathcal{W}}| = |\boldsymbol{\mathcal{F}}| = |\mathcal{C}| = |\mathcal{O}| = M$. Further, let $\mathbf{F} \in \mathbb{C}^{N \times M}$ and $\mathbf{W} \in \mathbb{C}^{N \times M}$ be the matrices that consist of the codebooks beams of $\boldsymbol{\mathcal{F}}$ and $\boldsymbol{\mathcal{W}}$. Then, the proposed sensing beamforming-combining pair codebook $\boldsymbol{\mathcal{P}}$ can be expressed as

$$\boldsymbol{\mathcal{P}} = \Big\{ (\mathbf{f}_m, \mathbf{w}_m) \in \mathbb{C}^{N \times 1} \times \mathbb{C}^{N \times 1} : \mathbf{f}_m = [\mathbf{F}]_{:,m}, \quad \mathbf{w}_m = [\mathbf{W}]_{:,m}, m \in \{1, \ldots, M\} \Big\}. \tag{3.17}$$

A comparison between the classical and the proposed beam codebook design is demonstrated in Fig. 3.7 for a scene of 100° field of view and $16/9$ aspect ratio, using 16×16 UPAs. The top figures are the 3D codebook radiation patterns, while the

bottom figures are the 2D codebook grids at a plane within 13.32mm depth. As shown, the proposed beam codebook eliminates any grid mismatch distortion.

3.7.2 Sidelobe Reduction Approach

As discussed in Section 3.6.2, to rectify the inter-path interference problem, the sensing framework needs to filter out the undesired channel paths. As illustrated in Fig. 3.6, one type of undesired channel paths is the type of paths transmitted from/received by the sidelobes of a codebook beam. For this reason, we propose an efficient sidelobe reduction (SLR) approach. In [96, 97], an SLR approach was proposed for low sidelobe beamforming in uniform circular arrays. Inspired by their work, we propose a new efficient sidelobe reduction approach to uniform planar arrays (UPAs) to reduce beamforming/combining sidelobe levels.

The key idea of this approach is when applying different weights on the beamforming/combining vector elements, the sensing framework can control the beam radiation pattern in a way to increase the power difference between the mainlobe and the sidelobes. Specifically, let $\mathbf{c}_{\mathrm{H}} \in \mathbb{R}^{N_{\mathrm{H}} \times 1}$ and $\mathbf{c}_{\mathrm{V}} \in \mathbb{R}^{N_{\mathrm{V}} \times 1}$ represent the horizontal and vertical weight vectors for sidelobe reduction. Let $\mathbf{b}_{\mathrm{H}}$ and $\mathbf{b}_{\mathrm{V}}$ denote the horizontal and vertical constituent beamforming vectors after sidelobe reduction. The updated beamforming codebook, $\boldsymbol{\mathcal{F}}$, for an $N_{\mathrm{H}} \times N_{\mathrm{V}}$ transmit UPA, can then be rewritten as

$$
\begin{aligned}
&\boldsymbol{\mathcal{F}} = \left\{ \mathbf{f} \in \mathbb{C}^{N \times 1} : \mathbf{f} = \mathbf{b}_{\mathrm{V}}\left(\theta_{\mathrm{z}}\right) \circ \mathbf{b}_{\mathrm{H}}\left(\theta_{\mathrm{x}}\right), \left(\theta_{\mathrm{z}}, \theta_{\mathrm{x}}\right) \in \mathcal{O} \right\}, \qquad (3.18)\\
&\mathbf{b}_{\mathrm{V}}\left(\theta_{\mathrm{z}}\right) = \widetilde{\mathbf{b}}_{\mathrm{V}}\left(\theta_{\mathrm{z}}\right) \odot \mathbf{c}_{\mathrm{V}},\ \mathbf{b}_{\mathrm{H}}\left(\theta_{\mathrm{x}}\right) = \widetilde{\mathbf{b}}_{\mathrm{H}}\left(\theta_{\mathrm{x}}\right) \odot \mathbf{c}_{\mathrm{H}},\\
&\left[\mathbf{c}_{\mathrm{V}}\right]_{r_{\mathrm{V}}} = \mathrm{e}^{\frac{-(r_{\mathrm{V}}-\mu_{\mathrm{V}})^2}{2\sigma_{\mathrm{V}}^2}}, \mu_{\mathrm{V}} = \tfrac{N_{\mathrm{V}}}{2}, \sigma_{\mathrm{V}} = \tfrac{N_{\mathrm{V}}}{\delta_{\mathrm{V}}}, r_{\mathrm{V}} \in \{1, \ldots, N_{\mathrm{V}}\}\\
&\left[\mathbf{c}_{\mathrm{H}}\right]_{r_{\mathrm{H}}} = \mathrm{e}^{\frac{-(r_{\mathrm{H}}-\mu_{\mathrm{H}})^2}{2\sigma_{\mathrm{H}}^2}}, \mu_{\mathrm{H}} = \tfrac{N_{\mathrm{H}}}{2}, \sigma_{\mathrm{H}} = \tfrac{N_{\mathrm{H}}}{\delta_{\mathrm{H}}}, r_{\mathrm{H}} \in \{1, \ldots, N_{\mathrm{H}}\}
\end{aligned}
$$

where $\delta_{\mathrm{H}}, \delta_{\mathrm{V}}$ denote the sidelobe reduction control variables; the higher the values, the greater reduction in the sidelobe power levels compared to the mainlobe power level.

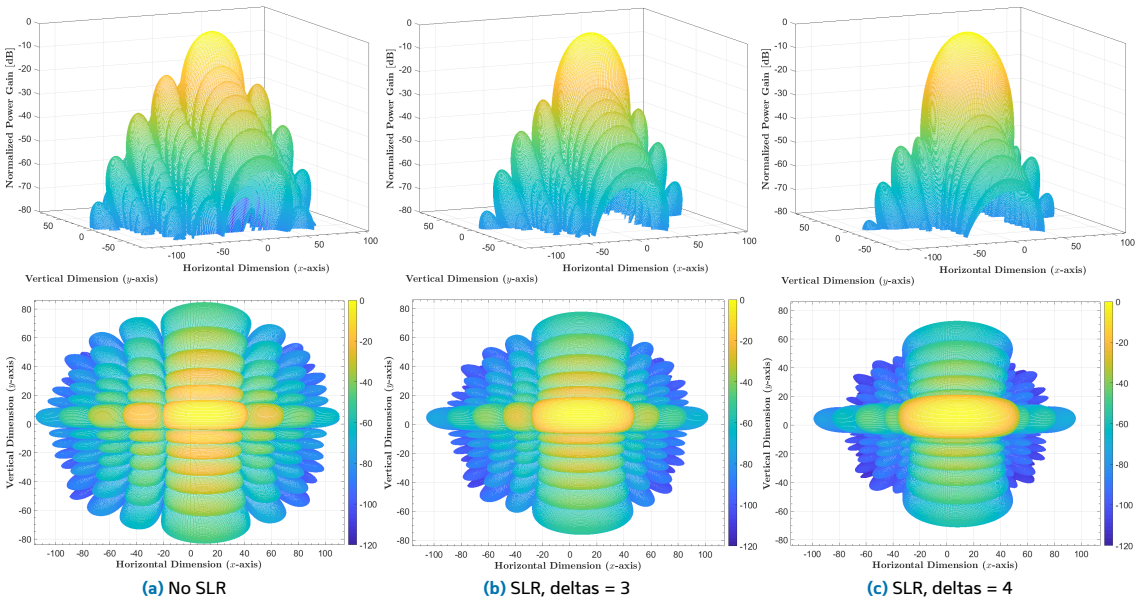
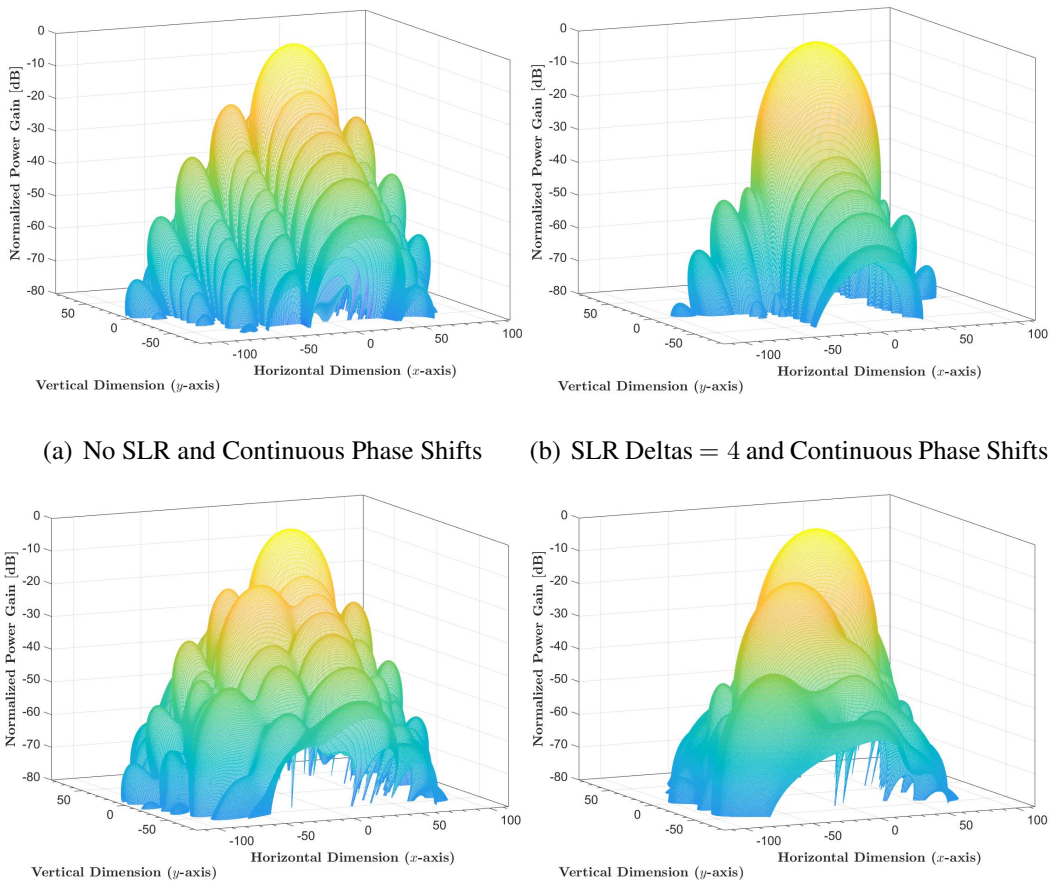


Figure 3.8: Normalized Power Radiation Pattern Comparison Between (a) the Case Without the Sidelobe Reduction (SLR) Approach, (b) the Case with the SLR Approach Where $\delta_{\mathrm{H}} = \delta_{\mathrm{V}} = 3$, and (c) Where $\delta_{\mathrm{H}} = \delta_{\mathrm{V}} = 4$. As Shown, Increasing the Values of the Control Variables (the Deltas) Increases the Gap Between the Mainlobe Level and the Sidelobes Levels. The Top Figures Are the 3D Views of the Patterns While the Bottom Figures Are the Top Views.

The updated combining codebook $\boldsymbol{\mathcal{W}}$ can be similarly defined. The beam codebook $\boldsymbol{\mathcal{P}}$ follows the same definition in (3.17).

Fig. 3.8 illustrates the radiation pattern in dB for one beamforming vector out of the updated beamforming codebook, $\boldsymbol{\mathcal{F}}$, for different values of the sidelobe reduction control variables, $\delta_{\mathrm{H}}, \delta_{\mathrm{V}}$. As depicted, increasing the values of the control variables increases the power gap between the mainlobe level and the sidelobes levels. To take into consideration the phase quantization of the RF phase shifters in the AR/VR transceiver architecture previously shown in Fig. 3.2, we examine the effect of 2-bit phase quantization on the power radiation pattern. The 2-bit discrete phase shift set is $\left\{0, \frac{\pi}{2}, \pi, \frac{3\pi}{2}\right\}$. Fig. 3.9 compares the normalized power radiation pattern between the case of continuous phase shifts and the case of 2-bit quantized phase shifts. As depicted, the phase quantization affects the beam pattern shape of the sidelobes.



(a) No SLR and Continuous Phase Shifts

(b) SLR Deltas $= 4$ and Continuous Phase Shifts

(c) No SLR and 2-bit Phase Shifts

(d) SLR Deltas $= 4$ and 2-bit Phase Shifts

Figure 3.9: Normalized Power Radiation Pattern Comparison Between the Case with No Phase Quantization and the Case with 2-bit Phase Quantization, for Two Scenarios: Without or with the Sidelobe Reduction (SLR) Approach Where $\delta_{\mathrm{H}} = \delta_{\mathrm{V}} = 4$.

One main advantage of this approach is its computational efficiency; as formulated, only two element-wise multiplication between the weight vectors and the constituent beamforming vectors, $\widetilde{\mathbf{b}}_{\mathrm{H}}, \widetilde{\mathbf{b}}_{\mathrm{V}}$, are needed to update the beam radiation pattern. This multiplication, however, requires an analog beamforming architecture with the capability of changing both the phase and magnitude. In the results section, we only used this SLR-based beam codebook in the simulations of Fig. 3.25. In the future work, it is interesting to explore phase-only approximations of this SLR-based beam codebook structure. By contrast, reducing the sidelobe levels dramatically increases

the beamwidth of the mainlobe, as depicted in Fig. 3.8. The increased mainlobe beamwidth, however, can be mitigated by the other solutions proposed for rectifying the inter-path interference problem, e.g. the successive interference cancellation (SIC) algorithm and the joint processing (JP) solution, as will be described in the following section.

3.8 Proposed Scene Range/Depth Estimation

In this section, given a pre-designed beamforming/combining codebook, $\boldsymbol{\mathcal{P}}$, we propose an efficient approach for the scene range/depth estimation in AR/VR devices. As depicted in Fig. 3.4, once the beamforming/combining codebook has been designed, the AR/VR transmits the sensing signal while sweeping over all the beamforming-combining vector pairs. Specifically, for a beamforming-combining vector pair $(\mathbf{f}_m, \mathbf{w}_m)$, where $m \in \{1, \ldots, M\}$, the receive *sensing* signal, $\mathbf{y}_m \in \mathbb{C}^{N^{\mathrm{P}}+L_{\mathrm{d}}}$, can be modeled as in (3.6) and (3.7). After reception, the acquired sensing signals are processed to estimate the range and depth maps, as will be thoroughly explained in this section. Our proposed post-processing solution has three main elements: (i) The use of oversampled/overlapped beams, (ii) the successive interference cancellation based management of inter-target and inter-path interference, and (iii) the joint processing of the signals received using the codebook beams to realize high-resolution and accurate depth maps. Next, we explain these three elements in Sections 3.8.1-3.8.3 before presenting the scene range/depth map construction approach in Section 3.8.4.

3.8.1 Overlapped Beams

With the objective of increasing the resolution of the mmWave MIMO based depth maps, we propose to adopt oversampled sensing codebooks to scan the surrounding environment. In particular, for the sensing codebook, we adopt the developed code-

book in Section 3.7.1 with oversampling factors of $F_{\mathrm{H}}^{\mathrm{OS}}$ and $F_{\mathrm{V}}^{\mathrm{OS}}$ in the azimuth and elevation directions. While the oversampled codebook has the potential of enhancing the depth map resolution, it is important to note that advanced post-processing (for the receive signals using these oversampled beams) needs to be incorporated to achieve this goal. The reason mainly goes back to the wide beamwidth (and low spatial resolution) of the codebook beams, which is fundamentally limited by the number of AR/VR antennas. This wide beamwidth leads to a number of challenges: (i) The spatial regions scanned by the oversampled beams have high overlap. This makes it hard to differentiate between the depths of the different objects in the depth map pixels, which challenges the objective of realizing high-resolution depth maps. (ii) Since the codebook beams still have wide beamwidth, the inter-target interference problem discussed in Section 3.6.2 still exists.

To address these challenges, we propose a novel post-processing approach based on successive interference cancellation and joint-beam processing. This approach in summarized in two main steps as follows. In the first step, a successive interference cancellation (SIC) based algorithm is used to detect the most dominant channel paths contributing to the range/depth estimation of the region covered by each codebook beam. These paths form a set of candidate ranges/depths for the scene range/depth estimation. In the second step, a developed joint-beam processing solution selects one range/depth out of the set of candidate ranges/depths formed by the SIC algorithm. These two sequential algorithms are discussed in detail in the following two subsections.

3.8.2 Successive Interference Cancellation

The main goal of the successive interference cancellation (SIC) algorithm is to detect all the dominant paths that might contribute to the range estimation of the

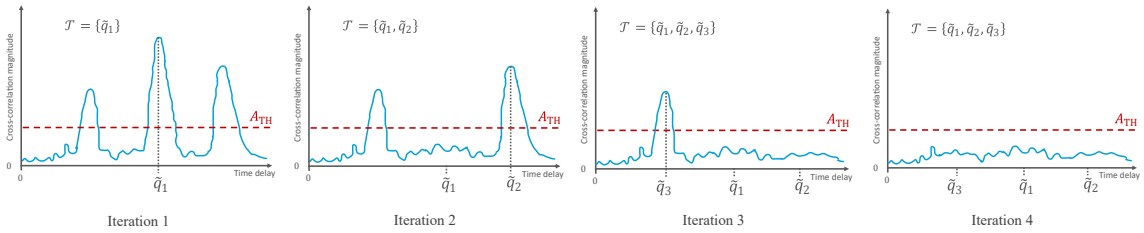


Figure 3.10: The Operation of the Successive Interference Cancellation (SIC) Algorithm Is Illustrated. The Delay Position of the Maximum Cross-correlation Is First Detected. The SIC Algorithm Then Encodes a Signal Shifted at This Delay Position and Subtracted It from the Receive Signal. After That, the Algorithm Repeats Itself until All the Local Maxima above the Threshold Value Are Detected.

region of interest. This is motivated by its good performance in multi-target detections problems [98]. The SIC algorithm is applied in the discrete-time domain and is summarized in Algorithm 3. The algorithm is described as follows. Let the length of the receive sensing sequence $y_m[n]$ be $L_{\mathsf{y}} = N^{\mathrm{p}} + L_{\mathsf{d}}$ symbols. First, as shown in Fig. 3.10, for every codebook beam, the delay position of the maximum cross-correlation magnitude value is detected. $\mathcal{Q}$ is the set of possible delays. Second, the SIC algorithm encodes the transmit preamble signal to be shifted to this delay position and subtracted it from the received signal. Afterwards, the algorithm repeats itself to detect the second local maximum above the threshold value. Finally, The SIC algorithm stops iterating when all the local maxima above the threshold value are detected. The output of this algorithm is a set of candidate delays for every codebook beam. These sets pass as input to the next algorithm, the joint processing solution, as will be explained in the next part. In Fig. 3.10, note that the cross-correlation magnitude plot appears to be drawn as a continuous plot, only for illustration purposes. The actual cross-correlation magnitude, however, is expressed in discrete time delays.

Algorithm 3 Successive Interference Cancellation

Inputs: Receive sensing signal $y_m[n]$, transmit preamble signal $s^{\mathrm{p}}[n]$, threshold level A_{TH}, beamforming-combining pair codebook $\boldsymbol{\mathcal{P}}$.

Outputs: Candidate delay set for each beam, $\mathcal{T}_m, \forall m \in \{1, \ldots, M\}$,.

1: **for** $m = 1$ **to** M **do**

2: **Initialize:** Updated signal $\tilde{y}_m[n] \leftarrow y_m[n]$, solution set $\mathcal{T}_m \leftarrow \emptyset, \forall m$, and $\tilde{A} \leftarrow A_{\mathrm{TH}}$.

3: **while** $\tilde{A} \geq A_{\mathrm{TH}}$ **do**

4: Calculate the delay of the path with maximum cross-correlation

$$\tilde{q} \leftarrow \underset{q:q\in\mathcal{Q}}{\arg\max} \left| \sum_{n=L_{\mathrm{y}}-N_{\mathrm{y}}}^{L_{\mathrm{y}}-1} s^{\mathrm{p}}[n] \times (\tilde{y}_m[n-q])^* \right|^2 .$$

5: Add the candidate delay $\tilde{q}$ to the solution set

$$\mathcal{T}_m \leftarrow \mathcal{T}_m \cup \{\tilde{q}\} .$$

6: Calculate the energy of the transmit signal up to $\tilde{q}$

$$E_{\mathrm{Q}} \leftarrow \sum_{n=L_{\mathrm{y}}-N_{\mathrm{y}}}^{L_{\mathrm{y}}-\tilde{q}-1} |s^{\mathrm{p}}[n]|^2 .$$

7: Perform interference cancellation at $\tilde{q}$

$$\tilde{A} \leftarrow \left[\sum_{n=L_{\mathrm{y}}-N_{\mathrm{y}}}^{L_{\mathrm{y}}-1} s^{\mathrm{p}}[n] \times (\tilde{y}_m[n-\tilde{q}])^* \right] .$$

$$\tilde{y}_m[n] \leftarrow \tilde{y}_m[n] - \frac{\tilde{A}}{E_{\mathrm{Q}}} s^{\mathrm{p}}[n-\tilde{q}].$$

8: Calculate the next max. in the cross-correlation

$$\tilde{A} \leftarrow \max_{q:q\in\mathcal{Q}} \left| \sum_{n=L_{\mathrm{y}}-N_{\mathrm{y}}}^{L_{\mathrm{y}}-1} s^{\mathrm{p}}[n] \times (\tilde{y}_m[n-q])^* \right|^2 .$$

Algorithm 4 Joint Processing Solution

Inputs: Candidate delay set for each beam, $\mathcal{T}_{h,v}, \forall h \in \{1, \ldots, \overline{N}_{\mathrm{H}}\}, \forall v \in \{1, \ldots, \overline{N}_{\mathrm{V}}\}$.

Outputs: Scene range estimate $(\hat{\rho}_{h,v})^{\mathrm{SRE}}, \forall h, v$.

1: **Initialize:** Common adjacent set $\mathcal{N}_{h,v} \leftarrow \emptyset, \forall h, v$, difference set

$$\mathcal{M}_{h,v} \leftarrow \emptyset, \forall h, v.$$

2: **for** $v = 1$ **to** $\overline{N}_{\mathrm{V}}$ **do**

3: **for** $h = 1$ **to** $\overline{N}_{\mathrm{H}}$ **do**

4: Construct the common adjacent set

$$\mathcal{N}_{h,v} \leftarrow (\mathcal{T}_{h-1,v} \cup \mathcal{T}_{h,v-1} \cup \mathcal{T}_{h-1,v-1} \cup \mathcal{T}_{h+1,v-1}).$$

5: Construct the difference set

$$\mathcal{M}_{h,v} \leftarrow \mathcal{T}_{h,v} \setminus \mathcal{N}_{h,v}.$$

6: **if** $\mathcal{M}_{h,v} \neq \emptyset$ **then**

7: Choose the least delay from the difference set

$$(\hat{\rho}_{h,v})^{\mathrm{SRE}} \leftarrow \frac{\varsigma T_{\mathrm{S}}}{2} \min \mathcal{M}_{h,v}.$$

8: **else**

9: Choose the least delay from the candidate set

$$(\hat{\rho}_{h,v})^{\mathrm{SRE}} \leftarrow \frac{\varsigma T_{\mathrm{S}}}{2} \min \mathcal{T}_{h,v}.$$

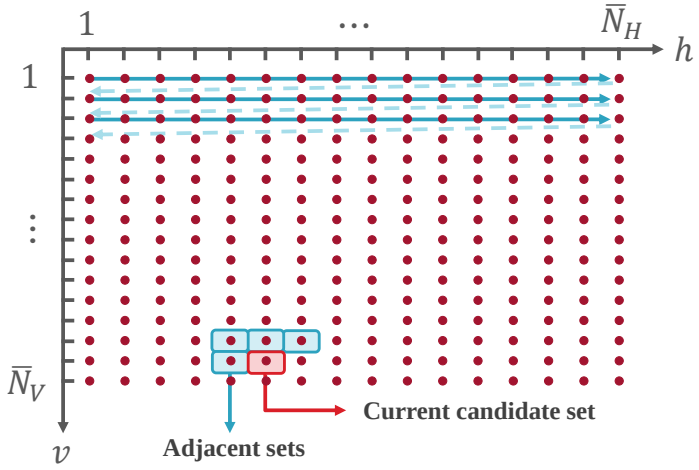


Figure 3.11: This Figure Illustrates the Basic Operation of the Joint Processing (JP) Solution for Overlapped Beams. The JP Solution Sweeps from Left to Right, Then from Top to Bottom. The JP Solution Decides on Which Path to Choose from the Current Candidate Set by a Simple Comparison with the Sets of the Surrounding Grid Points.

3.8.3 Joint Processing Solution

The purpose of the joint processing (JP) solution between the overlapped beams is to estimate the transitions in depth/range maps more accurately. The proposed JP solution is summarized in Algorithm 4. The algorithm is described in detail as follows. First, the JP solution works on the candidate delay sets, the output from the SIC algorithm, $\{\mathcal{T}_m\}_{m=1}^{M}$, to choose one range estimate out of the candidate delay set. This processing, however, is employed relative to the 2D codebook grid, as illustrated in Fig. 3.11. Following this notion, the linear indices in $\mathcal{T}_m$ is now converted into matrix subscripts $\mathcal{T}_{h,v}$ through the transformation $m = (v-1)\overline{N}_{\mathrm{H}} + h$, such that $\mathcal{T}_m = \mathcal{T}_{h,v}$, where v is the elevation beam index (vertical grid index) and h is the azimuth beam index (horizontal grid index). The objective is to calculate the scene range estimates across all beam directions, $(\hat{\rho}_{h,v})^{\mathrm{SRE}}, \forall h, v$.

As shown in Fig. 3.11, the JP solution sweeps from left to right, then from top to bottom. For each grid point, the JP solution uses (i) the set of the current grid point, named as the "current set", and (ii) the sets of the previous adjacent grid points to

construct a "common adjacent set". This common adjacent set is the union of the sets of all previous adjacent grid points. Then, to investigate if a new object/surface transition appears, this current set is compared with the common adjacent set to detect if there is any set difference. This is based on the notion that the difference set can probably be the new edges that will appear in the range map while sweeping. If the set difference is not empty, then the solution chooses the path with the least time-of-flight from the set difference. Otherwise, if the set difference is empty, then the solution chooses the path with the least time-of-flight from the current set.

Algorithm 5 mmWave MIMO Sensing Based Range/Depth Estimation Framework

Inputs: Field of view FoV, aspect ratio A_{R}, number of horizontal and vertical beams $\overline{N}_{\mathrm{H}}, \overline{N}_{\mathrm{V}}$.

Outputs: Range map $\widehat{\mathbf{R}}_{\mathrm{map}}$, depth map $\widehat{\mathbf{D}}_{\mathrm{map}}$.

1: Design the beamforming-combining pair codebook, $\boldsymbol{\mathcal{P}}$, following Section 3.7.

2: **for** $m = 1$ **to** M **do** ▷ For each pair $(\mathbf{f}_m, \mathbf{w}_m)$.

3: Acquire receive *sensing* signal, $y_m[n]$, as in (3.6), $\forall n \in \{0, 1, \ldots, N^{\mathrm{p}} + L_{\mathsf{d}} - 1\}$.

4: Calculate the candidate delay set for each beam, $\mathcal{T}_m$, $\forall m$, as in Algorithm 3.

5: Calculate the scene range estimate, $(\hat{\rho}_m)^{\mathrm{SRE}}$, $\forall m$, as in Algorithm 4.

6: Calculate fine range estimates, following Section 3.5.2.

$$(\hat{\rho}_m)^{\mathrm{MC}} \leftarrow (\hat{\rho}_m)^{\mathrm{SRE}} + (\hat{\rho}_m)' \, , \forall m.$$

7: Construct the range map, $\widehat{\mathbf{R}}_{\mathrm{map}}$, from (3.20).

8: Construct the depth map, $\widehat{\mathbf{D}}_{\mathrm{map}}$, from (3.21).

3.8.4 Range/Depth Map Construction

In this section, we formulate the depth map construction approach, the last step in Fig. 3.4. In summation of the broader view, the mmWave MIMO sensing based range/depth map estimation framework is outlined in Algorithm 5. The algorithm steps are summarized as follows. *Step 1* refers to the design of the beamforming-combining pair codebook $\boldsymbol{\mathcal{P}}$ was covered in Section 3.7. *Step 4* refers to the successive interference cancellation described in Section 3.8.2. *Step 5* refers to the joint processing solution detailed in Section 3.8.3. After that, in *Step 6*, the fine range estimate can be calculated, such that $(\hat{\rho}_m)^{\mathrm{MC}} = (\hat{\rho}_m)^{\mathrm{SRE}} + (\hat{\rho}_m)'$, where $(\hat{\rho}_m)'$ is computed from the algorithm described in Section 3.5.2. Next, after calculating the range estimates, the upcoming steps (*Steps 7,8*) are focused on constructing the range and depth maps. Note that the range of an object is actually the radial distance in spherical coordinates. Fortunately, the (x, y, z) rectangular coordinates of the sensor grid points on the camera plan were already calculated for the design of the beamforming-combining pair codebook using (3.14). These rectangular coordinates in (3.14) can then be converted to spherical coordinates, such that

$$\mathcal{S} = \Bigg\{ (\theta_{\mathrm{z}}, \Phi) : \theta_{\mathrm{z}} = \left[\tfrac{\pi}{2} - \arctan\left(\tfrac{z}{\sqrt{x^2+y^2}} \right) \right], \quad \Phi = \arctan\left(\tfrac{y}{x}\right), (x, y, z) \in \mathcal{C} \Bigg\}. \tag{3.19}$$

In order to construct the matrices for the range and depth maps, let $\boldsymbol{\Theta}, \boldsymbol{\Phi} \in \mathbb{R}^{\overline{N}_{\mathrm{V}} \times \overline{N}_{\mathrm{H}}}$ be the matrices that represent the angles of the spherical coordinates $(\theta_{\mathrm{z}}, \Phi) \in \mathcal{S}$, respectively. Following *Step 6* in Algorithm 5, the range map estimate $\widehat{\mathbf{R}}_{\mathrm{map}} \in \mathbb{R}^{\overline{N}_{\mathrm{V}} \times \overline{N}_{\mathrm{H}}}$ can be expressed as

$$\left[\widehat{\mathbf{R}}_{\mathrm{map}}\right]_{v,h} = (\hat{\rho}_m)^{\mathrm{MC}}, \ m = (v-1)\overline{N}_{\mathrm{H}} + h, \tag{3.20}$$

where $m \in \{1, \ldots, M\}$, $h \in \{1, \ldots, \overline{N}_{\mathrm{H}}\}$, $v \in \{1, \ldots, \overline{N}_{\mathrm{V}}\}$. Given the angles in

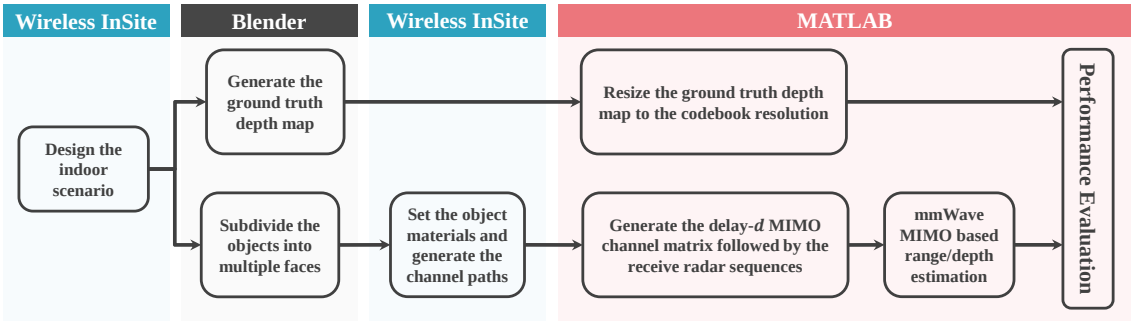


Figure 3.12: This Figure Demonstrates the Adopted Simulation Framework for Scene Depth Estimation. The Framework Consists of Designing the Indoor Setup, Generating the Ground Truth Range/Depth Maps, and Constructing the Estimated Maps for Performance Evaluation. For More Complex Setups, Designing the Indoor Scenarios Jointly in Wireless InSite and Blender Can Be More Effective.

spherical coordinates and the range map estimate, the depth map estimate $\widehat{\mathbf{D}}_{\text{map}} \in \mathbb{R}^{\overline{N}_{\text{V}} \times \overline{N}_{\text{H}}}$ can then be expressed as

$$\left[\widehat{\mathbf{D}}_{\text{map}}\right]_{v,h} = \left|\hat{\rho} \sin\left(\theta_{\text{z}}\right) \sin\left(\Phi\right)\right|, \tag{3.21}$$

$$\text{where} \quad \hat{\rho} = \left[\widehat{\mathbf{R}}_{\text{map}}\right]_{v,h}, \ \theta_{\text{z}} = \left[\boldsymbol{\Theta}\right]_{v,h}, \ \Phi = \left[\boldsymbol{\Phi}\right]_{v,h}, \ \forall v, h.$$

Finally, since the range and depth map resolutions are set to $\overline{N}_{\text{H}} \times \overline{N}_{\text{V}}$, two-dimensional image interpolation can be employed to scale the maps to the desired resolutions, $M_h \times M_w$. Examples of interpolation methods are the nearest neighbor interpolation and the bicubic interpolation. Although the bicubic interpolation can probably be the interpolation method of choice for achieving more estimation accuracy, the nearest neighbor interpolation is more computationally efficient. In the simulation results of Section 3.9, we evaluate the two interpolation approaches for our mmWave MIMO based depth map construction problem.

3.9 Simulation Results

In this section, we evaluate the performance of the proposed mmWave based depth estimation approach. First, we describe the adopted simulation framework

in Section 3.9.1 before extensively studying the estimation accuracy of the proposed approach under various scenarios and system parameters. The simulation results presented can be of great usefulness for various applications; they can be generally applied to AR/VR devices, smart home devices, or auto drive devices.

3.9.1 Simulation Framework

Since the depth estimation heavily depends on the environment under test, it is crucial to evaluate the performance of the proposed solution based on realistic channels. This motivates using channels generated by accurate ray-tracing to capture the sensing dependence on the environment geometry, scatterers' materials, AR/VR position, etc. This is why we designed the simulations models using Remcom Wireless InSite [1], which is an accurate 3D ray-tracing simulator. Further, to efficiently incorporate diffuse scattering models, we need to have highly detailed floor plans with a sufficient number of faces. To achieve this objective, we resorted to the high-fidelity game engine, Blender [3], to build accurate floor plans. These plans/models are then exported to Wireless InSite to obtain the ray-tracing outputs, and finally to MATLAB to construct the channel models in (3.4) and implement the proposed depth estimation approach. The proposed evaluation framework is illustrated in Fig. 3.12. For benchmarking, we also use the Blender floor plans to obtain the ground truth depth maps, which are essential to evaluate the accuracy of our solutions. The ground truth maps are generated by placing a Blender camera at the same position of the UPA reference antenna element, and adjusting the Blender camera parameters to capture the same field of view.

Signal model: We adopt the signal model described in Section 3.3 with a focus on the sensing system performance. The AR/VR device is assumed to be fixed in position. Unless otherwise mentioned, the UPA size is 16×16 antennas ($N_{\mathrm{H}} =$

$N_{\mathrm{V}} = 16$) at the mmWave 60GHz operating band with transmission bandwidth of 2GHz. The antenna elements have a gain of 0dBi with half-wavelength antenna spacing. The transmit power is set to 30dBm. The preamble sequence is the same as the one in the single carrier PHY packet preamble of the IEEE 802.11ad standard (3328 symbols). M preamble sequences are used to sense the environment via M beamforming-combining pairs. For the sake of calculating a rough estimate of the time allocated for environment sensing through transmission and reception, assume that all the M preamble sequences are transmitted sequentially with guard intervals in between. The highest M value reported in the upcoming simulation results is 4096 beams. Assuming a sampling rate of 2Gsps, the sensing time estimate is $\approx$ 7ms.

Channel generation: The channel matrix, $\mathbf{H}_d$, is generated in two steps. The first step is generating the channel rays using the ray-tracing software, Wireless InSite. The Wireless InSite propagation model is set to 'X3D' with 0.1° ray-spacing and enabled mode of diffuse scattering. Up to three reflections, one diffraction, and one transmission properties are allowed for each ray in the Wireless InSite simulation. The diffuse scattering model used is "directive with backscatter"; this model is fixed across all materials in all the testing scenarios. The chosen diffuse scattering model creates two scattering lobes; a forward lobe of diffuse scattered power centered on the direction of specular reflection and a backward lobe centered on the opposite direction of incidence. The diffuse scattering parameters of the different materials are summarized in Table 3.1. The values reported in Table 3.1 follow the ITU default parameter values at 60GHz. The second step in the sensing channel generation is calculating the delay-d channel matrix out of the channel paths using the DeepMIMO dataset generation code [2]. Using these channels and following (3.3)-(3.4), the noisy receive sensing sequences are generated. The noise power is calculated based on a 2GHz bandwidth and a receiver noise figure of 7dB.

Table 3.1: The Adopted Diffuse Scattering Parameters for Different Materials

Diffuse Scattering Parameter	**Concrete**	**Ceilingboard**	**Wood**	**Floorboard**	**Drywall**	**Glass**
Scattered to incident electric field ratio	40%	30%	15%	15%	10%	0%
Forward to backward scattering power ratio	75%	75%	75%	75%	75%	75%
Cross-polarization ratio	40%	40%	40%	40%	40%	40%
Narrowness of the scattering lobes	40%	40%	40%	40%	40%	40%

Table 3.2: The Estimation Error Results of the One Wall Scenario for Different Wall Materials

Estimation Error (m)	**Concrete**	**Ceilingboard**	**Wood**	**Floorboard**	**Layered Drywall**	**Glass**
Basic Correlator	0.101	0.099	0.101	0.0984	0.103	12.697
Massive Correlator	0.0983	0.097	0.0983	0.0983	0.101	15.498

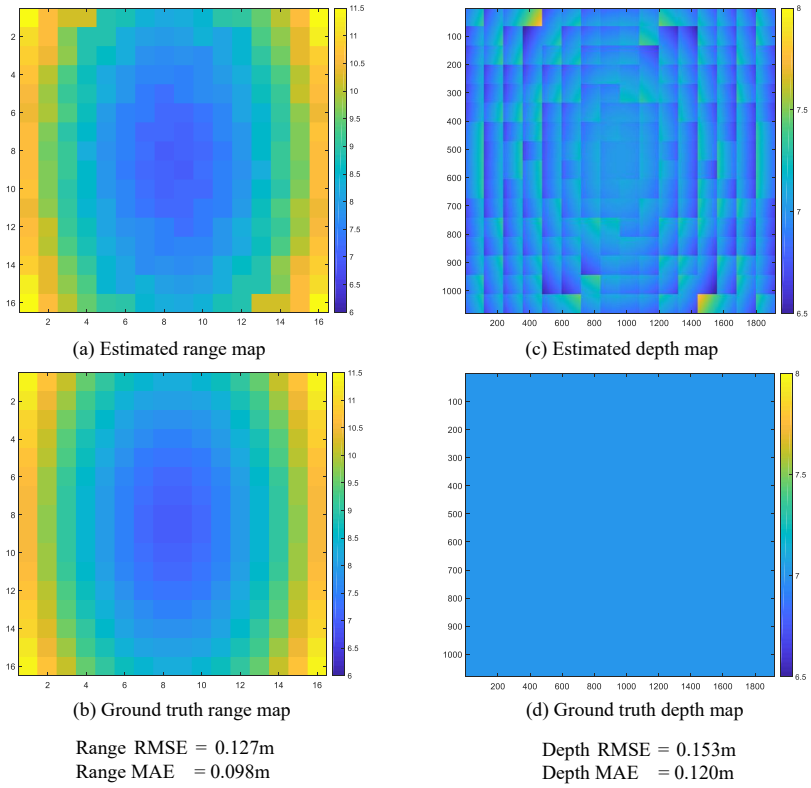


Figure 3.13: The Maps for the One Wall Scenario Are Depicted for a Separation Distance of 7 Meters from the AR/VR Device with 16×16 UPAs. The Depicted Maps Are the Estimated Maps (at the Top), Ground Truth Maps (at the Bottom), Range Maps (on the Left Side), and 1080p Depth Maps (on the Right Side). Comparing (a) with (b), the Range Map Estimation Error: MAE $= 0.098$m. Comparing (c) with (d), the Depth Map Estimation Error: MAE $= 0.12$m.

mmWave based depth estimation parameters: The beamforming-combining pair codebook is designed based on a 100° field of view centered on the antenna array boresight, a 16/9 scene aspect ratio, and horizontal and vertical oversampling factors of unity. The ground truth depth maps are generated from Blender using a Blender camera with a 100° field of view, a focal length of 13.43mm corresponding to a sensor width of 32mm. The ground truth depth map image quality is set to 1080p resolution; i.e., 1920×1080 pixels. Concerning the massive correlator, f_{est} is set to 100 multiple of the sampling frequency f_{S}; i.e., $\delta = \frac{f_{\text{est}}}{2f_{\text{S}}} = 50$. Unless mentioned otherwise, the massive correlator is adopted for range estimation. Throughout this chapter, two

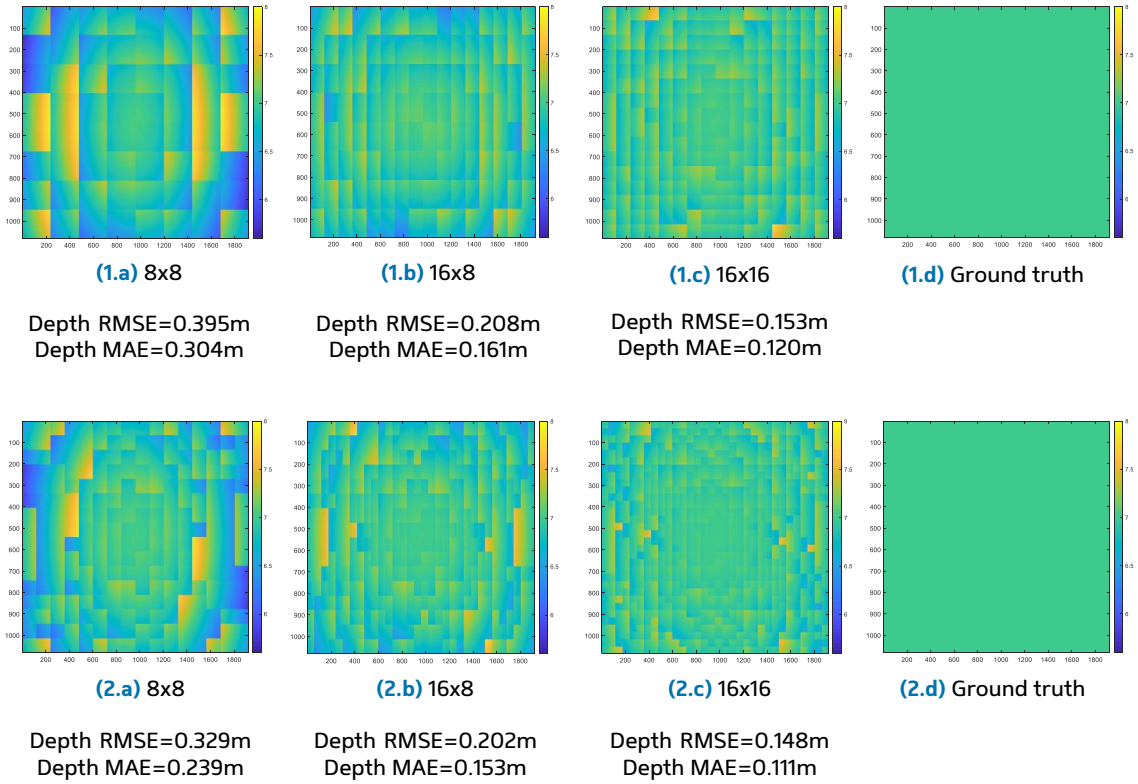
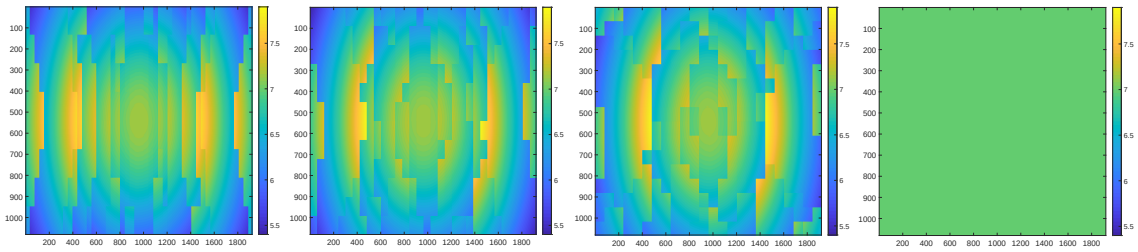


Figure 3.14: The Depth Maps for the One Wall Scenario Are Depicted for Different Antenna Configurations and Codebook Resolutions, for a Separation Distance of 7 Meters. Figures (a), (b), and (c) Illustrate the Estimated 1080p Maps for 8×8, 16×8, and 16×16 UPAs. Figures (d) Illustrate the Ground Truth Maps. The Top Maps Are with No Codebook Oversampling While the Bottom Maps Are with Codebook Oversampling Factors of Two.

performance metrics are used: (i) root-mean-square-error (RMSE) between the estimated map and the ground truth map to indicate the standard deviation of the estimation error, and (ii) mean-absolute-error (MAE) to denote the expected value of the estimation error. The two metrics are defined in (3.9). Next, we evaluate the performance of our proposed mmWave MIMO depth estimation approach in four main scenarios: (i) A one wall scenario in Section 3.9.2, (ii) a two walls scenario in Section 3.9.3; (iii)) a room with two pillars scenario in Section 3.9.4, and (iv) a conference room scenario in Section 3.9.5.



(a) 12×2 UPA (b) 8×3 UPA (c) 6×4 UPA (d) Ground Truth

Depth RMSE $= 0.505$m Depth RMSE $= 0.512$m Depth RMSE $= 0.473$m

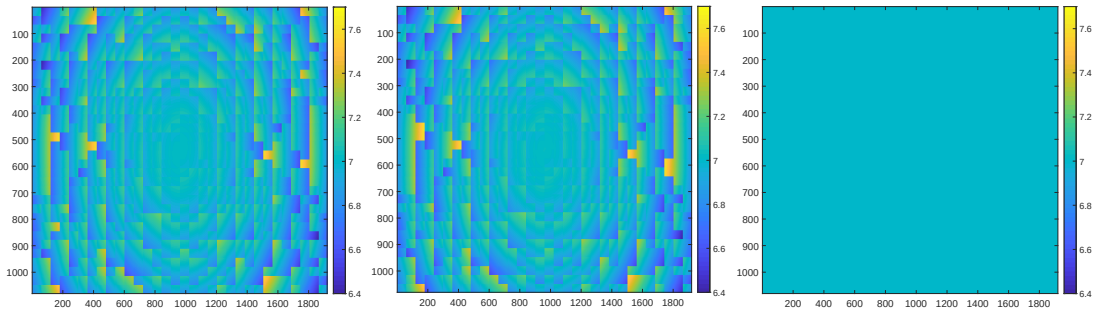
Depth MAE $= 0.392$m Depth MAE $= 0.397$m Depth MAE $= 0.372$m

Figure 3.15: The 1080p Depth Maps for the One Wall Scenario Are Depicted at Different Antenna Configurations, for a Separation Distance of 7 Meters. The Same Number of Antenna Elements Is Used (24 Elements) and Codebook Oversampling Factors of Four Are Employed. Figures (a), (b), and (c) Illustrate the Estimated Maps for 12×2, 8×3, and 6×4 UPAs. Figure (d) Illustrates the Ground Truth Depth Map.

3.9.2 One Wall Scenario

The one wall scenario consists of an AR/VR transceiver facing a wall in free space propagation. Unless otherwise mentioned, the separation distance between the wall and the transceiver is 7 meters and the wall building material is concrete. In Fig. 3.13, we show the estimated range and depth maps for the one wall scenario compared to the ground truth maps. Fig. 3.13(a) and Fig. 3.13(b) show that the range map estimation error has an average MAE of 0.098m and RMSE of 0.127m. Further, the depth map estimation error Fig. 3.13(c) and Fig. 3.13(d) has an average MAE of 0.12m and RMSE of 0.153m. Overall, these figures show that the proposed approaches can accurately estimate the range/depth maps for a wall at 7m distance from the AR/VR device with around 10cm error, which highlights the effectiveness of this approach.

Impact of the important system parameters: Next, we briefly evaluate the impact of the various system parameters on the performance of the proposed mmWave depth map estimation solution.



(a) Continuous Phase Shifts
Depth RMSE = 0.148m
Depth MAE = 0.1106m

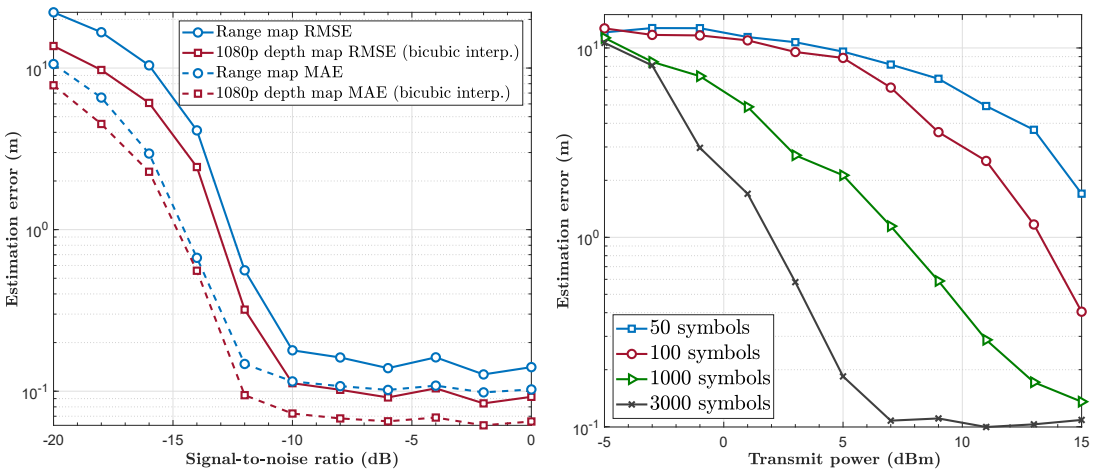
(b) 2-bit Phase Shifts
Depth RMSE = 0.149m
Depth MAE = 0.1111m

(c) Ground Truth

Figure 3.16: The 1080p Depth Maps for the One Wall Scenario at 7m Separation Distance Are Estimated for Two Cases of the RF Phase Shifters at the AR/VR Device: (a) Continuous Phase Shifts and (b) 2-bit Quantized Phase Shifts. 16×16 UPA Is Employed with Codebook Oversampling Factors of Two. Figure (c) Illustrates the Ground Truth Depth Map.

- **Number of antennas and sensing codebook beams:** In Fig. 3.14, we plot the estimated and ground-truth depth maps for a different number of antennas and codebook oversampling factors. As illustrated, the depth estimation accuracy can generally improve by increasing the number of antennas and/or the codebook oversampling factors. This comes with the cost of deploying more antennas at the AR/VR device or employing more beams, which translates to a longer sensing time. In Fig. 3.15, we plot the estimated and ground-truth depth maps for different antenna configurations using the same number of antenna elements. As depicted, the depth estimation accuracy depends on the UPA configuration, with the best configuration being the 6×4 UPA because of its closeness to the 1080p aspect ratio.

- **RF phase shift quantization:** As previously described in Section 3.7.2, the phase quantization of the RF phase shifters in the AR/VR transceiver architecture produces a noticeable change in the radiation pattern shape of the sidelobes.



(a) Signal-to-noise Ratio (b) Transmit Power

Figure 3.17: For the One Wall Scenario, the Error Performance of the Proposed mmWave MIMO Based Depth Estimation Solution Is Evaluated under Different Error Metrics in (a) and Is Evaluated for Different Preamble Sequence Lengths in (b). The Wall Is 7 Meters Away from theAR/VR Device with 16×16 UPAs. The Figures Show the Robustness of the Developed Approach under a Relatively Low SNR Regime. Note That the Displayed Transmit Power Range in (b) Corresponds to an Average SNR Range of -20.7dB to -0.7dB.

To examine the effect of this phase quantization on the estimated depth maps, Fig. 3.16 shows the comparison of the estimated depth maps for two cases of the RF phase shifters at the AR/VR device: (a) continuous phase shift and (b) 2-bit quantized phase shifts. As depicted, the phase quantization contributes with a small negative impact on the depth map estimation accuracy for the one wall scenario at a separation distance of 7 meters.

- **Transmit sensing power:** In Fig. 3.17(a), we investigate the effect of changing the transmit power on the depth map estimation accuracy. The SNR value of 0dB corresponds to a transmit power of 15dBm. This figure shows that a transmit power of 5dBm (SNR of -10dB) could be sufficient to reach around 10cm error for the depth estimation accuracy.

- **Preamble sequence length**: The estimation error versus transmit power is de-

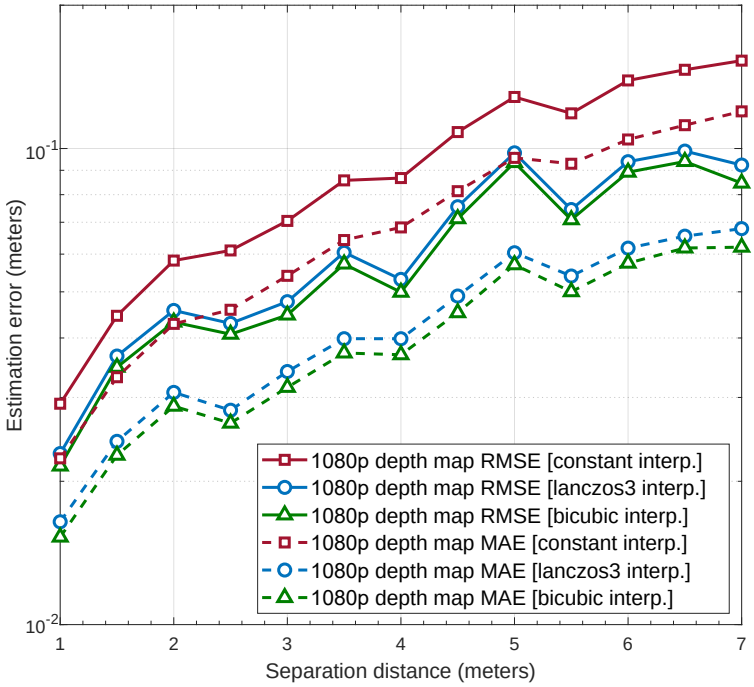


Figure 3.18: The Error Performance of the Proposed mmWave MIMO Based Depth Estimation Solution Is Evaluated Across Different Separation Distances for the One Wall Scenario. The Estimation Error Starts from $\approx$ 1.5m at a 1m Distance and Reaches Around 10cm at a 7m Distance.

picted in Fig. 3.17(b) for different values of preamble sequence lengths, namely preambles with 50, 100, 1000, and 3000 symbols. As shown in this figure, increasing the preamble sequence length improves the depth estimation accuracy at the expense of increased sensing time and post-processing complexity.

- **Separation distance between the AR/VR device and the facing wall:** Fig. 3.18 investigates the impact of increasing the depth value on the depth estimation accuracy. As shown in this figure, the larger the distance between the AR/VR device and the facing surface, the larger the error in the depth estimate, which is expected. This figure also highlights some advantage for the bicubic interpolation compared to the other interpolation methods.

- **The surface material:** Now, we evaluate the performance of the proposed ap-

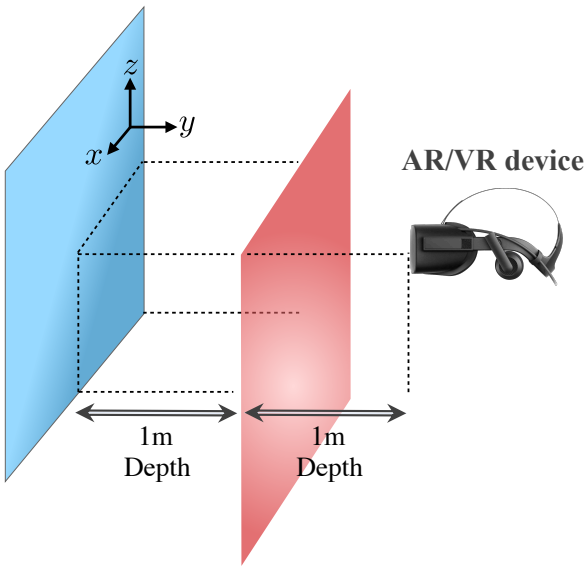


Figure 3.19: The Adopted Two Walls Scenario Is Illustrated.

proach for different surface materials. More specifically, we summarize in Table 3.2 the range map MAE for different candidates of the wall material. Overall, we can notice some correlation between the estimation accuracy and the *scattered to incident power ratio* property of the materials, which are summarized in Table 3.1.

3.9.3 Two Walls Scenario

The two walls scenario consists of one AR/VR device facing two walls in free space propagation as depicted in Fig. 3.19. The separation distance between the front wall and the AR/VR device is 1m while the separation between the back wall and the AR/VR device is 2m. The walls' building material is concrete. Each wall consists of $2,048$ faceted faces, and each face contributing with at most one backscattered ray. The purpose behind studying this scenario is to test the alignment of the estimated map compared to the ground truth depth map. The results of this test are illustrated

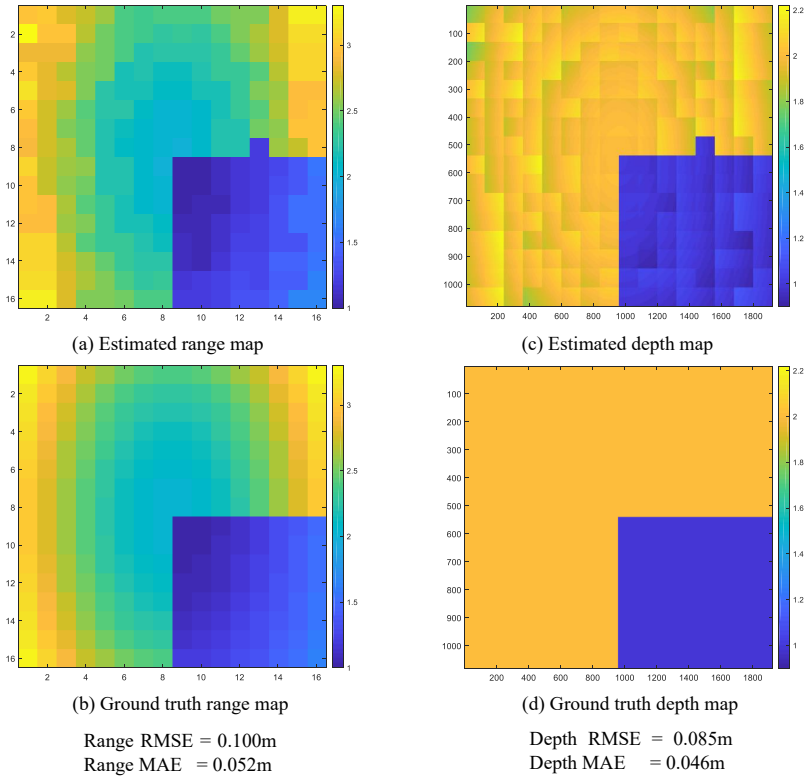
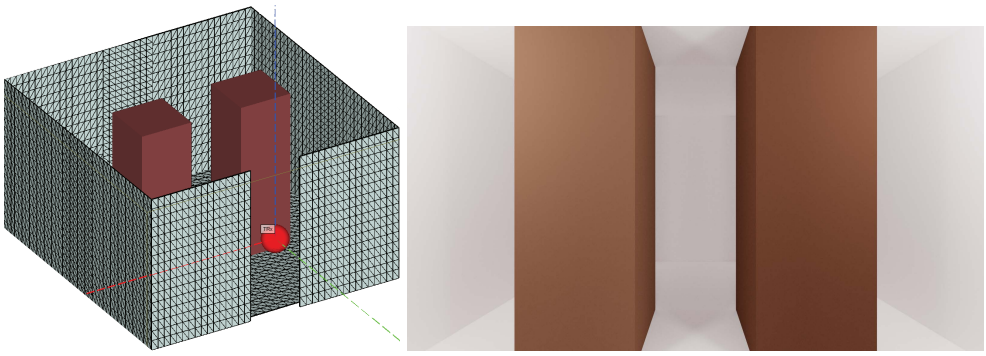


Figure 3.20: The Maps for the Two Walls Scenario Are Depicted. The AR/VR Device Is Employed with 16×16 UPAs. The Depicted Maps Are the Estimated Maps (at the Top), Ground Truth Maps (at the Bottom), Range Maps (on the Left Side), and 1080p Depth Maps (on the Right Side). Comparing (a) with (b), the Range Map Estimation Error: MAE $= 0.052$m. Comparing (c) with (d), the Depth Map Estimation Error: MAE$= 0.046$m.

in Fig. 3.20, where the estimated range and depth maps are compared to the ground truth maps. As shown in Fig. 3.20, the two edges of the front wall in the estimated maps align reasonably well with the one displayed in the ground truth maps. This highlights the promising performance of proposed mmWave based depth estimation solution.

3.9.4 A Room with Two Pillars

In this scenario, we consider a 5m×5m room where one AR/VR device is centered at the front door of the room, as depicted in Fig. 3.21. The room consists of a concrete floor plan with two wood pillars in the middle of the room. The wood pillars are at

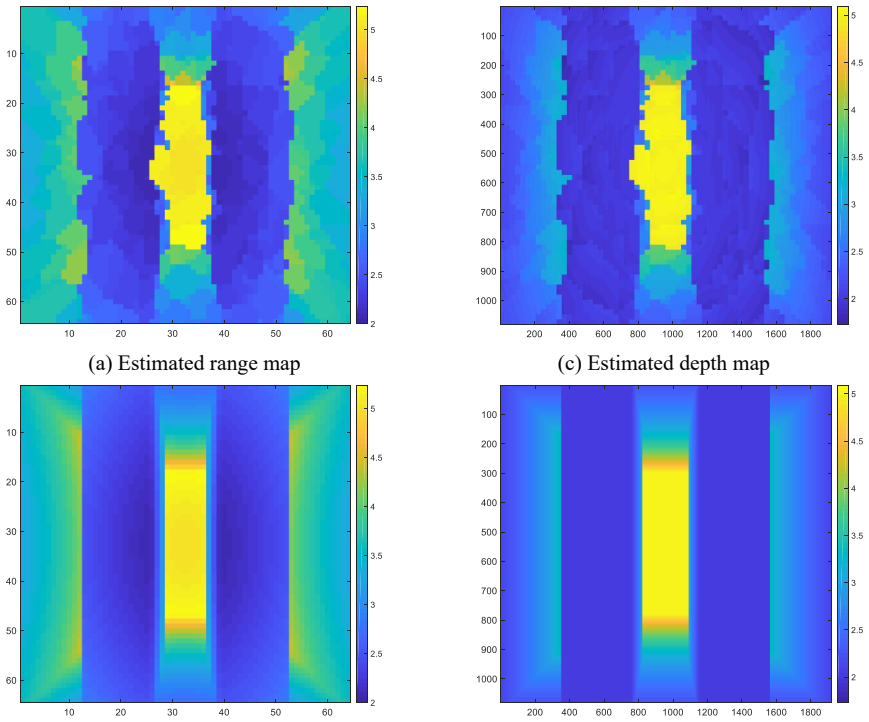


(a) The Room with Two Pillars (b) The Room Scene from the Door Position

Figure 3.21: Figure (a) Illustrates the Bird View of the Room with Two Pillars. Figure (b) Shows the Scene from the AR/VR Device Position, Centered at the Front Door. The 5m×5m Room Consists of a Concrete Floor Plan with Two Wood Pillars in the Middle of the Room. The Wood Pillars Are at 2 Meters Distance from the AR/VR Device.

2 meters distance from the AR/VR transceiver. The floor plan consists of $15,488$ faceted faces whereas each of the wood pillars consists of $3,072$ faceted faces. Note that the ceiling of the floor plan is set to the invisible mode for visibility purposes only. For the estimation error assessment of the indoor space scenario, Fig. 3.22 shows the comparison between estimated and ground truth maps for 16×16 UPA antennas with a codebook oversampling factors of four in both azimuth and elevation dimensions.

First, Fig. 3.22(a) with Fig. 3.22(b) show the estimate and ground truth range maps, which have a MAE of 0.139m and RMSE of 0.355m. For the depth maps, Fig. 3.22(c) with Fig. 3.22(d) represent 1080p maps with estimation error of(i) 0.126m for the MAE and 0.356m for the RMSE with nearest neighbor interpolation, and (ii) 0.123m for the MAE and 0.328m for the RMSE with bicubic interpolation. From observing the difference in maps, the mmWave reasonably recover most of the depth information of the scene with low codebook resolution (16×16) compared to the ground truth 1080p resolution. With narrower transmit and receive beams, i.e. more antenna elements, the estimation accuracy is expected to further improve.



(a) Estimated range map

(c) Estimated depth map

(b) Ground truth range map

(d) Ground truth depth map

Range RMSE = 0.355m
Range MAE = 0.139m

Depth RMSE = 0.356m
Depth MAE = 0.126m

Figure 3.22: The Maps for the Room with Two Pillars Are Depicted. 16×16 UPAs Are Employed with Codebook Oversampling Factors of Four. The Depicted Maps Are the Estimated Maps (at the Top), Ground Truth Maps (at the Bottom), Range Maps (on the Left Side), and 1080p Depth Maps (on the Right Side). Comparing (a) with (b), the Range Map Estimation Error: MAE $= 0.139$m. Comparing (c) with (d), the Depth Map Estimation Error: MAE $= 0.126$m.

The depth map estimation accuracy for this scenario is also evaluated at different SNRs in Fig. 3.23. In this figure, we adopt the model and system parameters used in Fig. 3.22 with 16×16 UPAs and oversampling factors of four. It is also worth mentioning that 0dB SNR corresponds to -20dBm transmit power in our setup. As shown in Fig. 3.23, the estimated depth maps have MAE of almost 10cm at 0dB, which highlights the promising performance of our proposed depth map estimation approach at relatively low SNRs and in an indoor room with several surfaces and different materials. This will be further emphasized in the following subsection.

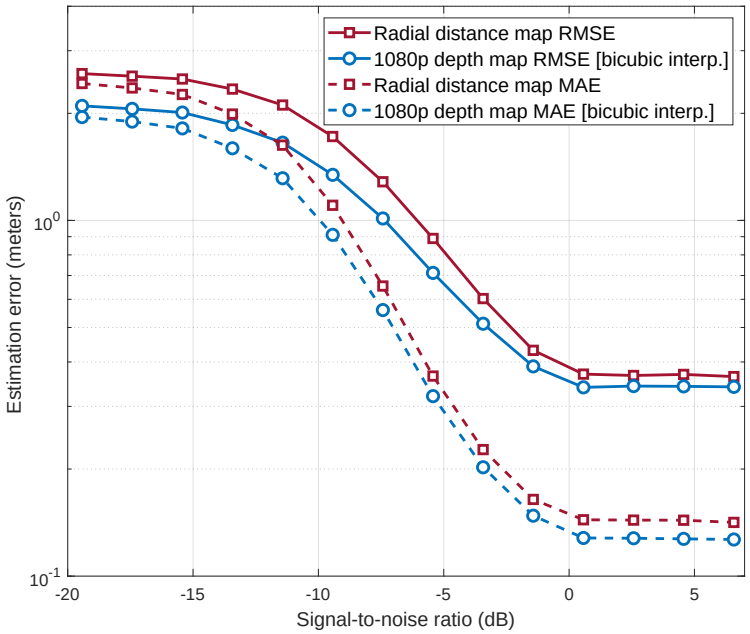


Figure 3.23: For the Room with Two Pillars, the Error Performance of the Proposed mmWave MIMO Based Depth Estimation Is Evaluated for Different Error Metrics. 16×16 UPAs Are Employed with a Codebook Oversampling Factors of Four in Both Dimensions. This Figure Shows the Robustness of the Proposed mmWave MIMO Based Depth Estimation under a Relatively Low SNR Regime.

3.9.5 Conference Room Scenario

In this scenario, we consider the conference room shown in Fig. 3.24. The ceiling of the indoor space is set to the invisible mode for visibility purpose only. The 10m$\times$10m indoor space has a 6m$\times$6m conference room with glass walls. The indoor space walls are made of layered drywall, the ceiling is made of ceiling board, and the floor is made of floorboard. The conference room chairs and tables are made of wood. The conference room door opening is 1m in width and 2.7m in height. The number of facets for each item in the indoor space is as follows: $2,048$ facets for the layered drywall, $2,048$ facets for the floorboard, and $2,048$ facets for the ceiling board. In addition, the number of facets for each item in the conference room is as follows: $1,568$ facets for the glass wall, $4,446$ facets for the table, $21,192$ facets for the office chairs. The conference room scenario consists of two AR/VR devices for two scenes

(a) Conference room scenario

(b) First scene

(c) Second scene

Figure 3.24: (a) the Bird View of the Conference Room Scenario; (b) and (c) the Scenes under Study. The 10m×10m Indoor Space Contain a 6m×6m Conference Room in Glass. The Indoor Space Walls Are Made from Layered Drywall, the Ceiling Is Made from Ceiling Board and the Floor Is Made from Floorboard. The Conference Room Chairs and Tables Are Made from Wood.

under study — the first device is centered at the front door of the conference room while the second transceiver is placed outside of the conference room facing the other glass facet. The scenes captured by the AR/VR camera for the two cases are shown in Fig. 3.24(b) and Fig. 3.24(c).

One main motivation for leveraging mmWave MIMO to estimate the depth maps (compared to RGB based depth estimation approaches) is the expected higher efficiency in detecting transparent and dark objects. In Fig. 3.25, we compare our mmWave MIMO based depth estimation approach with the RGB based depth estimation approach, detailed in [8], for the two considered conference room scenarios. It's worth emphasizing here that the algorithms in [8] achieve considerably good depth accuracy when tested on the NYU depth V2 dataset [99]. As shown in Fig. 3.25, the mmWave MIMO based estimator outperforms the RGB based estimator in recogniz-

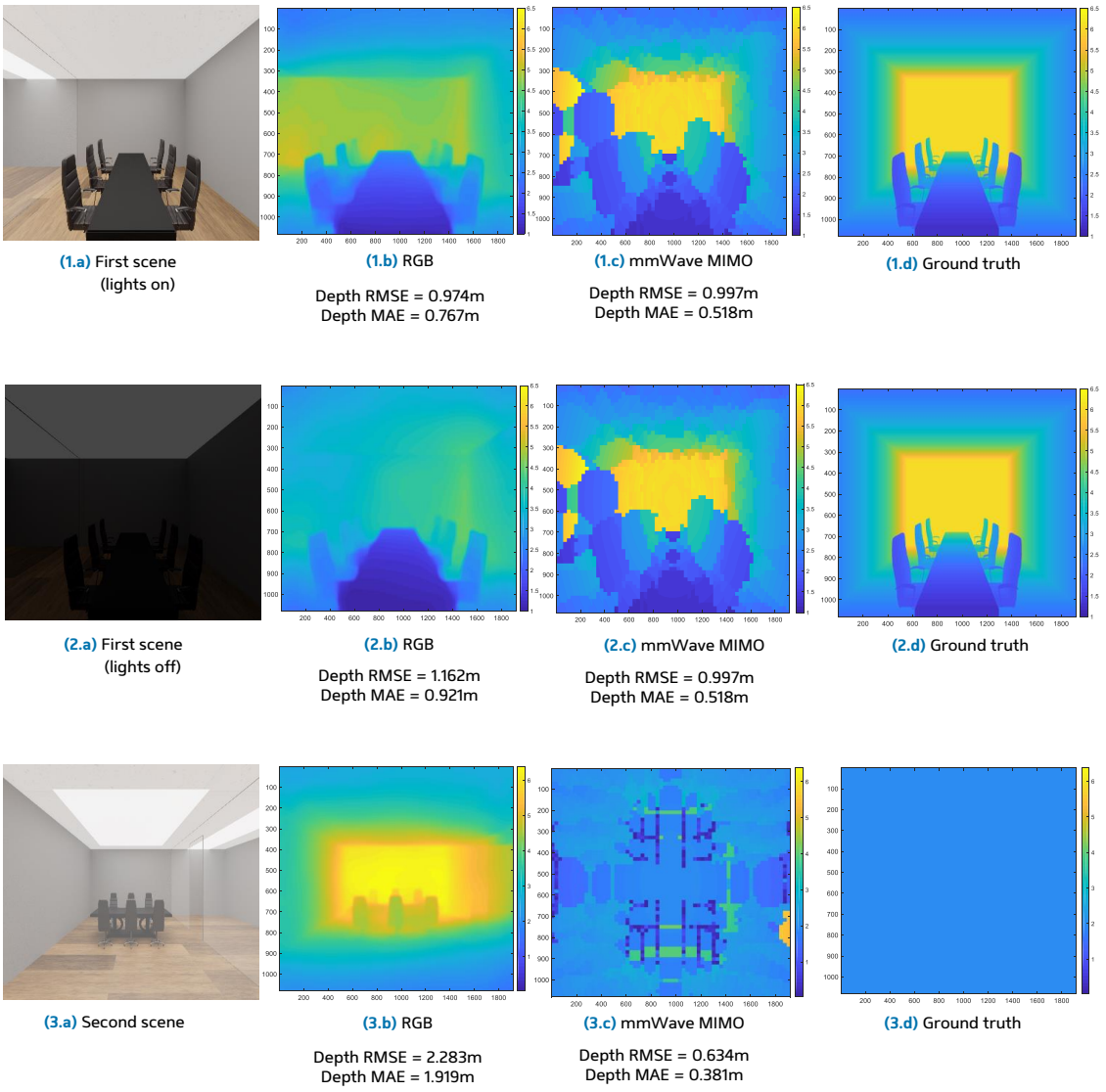


Figure 3.25: For the Conference Room Scenario, the Proposed mmWave MIMO Based Depth Estimation Is Compared with the RGB Based Depth Estimation in [8]. 16×16 UPAs Are Employed with Codebook Oversampling Factors of Four. The Depicted Maps Are the Maps of the First Scene with Lights On/Off (the Top Two Rows) and the Second Scene (the Bottom Row). (a) the Scenes under Study; (b) the Estimated Maps from Monocular RGB Images; (c) the Estimated Maps from Our Proposed Solution; (d) the Ground Truth Depth Maps.

ing transparent and dark objects. For the first scene, the glass wall was not detected by the RGB estimator. Also, in the presence of a scene with low illumination, the mmWave MIMO based estimator performance shows robustness in the estimation accuracy compared to the RGB based estimator. Figure 1.c) and 2.c) were generated with the aid of the SLR approach in Section 3.7.2, with $\delta_{\mathrm{H}} = 2, \delta_{\mathrm{V}} = 3$. For this reason, the depth maps constructed by the mmWave MIMO system seem coarser than the one constructed by RGB cameras, which can be resolved using morphological image processing operations, e.g., the erosion operation. As for the second scene, the RGB based estimator is unable to detect the transparent glass compared to the mmWave MIMO based estimator. Interestingly, despite the fact that the glass scattering ratio is 0% based on Table 3.1, the conference room glass wall is partially recovered by the mmWave MIMO based estimator because of the boresight reflection path. This makes the wireless AR/VR experience safer by providing the ability to detect transparent surfaces. All these promising results highlight the potential of leveraging the proposed mmWave MIMO based depth map estimation approaches for immersive AR/VR experience.

3.10 Conclusion

In this chapter, we considered the problem of estimating accurate depth maps for AR/VR devices, which is an essential goal for immersive mixed-reality experience. For this problem, we proposed leveraging the mmWave communication systems that are deployed on the AR/VR devices to estimate and build high-resolution depth maps. We formulated the communication-constrained depth map sensing problem and proposed a comprehensive framework for realizing this objective. The proposed framework includes (i) the construction of depth map specific sensing codebooks using practical mmWave antenna arrays and (ii) the development of efficient post-processing

solutions for jointly processing the receive signals from the multiple sensing beams and estimating high-resolution depth maps. Simulations using accurate 3D ray-tracing models confirmed the promising accuracy of our proposed mmWave based depth map estimation approach in various environment scenarios. In particular, the results show that the proposed approach can construct relatively high-resolution depth maps with less than 10cm error using practical mmWave systems. This highlights the potential of leveraging this solution to complement RGB-D based depth maps and realize immersive depth perception for wireless virtual/augmented reality systems.

Chapter 4

RECONFIGURABLE INTELLIGENT SURFACE AIDED WIRELESS SENSING FOR SCENE DEPTH ESTIMATION

4.1 Abstract

Current scene depth estimation approaches mainly rely on optical sensing, which carries privacy concerns and suffers from estimation ambiguity for distant, shiny, and transparent surfaces/objects. Reconfigurable intelligent surfaces (RISs) provide a path for employing a massive number of antennas using low-cost and energy-efficient architectures. This has the potential for realizing RIS-aided wireless sensing with high spatial resolution. In this chapter [1] , we propose to employ RIS-aided wireless sensing systems for scene depth estimation. We develop a comprehensive framework for building accurate depth maps using RIS-aided mmWave sensing systems. In this framework, we propose a new RIS interaction codebook capable of creating a sensing grid of reflected beams that meets the desirable characteristics of efficient scene depth map construction. Using the designed codebook, the received signals are processed to build high-resolution depth maps. Simulation results compare the proposed solution against RGB-based approaches and highlight the promise of adopting RIS-aided mmWave sensing in scene depth perception.

[1] This chapter is based on the work submitted to IEEE and published in the preprint paper: A. Taha, H. Luo, and A. Alkhateeb, "Reconfigurable Intelligent Surface Aided Wireless Sensing for Scene Depth Estimation," in arXiv preprint arXiv:2211.08210, Nov. 2022. [Online]. Available: https://arxiv.org/abs/2211.08210. This work was supervised by Prof. Ahmed Alkhateeb. Hao Luo provided important ideas for the reconfigurable intelligent surface aided wireless sensing design that greatly improved the work.

4.2 Introduction

Because of their promising coverage and spectral efficiency gains [28, 100, 101], the use of reconfigurable intelligent surfaces (RISs) is envisioned as a key enabler for next-generation communication systems. These surfaces comprise massive numbers of nearly passive elements that interact with the incident signals in a smart way to improve the performance of such systems. RISs have recently started gaining interest in improving some of the wireless sensing systems [102–108], with no application yet in scene depth estimation. Current scene depth estimation approaches are mainly rooted in optical sensing. While optical sensors can generally provide good accuracy, they suffer from some critical limitations. These limitations stem from the fundamental properties of the way light propagates and interacts with the elements of an environment. The accuracy of optical sensors normally degrades in scenarios with unfavorable light conditions, in the presence of shiny, dark, or transparent objects/surfaces, and in the presence of non-line-of-sight (NLoS) objects/surfaces. Optical sensors also suffer from key privacy concerns and range/velocity estimation ambiguity for distant objects/surfaces.

To overcome these limitations, millimeter-wave (mmWave) wireless sensing is a promising technology for complementing optical sensors in accurately sensing the environment. mmWave signal propagation is not affected by interference from light sources. These signals exhibit different propagation properties that can aid in recognizing transparent, shiny, dark, and distant objects/surfaces. In addition, wireless sensing systems have fewer privacy concerns and can be well integrated with the wireless communication framework [109].

mmWave wireless sensing has been investigated in the literature for various sensing problems. For instance, in [110], an imaging algorithm using frequency-modulated

continous-wave (FMCW) mmWave radar is developed by leveraging the conditional generative adversarial networks. In [31], a mmWave MIMO based sensing framework is developed for estimating scene depth maps, under the constraints of a mmWave communication system. In [111], a 3D body reconstruction dataset is built with the proposed automatic annotation system. The effectiveness of the mmWave radar based body reconstruction is demonstrated by a point-cloud based algorithm. Scaling mmWave MIMO antenna arrays, however, is associated with large computational/hardware complexity and energy consumption. This limitation poses a critical challenge in scaling the spatial resolution, which motivates leveraging RISs to assist mmWave wireless sensing systems. In addition, RIS aided sensing systems can filter out in the signal reception more undesired channel paths than the ones filtered out by mmWave MIMO based sensing systems, as will be discussed in detail later.

RIS-aided sensing systems is gaining interest in the literature. Several RIS-aided sensing systems have been studied to improve the sensing performance in target detection [102–104]. In [102], RIS is used to extend the coverage of the radar surveillance in non-line-of-sight (NLOS) scenarios. In [103], an RIS-aided radar system with multiple targets is proposed, where the radar waveforms and the phase shifts of the RIS are jointly optimized. In [104], a general signal model for RIS-aided target detection is studied by considering monostatic, bistatic, LOS, and NLOS scenarios. In addition, RISs can be leveraged in addressing wireless imaging challenges [105–108].

In [105, 106], RIS-aided microwave imaging systems are proposed, where the image of the targets can be reconstructed from the receive signals. In [107], the authors propose a WiFi-based RIS-aided imaging system, and the beamforming of the RIS is designed for sensing all regions of interest. In [108], an RIS-aided RF sensing system for semantic segmentation is proposed. The semantic recognition is conducted based on the point cloud of the objects, which is reconstructed from the receive

signals. To the best of our knowledge, RIS-aided sensing systems have not yet been investigated for scene depth estimation. Accurate scene depth perception can enable some key emerging applications, including augmented and virtual reality (AR/VR) and automotive vehicles among others.

In this chapter, we investigate the RIS aided wireless sensing based scene depth estimation problem. The contributions of this chapter can be summarized as follows.

- *RIS sensing based scene depth estimation framework:* We formulate the RIS wireless sensing based scene depth estimation problem and propose a comprehensive framework for building scene depth maps using RIS aided wireless sensing systems.

- *Depth map suitable RIS sensing codebook:* We propose a novel RIS interaction codebook design capable of creating a sensing grid of reflected beams that meets the desirable characteristics of efficient scene depth map construction. Given the designed RIS interaction codebook, we develop a post-processing solution on the receive signals to build high-resolution depth maps.

Based on accurate 3D ray-tracing Wireless InSite [1] channels and ground truth Blender [112] depth maps, the simulation results show the promise of adopting RIS aided mmWave sensing for scene depth estimation.

4.3 System and Channel Models

In this section, we present the adopted system and channel models for RIS aided wireless sensing systems.

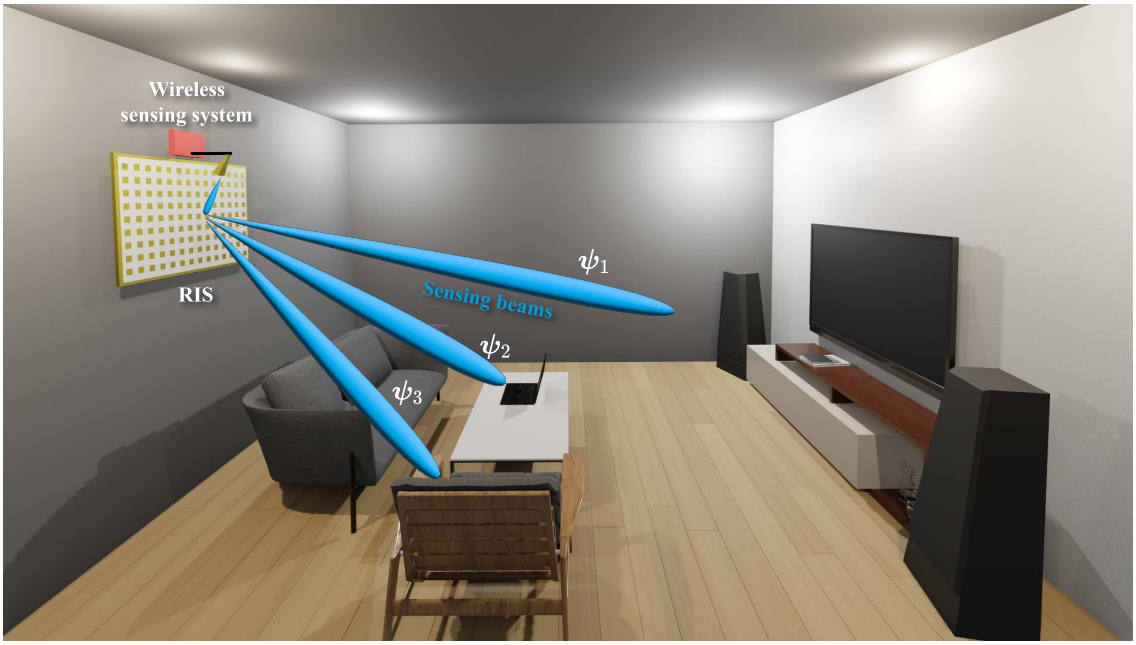


Figure 4.1: The RIS-Aided Wireless Sensing System Is Shown. The Sensing Signals Are Transmitted to the RIS Through a Feeding Antenna. The RIS Then Reflects the Incident Signals to the Environment. The Backscattered/Reflected Signals Are Then Reflected by the RIS Back to the Sensing System, Using a Sensing Codebook, for Depth Perception.

4.3.1 System Model

In this chapter, we adopt a reconfigurable intelligent surface (RIS) aided mmWave wireless sensing system, as shown in Fig. 4.1. The sensing system consists of a transmitter and a receiver; both are connected through a self-isolation circuitry [31] to a shared single antenna, for ease of exposition. This single antenna acts as a feeding antenna that illuminates the RIS for sensing purposes. The proposed solution and the results in this chapter can be extended though to multi-antenna sensing transceivers. The RIS is equipped with N reconfigurable elements, where each element can be modeled as a phase shifter. Denote the RIS interaction (reflection beamforming) matrix by $\boldsymbol{\Psi} = \text{diag}\left(\boldsymbol{\psi}\right) \in \mathbb{C}^{N\times N}$, where $\boldsymbol{\psi} = \left[\mathrm{e}^{\jmath\phi_1}, \ldots, \mathrm{e}^{\jmath\phi_N}\right]^T$ is the interaction vector with unit modulus entries. The phase shift induced by the n^{th} RIS elements on the incident signals is represented by ϕ_n, $n \in \{1, \ldots, N\}$.

The sensing process proceeds as follows: (a) the sensing system transmits sensing signals to the RIS; (b) the RIS reflects these signals towards the surrounding environment, which contains G_{tar} mobile targets; (c) the signals are reflected back to the surface by the targets; (d) the RIS reflects back these incident signals to the sensing system; (e) the sensing system processes the receive signals to achieve a sensing objective.

In this chapter, our sensing objective is to estimate the depth map of the environment. For that objective, we make the following assumptions: (a) The RIS elements are not mutually correlated; (b) the channel between the sensing system and the RIS is in the near field region whereas the channel between the RIS and the targets is in the far field region; (c) the channel between the sensing system and the targets is neglected, assuming the feeding antenna radiation pattern is directional towards the RIS; (d) the RIS interaction is reciprocal when interchanging the incident signal directions with the reflected signal directions. Next, we describe the transmit and receive signal models and channel model in detail.

Transmit Signal Model: The adopted sensing system is a wideband frequency-modulated continous-wave (FMCW) radar transceiver, with a complex-baseband architecture, as detailed in [113]. Let the FMCW radar transmit signal be a radar frame, which consists of a sequence of M_{chirp} repeated chirp signals with a chirp repetition interval of T_{PRI} seconds. Let $a_{\text{BP}}(t) \in \mathbb{R}$ be the real-valued bandpass transmit signal of a single chirp, with a duration of T_{active} seconds, a transmission bandwidth of $\text{BW} = ST_{\text{active}}$, a chirp slope of S, and a starting chirp frequency of f_0. The time-varying frequency of the transmit signal is $f(t) = f_0 + St, 0 \leq t \leq T_{\text{active}}$. The time-varying phase can then be expressed as $\chi(t) = 2\pi \int_0^t (f_0 + S\eta)\, d\eta = 2\pi f_0 t + \pi S t^2$. The signal $a_{\text{BP}}(t)$ can

then be formulated as

$$a_{\mathrm{BP}}(t) = \begin{cases} \cos\left(2\pi f_0 t + \pi S t^2\right) & 0 \leq t \leq T_{\mathrm{active}}, \\ 0 & \text{otherwise.} \end{cases} \tag{4.1}$$

The real-valued bandpass transmit signal of a single radar frame, $x_{\mathrm{BP}}(t)$, can then be formulated as

$$x_{\mathrm{BP}}(t) = \sqrt{\mathcal{E}_{\mathrm{T}}} \sum_{c=0}^{M_{\mathrm{chirp}}-1} a_{\mathrm{BP}}(t - cT_{\mathrm{PRI}}) \tag{4.2}$$

$$= \mathrm{Re}\left(x(t)\, \mathrm{e}^{\jmath 2\pi f_0 t}\right), t \in \mathbb{R}_{\geq 0}, \tag{4.3}$$

where $\mathcal{E}_{\mathrm{T}}$ is the transmit signal energy and $x(t) \in \mathbb{C}$ is the complex-valued lowpass-equivalent transmit signal.

Receive Signal Model: For the channel model, we adopt the extended Saleh-Valenzuela wideband geometric channel model [31]. After the transmit signal travels through the bandpass channel and experiences additive noise at the receiver, the receive bandpass signal $y_{\mathrm{BP}}(t) = \mathrm{Re}(y(t)\mathrm{e}^{\jmath 2\pi f_0 t})$ can be modeled in terms of its complex-valued lowpass-equivalent signal $y(t) \in \mathbb{C}$, which can be defined as

$$y(t) = x(t) * h(t) + w(t) = \sum_{g=1}^{G_{\mathrm{tar}}} \sum_{\ell=1}^{L_g} \mathsf{h}_{g,\ell}(t) x(t - \xi_{g,\ell}(t)) + w(t), \tag{4.4}$$

where $h(t) \in \mathbb{C}$ is the complex-valued lowpass-equivalent channel. $w(t) \sim \mathcal{N}(0, \sigma_w^2)$, $w(t) \in \mathbb{C}$ is the receive noise with variance σ_w^2. L_g is the number of channel paths interacting with the g^{th} target. $\mathsf{h}_{g,\ell} \in \mathbb{C}$ is the complex-valued channel path gain of the ℓ^{th} channel path of the g^{th} target, which is modeled in detail in Section 4.3.2.

The propagation delay is denoted by $\xi_{g,\ell}(t) \in \mathbb{R}$, which can be formulated as

$$\xi_{g,\ell}(t) = \frac{R_{g,\ell}(t)}{\varsigma} = \frac{R_{0,g,\ell}}{\varsigma} + \frac{\nu_{g,\ell} t}{\varsigma} + \frac{\mathsf{a}_{g,\ell} t^2}{2\varsigma}, \tag{4.5}$$

where ς is the speed of light. $R_{g,\ell}(t)$ is the time-varying total propagation distance at time t, traveled by the ℓ^{th} channel path of the g^{th} target (with one or multiple

interactions with the environment), starting from the radar transmitter and ending at the radar receiver. $R_{0,g,\ell}$ is the initial propagation distance observed at the sensing start time, at $t = 0$. $\nu_{g,\ell}$ and $\mathsf{a}_{g,\ell}$ represent the Doppler velocity and acceleration of the ℓ^{th} channel path contributed by the g^{th} target, respectively.

To construct the receive baseband signal [113], the receive signal $y_{\text{BP}}(t)$ is first mixed with two versions of the transmit signal $x_{\text{BP}}(t)$, one with a -90° phase shift difference. Then, the outputs of the mixers pass through low-pass filters and analog-to-digital converters (ADCs) to generate the in-phase signal $I[s, c]$ and the quadrature-phase signal $Q[s, c]$, for the ADC sample $s \in \mathcal{S}, \mathcal{S} = \{0, 1, \ldots, (M_{\text{sample}} - 1)\}$, and for the chirp $c \in \mathcal{C}, \mathcal{C} = \{0, 1, \ldots, (M_{\text{chirp}} - 1)\}$. M_{sample} is the number of ADC samples per chirp. Let $b[s, c]$ denotes the discrete-time equivalent of a continuous-time signal $b(t)$, sampled at time $t = sT_{\text{S}} + cT_{\text{PRI}}$, $T_{\text{S}} = 1/F_{\text{S}}$. T_{S} and F_{S} are the ADC sampling time period and the ADC sampling frequency, respectively. The receive baseband digital signal, $z[s, c] = I[s, c] + \jmath\, Q[s, c]$, can be formulated as

$$z[s,c] = \left(\sum_{g=1}^{\mathsf{G}_{\text{tar}}} \sum_{\ell=1}^{L} \sqrt{\rho_{g,\ell}[s,c]}\, \mathrm{e}^{-\jmath \vartheta_{g,\ell}[s,c]}\, \mathrm{e}^{+\jmath\, \Xi_{g,\ell}[s,c]} \right) + w[s,c] \mathrm{e}^{\jmath \chi[s]}, \tag{4.6}$$

where $\chi[s] = 2\pi f_0 t_{\text{fast}} + \pi S t_{\text{fast}}^2$ and $t_{\text{fast}} = sT_{\text{S}}$. The channel path receive power and phase are $\rho_{g,\ell}[s, c] = \mathcal{E}_{\text{T}} |\mathsf{h}_{g,\ell}[s, c]|^2$ and $\vartheta_{g,\ell}[s, c] = \arg\left(\mathsf{h}_{g,\ell}[s, c]\right)$, respectively. The phase term $\Xi_{g,\ell}[s, c]$ contains range and Doppler information of the targets, which can be defined as

$$\Xi_{g,\ell}[s,c] = 2\pi \big(f_0 \xi_{g,\ell}[s,c] + S t_{\text{fast}} \xi_{g,\ell}[s,c] - \tfrac{S}{2} \xi_{g,\ell}^2[s,c] \big). \tag{4.7}$$

The discrete time varying propagation delay is denoted by $\xi_{g,\ell}[s, c]$, which can be formulated as

$$\xi_{g,\ell}[s,c] = \frac{R_{g,\ell}[s,c]}{\varsigma} = \frac{R_{0,g,\ell} + \nu_{g,\ell} t_{\text{slow}} + \frac{\mathsf{a}_{g,\ell}}{2} t_{\text{slow}}^2}{\varsigma}, \tag{4.8}$$

$$t_{\text{slow}} = t_{\text{fast}} + cT_{\text{PRI}} = sT_{\text{S}} + cT_{\text{PRI}}, \tag{4.9}$$

where t_{slow} represents the discrete time delay of each ADC sample in each chirp out of the M_{chirp} chirps. Next, we describe the time-varying complex-valued channel gain model $\mathsf{h}_{g,\ell}(t)$.

4.3.2 Channel Model

For RIS aided radar channel modeling, we adopt and extend on the channel model of the non-line-of-sight monostatic radar configuration detailed in [104]. Different from the model in [104], we adopt a multi-path geometric channel model where each channel path can experience one or multiple interactions in the environment, which consists of multiple targets. The complex-valued channel path gain $\mathsf{h}_{g,\ell}(t) \in \mathbb{C}$ can be modeled as [104]

$$\mathsf{h}_{g,\ell}(t) = \underbrace{(\mathbf{g}^T \boldsymbol{\Psi} \mathbf{v}(\bar{\theta}_{g,\ell}(t)) \bar{\gamma}_{g,\ell}(t))}_{\text{Radar}\to\text{RIS}\to\text{Target}} \times \underbrace{(\mathbf{g}^T \boldsymbol{\Psi} \mathbf{v}(\ddot{\theta}_{g,\ell}(t)) \ddot{\gamma}_{g,\ell}(t))}_{\text{Target}\to\text{RIS}\to\text{Radar}}, \tag{4.10}$$

$$= \bar{\gamma}_{g,\ell}(t) \left((\mathbf{g} \odot \boldsymbol{\psi})^T \mathbf{v} \left(\bar{\theta}_{g,\ell}(t) \right) \right) \times \ddot{\gamma}_{g,\ell}(t) \left((\mathbf{g} \odot \boldsymbol{\psi})^T \mathbf{v} \left(\ddot{\theta}_{g,\ell}(t) \right) \right), \tag{4.11}$$

where $\mathbf{g} \in \mathbb{C}^N$ is the normalized near-field forward/backward channel vector between the radar feeding antenna and the RIS elements. The normalization is relative to the scalar channel passing through the RIS reference element, whose complex-valued gain is included in the definitions of $\bar{\gamma}_{g,\ell}(t), \ddot{\gamma}_{g,\ell}(t)$ [2] . The far-field transmit/receive RIS array response vector is $\mathbf{v}(\cdot) \in \mathbb{C}^N$. Let an angle notation of φ denote the set of the azimuth and zenith angles, $\varphi = \{\varphi^{\text{az}}, \varphi^{\text{ze}}\}$. $\bar{\theta}_{g,\ell}(t)$ (and $\ddot{\theta}_{g,\ell}(t)$) are the time-varying azimuth and zenith angles of departure (and arrival) of the ℓ^{th} channel path of the g^{th} target, relative to the RIS reference element. The time-varying nature of the path gains and the angles is caused by the mobility of the targets in the environment. $\mathcal{G}(\varphi)$ is the transmit/receive gain of the feeding antenna in the direction

[2] The diacritical marks $\bar{(\cdot)}$ and $\ddot{(\cdot)}$ are used to distinguish between the forward (Radar-Target) and backward (Target-Radar) sides, respectively.

φ. $\bar{\gamma}_{g,\ell}(t), \ddot{\gamma}_{g,\ell}(t) \in \mathbb{C}$ are the two-hop forward and backward complex-valued channel path gains, including the propagation between the radar transceiver and the RIS reference element, and the propagation between the RIS reference and the g^{th} target.

The normalized near-field channel vector $\mathbf{g}$, between the radar transceiver and the RIS elements, can be represented as [104]

$$[\mathbf{g}]_n = \left(\frac{\mathcal{G}(\bar{\Omega}_n)\zeta(\bar{\omega}_n, \bar{\theta}_{g,\ell})\delta_1^2}{\mathcal{G}(\bar{\Omega}_1)\zeta(\bar{\omega}_1, \bar{\theta}_{g,\ell})\delta_n^2} \right)^{1/2} \mathrm{e}^{-\jmath 2\pi(\delta_n - \delta_1)/\lambda}, \tag{4.12}$$

where $n \in \{1, \ldots, N\}$ and $\lambda = \frac{\varsigma}{f_0}$ is the operating wavelength. δ_n is the distance between the radar feeding antenna and the n^{th} RIS element, where δ_1 represents the distance with respect to the RIS reference element. Let the vector of distances between the radar feeding antenna and the RIS elements be $\boldsymbol{\delta} = [\delta_1, \ldots, \delta_N]^T$. $\bar{\Omega}_n$ (and $\ddot{\Omega}_n$) are the azimuth and zenith angles of departure (and arrival), relative to the radar feeding antenna, for the propagation between the radar transceiver and the n^{th} RIS element. $\bar{\omega}_n$ (and $\ddot{\omega}_n$) are the azimuth and zenith angles of arrival (and departure), relative to the n^{th} RIS element, for the propagation between the radar transceiver and the n^{th} RIS element. $\zeta(\varphi_{\text{in}}, \varphi_{\text{out}})$ is the radar cross-section gain of an RIS element towards the direction φ_{out}, when illuminated from the direction φ_{in}, which is modeled in [104].

The two-hop forward and backward channel path gains are defined as [104]

$$\bar{\gamma}_{g,\ell}(t) = \left(\frac{\mathcal{G}(\bar{\Omega}_1)\zeta(\bar{\omega}_1, \bar{\theta}_{g,\ell})}{(4\pi)^2\delta_1^2\bar{d}_{g,\ell}^2(t)\bar{\mathsf{L}}_{g,\ell}(t)} \right)^{1/2} \mathrm{e}^{-\jmath 2\pi(\delta_1 + \bar{d}_{g,\ell}(t))/\lambda}, \tag{4.13}$$

$$\ddot{\gamma}_{g,\ell}(t) = \left(\frac{\sigma_g\zeta(\ddot{\theta}_{g,\ell}, \ddot{\omega}_1)\mathcal{G}(\ddot{\Omega}_1)\lambda^2}{(4\pi)^3\ddot{d}_{g,\ell}^2(t)\delta_1^2\ddot{\mathsf{L}}_{g,\ell}(t)} \right)^{1/2} \mathrm{e}^{-\jmath 2\pi(\ddot{d}_{g,\ell}(t) + \delta_1)/\lambda}, \tag{4.14}$$

where $\bar{d}_{g,\ell}(t), \ddot{d}_{g,\ell}(t)$ are the time-varying forward and backward traveling distance of the ℓ^{th} path, between the RIS reference element and the g^{th} target, which can be related to the total propagation distance such that $R_{g,\ell}(t) = 2\delta_1 + \bar{d}_{g,\ell}(t) + \ddot{d}_{g,\ell}(t)$. σ_g

is the radar cross-section gain of the g^{th} target. $\bar{\mathsf{L}}_{g,\ell}(t), \ddot{\mathsf{L}}_{g,\ell}(t)$ are the time-varying forward and backward loss factors for any additional attenuation. The time-varying nature of these distances and these loss factors is caused by the mobility of the targets in the environment.

4.4 Problem Formulation

In this chapter, our objective is to efficiently estimate the depth map of the surrounding environment using the RIS-aided wireless sensing system described in Section 4.3.

4.4.1 Problem Definition

Following the depth map definition in [31], the depth map, $\mathbf{D}_{\text{map}} \in \mathbb{R}^{M_{\text{h}} \times M_{\text{w}}}$, can be defined as an image of resolution M_{w} pixels wide and M_{h} pixels high, where the value of each pixel denotes the smallest depth between the RIS reference element and the targets/surfaces in this pixel. The total number of pixels in the depth map is $M_{\text{res}} = M_{\text{w}} M_{\text{h}}$. Through (a) effectively scanning the environment using several RIS interaction vectors and (b) processing the receive sensing signals, the RIS aided sensing system can construct the RIS based estimated depth map.

To effectively scan the environment, we define a sensing codebook of RIS interaction vectors, $\boldsymbol{\mathcal{F}} = \{\boldsymbol{\psi}_m : m \in \mathcal{M}, \mathcal{M} = \{0, \ldots, M-1\}\}$. Each RIS interaction vector aids in the transmission and reception of a single chirp signal, when directed towards a certain direction in the environment. For the m^{th} RIS interaction vector, $\boldsymbol{\psi}_m$, the discrete-time complex-valued channel path gain $\mathsf{h}_{g,\ell}[s, m]$, $s \in \mathcal{S}$, $m \in \mathcal{M}$,

can be expressed as

$$\mathsf{h}_{g,\ell}[s,m] = \bar{\gamma}_{g,\ell}[s,m]\left((\mathbf{g}\odot\boldsymbol{\psi}_m)^T\,\mathbf{v}\left(\bar{\theta}_{g,\ell}[s,m]\right)\right)\times \\ \ddot{\gamma}_{g,\ell}[s,m]\left((\mathbf{g}\odot\boldsymbol{\psi}_m)^T\,\mathbf{v}\left(\ddot{\theta}_{g,\ell}[s,m]\right)\right). \quad (4.15)$$

The receive baseband digital signal can then be defined as

$$z[s,m] = \underbrace{\sum_{g=1}^{\mathsf{G}_{\text{tar}}}\sum_{\ell=1}^{L}\sqrt{\rho_{g,\ell}[s,m]}\mathrm{e}^{-\jmath\vartheta_{g,\ell}[s,m]}\mathrm{e}^{+\jmath\Xi_{g,\ell}[s,m]}}_{\text{Receive signal}} + \underbrace{w[s,m]\mathrm{e}^{\jmath\chi[s]}}_{\text{Noise}}, \quad (4.16)$$

where $\rho_{g,\ell}[s,m] = \mathcal{E}_{\text{T}}|\mathsf{h}_{g,\ell}[s,m]|^2$ is the channel path receive power and $\vartheta_{g,\ell}[s,m] = \arg\left(\mathsf{h}_{g,\ell}[s,m]\right)$ is the channel path phase, for the s^{th} ADC sample and using the m^{th} RIS interaction vector. $\Xi_{g,\ell}[s,m]$ can be formulated as

$$\Xi_{g,\ell}[s,m] = 2\pi\left(f_0\xi_{g,\ell}[s,m] + St_{\text{fast}}\xi_{g,\ell}[s,m] - \tfrac{S}{2}\xi^2_{g,\ell}[s,m]\right), \quad (4.17)$$

where the propagation delay $\xi_{g,\ell}[s,m]$ can be defined as

$$\xi_{g,\ell}[s,m] = \frac{R_{g,\ell}[s,m]}{\varsigma} = \frac{R_{0,g,\ell} + \nu_{g,\ell}t_{\text{slow}} + \frac{\mathsf{a}_{g,\ell}}{2}t^2_{\text{slow}}}{\varsigma}, \quad (4.18)$$

$$t_{\text{slow}} = t_{\text{fast}} + mT_{\text{PRI}} = sT_{\text{S}} + mT_{\text{PRI}}. \quad (4.19)$$

By stacking the S receive ADC samples, we can construct the receive sensing vector, $\mathbf{z}[m] \in \mathbb{C}^{M_{\text{sample}}}$, corresponding to the transmission of a single chirp signal using one RIS interaction vector, $\mathbf{z}[m] = [z[0,m],\ldots,z[M_{\text{sample}}-1,m]]^T$. If M radar chirps (a single radar frame) are transmitted and received via M RIS interaction vectors, the aggregated receive sensing signal matrix, $\mathbf{Z} \in \mathbb{C}^{M_{\text{sample}}\times M}$, can be expressed as

$$\mathbf{Z} = [\mathbf{z}[0],\mathbf{z}[1],\ldots,\mathbf{z}[M-1]]. \quad (4.20)$$

Next, to estimate the depth map, we define a post-processing function $\mathbf{p}(.)$. Given the receive sensing matrix $\mathbf{Z}$ and the RIS sensing codebook $\boldsymbol{\mathcal{F}}$, the estimated depth map can be formulated as

$$\widehat{\mathbf{D}}_{\text{map}} = \mathbf{p}(\mathbf{Z};\boldsymbol{\mathcal{F}}). \quad (4.21)$$

Our objective is to minimize the estimation error between the estimated depth map $\widehat{\mathbf{D}}_{\text{map}}$ and the actual depth map $\mathbf{D}_{\text{map}}$. For this reason, we adopt the root-mean squared error (RMSE) and the mean absolute error (MAE) as the performance metrics of depth sensing, which are defined as [31]

$$\Delta_{\text{RMSE}} = \left(\frac{1}{M} \|\mathbf{D}_{\text{map}} - \mathbf{p}(\mathbf{Z}; \boldsymbol{\mathcal{F}})\|_2^2 \right)^{1/2}, \tag{4.22}$$

$$\Delta_{\text{MAE}} = \frac{1}{M} \|\mathbf{D}_{\text{map}} - \mathbf{p}(\mathbf{Z}; \boldsymbol{\mathcal{F}})\|_1^2. \tag{4.23}$$

4.4.2 Main Challenges

Estimating scene depth maps using mmWave sensing systems suffer from the following challenges.

1. Codebook design: To build RIS-based depth maps capable of complementing RGB-D based depth maps, the RIS interaction codebook needs to be designed to reflect the incident signals in the directions of the full rectangular grid of typical depth optical sensors. Classical RIS codebooks [28], however, are designed based on DFT codebooks which forms parabolic grids instead of rectangular grids. In addition, mmWave MIMO based sensing codebooks, as detailed in [31], can not be adopted as RIS sensing codebooks.

2. Low-resolution depth maps: mmWave MIMO based depth map estimation has been investigated for wireless AR/VR systems [31]. Scaling mmWave antenna arrays, however, is associated with large computational/hardware complexity and energy consumption. This limitation poses a prominent challenge in scaling the spatial resolution of the depth maps.

3. Inter-target and inter-path interferences: When sensing the depth of a certain region of interest (represented by a single pixel), the best scenario is when only a single target exist in that region of interest, and that target backscatters a single-

bounce path to the receiver. In practice, however, it can be hard to differentiate the receive signals from multiple targets that are close to each others. In addition, the incident signals on each target can experience multiple bounces in different directions — directions away from the desired direction — before reaching the receiver. The challenge is how to design the RIS aided sensing solution to detect the desired channel path while filtering out the undesired channel paths [31]. In the next section, we present our proposed solution to address these challenges in estimating scene depth maps.

4.5 Proposed Solution

In this section, we introduce a comprehensive framework for scene depth estimation using RIS aided mmWave wireless sensing systems.

4.5.1 Key Idea

Because of the massive number of the nearly-passive RIS elements, these surfaces can adopt fine-grained reflection beams while scanning the environment, enabling high-resolution sensing grids using energy-efficient architectures [28]. In addition, RIS aided sensing systems can filter out more undesired paths than the ones filtered out by mmWave MIMO based sensing systems, without leveraging any elaborate post-processing functions (as opposed to the ones used in [31]). One possible reason is that an RIS interaction matrix is designed to focus the reflection in one desired direction and the reception from the same direction; any channel path arriving back to the RIS from a direction other than the desired direction is reflected away from the radar receiver. Also, for AR/VR systems, the post-processing sensing tasks can be offloaded from the AR/VR devices to the RIS aided wireless sensing systems — a significant advantage for AR/VR appplications. For these reasons, we propose an RIS aided

sensing based scene depth map estimation solution capable of further improving the depth perception of the surrounding environment compared to existing RGB based depth map estimation solutions [8, 9]. Next, we formulate the two main elements of our proposed RIS aided sensing framework for scene depth estimation, namely (a) the RIS sensing codebook design and (b) the scene depth estimation solution.

4.5.2 RIS Sensing Codebook Design

Our objective for the RIS interaction codebook design is to construct a sensing grid of reflected directions that fits the rectangular grid of a depth camera. For simplicity of the formulation, we first start by enunciating a set of reasonable assumptions as follows. Assume the M radar chirps (a single radar frame) are transmitted, received, and processed for depth map construction over a time interval of T seconds, during which the environment is assumed relatively static. In such case, the Doppler velocity and acceleration terms can be neglected from the previous definitions. We can also omit the time dependence notation previously used in defining some of the variables, e.g. $\xi_{g,\ell}(t)$, $R_{g,\ell}(t)$, $\bar{\theta}_{g,\ell}(t)$, $\ddot{\theta}_{g,\ell}(t)$, $\bar{\gamma}_{g,\ell}(t)$, $\ddot{\gamma}_{g,\ell}(t)$, $\bar{d}_{g,\ell}(t)$, $\ddot{d}_{g,\ell}(t)$, $\bar{\mathsf{L}}_{g,\ell}(t)$, $\ddot{\mathsf{L}}_{g,\ell}(t)$.

Assume the RIS is employing a uniform planar array (UPA) structure in the x-z plane. The RIS is then equipped with N_{H} elements on the x-axis (the horizontal axis) and N_{V} elements on the z-axis (the vertical axis), where $N = N_{\mathrm{H}} N_{\mathrm{V}}$. In such case, the far-field RIS array response vector $\mathbf{v}(\varphi)$, in the direction $\varphi = \{\varphi^{\mathrm{az}}, \varphi^{\mathrm{ze}}\}$, can then be formulated as

$$\mathbf{v}(\varphi) = \mathbf{v}_z(\varphi) \otimes \mathbf{v}_x(\varphi). \tag{4.24}$$

where $\mathbf{v}_x(.)$ and $\mathbf{v}_z(.)$ represent the elemental array response vectors in the x and z

directions, and are expressed as

$$\mathbf{v}_x\left(\varphi\right) = \left[1, \mathrm{e}^{\jmath\kappa \mathsf{d}\cos(\varphi^{\mathrm{az}})\sin(\varphi^{\mathrm{ze}})}, \ldots, \mathrm{e}^{\jmath\kappa \mathsf{d}(N_{\mathrm{H}}-1)\cos(\varphi^{\mathrm{az}})\sin(\varphi^{\mathrm{ze}})}\right]^T, \tag{4.25}$$

$$\mathbf{v}_z\left(\varphi\right) = \left[1, \mathrm{e}^{\jmath\kappa \mathsf{d}\cos(\varphi^{\mathrm{ze}})}, \ldots, \mathrm{e}^{\jmath\kappa \mathsf{d}(N_{\mathrm{V}}-1)\cos(\varphi^{\mathrm{ze}})}\right]^T, \tag{4.26}$$

where $\kappa = \frac{2\pi}{\lambda}$ is the wave number and d is the RIS element spacing in meters. For simplicity of scene definition, let the horizontal direction (the width) of the depth map be parallel to the x-axis, and its vertical direction (the height) be parallel to the z-axis. Let the RIS reference element — the focal point of the scene depth map — be the origin of the rectangular coordinate system. In such case, the depth of a target is measured by the y-coordinate of the x-z plane of that target, with respect to the RIS reference element. Consider an oversampled RIS interaction codebook of $M = \overline{N}_{\mathrm{V}}\overline{N}_{\mathrm{H}}$ beams, where $\overline{N}_{\mathrm{V}} = N_{\mathrm{V}}F_{\mathrm{V}}^{\mathrm{OS}}$ and $\overline{N}_{\mathrm{H}} = N_{\mathrm{H}}F_{\mathrm{H}}^{\mathrm{OS}}$. $F_{\mathrm{V}}^{\mathrm{OS}}$ and $F_{\mathrm{H}}^{\mathrm{OS}}$ are the oversampling factors in the horizontal and vertical dimensions, respectively.

Under these assumptions, The complex-valued channel path gain $\mathsf{h}_{g,\ell}[m]$, $m \in \mathcal{M}$, can now be expressed as

$$\mathsf{h}_{g,\ell}[m] = \bar{\gamma}_{g,\ell}\left(\left(\mathbf{g}\odot\boldsymbol{\psi}_m\right)^T\mathbf{v}\left(\bar{\theta}_{g,\ell}\right)\right) \times \ddot{\gamma}_{g,\ell}\left(\left(\mathbf{g}\odot\boldsymbol{\psi}_m\right)^T\mathbf{v}\left(\ddot{\theta}_{g,\ell}\right)\right). \tag{4.27}$$

The receive baseband digital signal can be redefined as

$$z[s,m] = \underbrace{\sum_{g=1}^{\mathsf{G}_{\mathrm{tar}}}\sum_{\ell=1}^{L}\sqrt{\rho_{g,\ell}[m]}\,\mathrm{e}^{-\jmath\vartheta_{g,\ell}[m]}\,\mathrm{e}^{+\jmath\Xi_{g,\ell}}}_{\text{Receive signal}} + \underbrace{w[s,m]\mathrm{e}^{\jmath\chi[s]}}_{\text{Noise}}, \tag{4.28}$$

where $\rho_{g,\ell}[m] = \mathcal{E}_{\mathrm{T}}|\mathsf{h}_{g,\ell}[m]|^2$ and $\vartheta_{g,\ell}[m] = \arg\left(\mathsf{h}_{g,\ell}[m]\right)$. The phase term $\Xi_{g,\ell}$ can then be formulated as

$$\Xi_{g,\ell} = 2\pi\left(f_0\xi_{g,\ell} + St_{\mathrm{fast}}\xi_{g,\ell} - \tfrac{S}{2}\xi_{g,\ell}^2\right), \tag{4.29}$$

where $\xi_{g,\ell} = R_{g,\ell}/\varsigma$ is the propagation delay. The receive sensing matrix, $\mathbf{Z}$, can then be constructed as in (4.20).

Now, we explain how to design the RIS interaction matrix to reflect the incident signal into a certain direction. From (4.28), the receive signal from a target in a certain direction can become more distinguishable from the ones received from targets in other directions by controlling their respective channel gains, $|\mathsf{h}_{g,\ell}|$. More specifically, to distinguish more the receive signal gain of the ℓ^{th} channel path of the g^{th} target, the RIS interaction vector $\boldsymbol{\psi}^{\star}$ can be designed as

$$\boldsymbol{\psi}^{\star} = \arg\max_{\boldsymbol{\psi}} \ |\mathsf{h}_{g,\ell}| = \arg\max_{\boldsymbol{\psi}} \left|((\mathbf{g} \odot \boldsymbol{\psi})^T \mathbf{v}(\bar{\theta}_{g,\ell}))((\mathbf{g} \odot \boldsymbol{\psi})^T \mathbf{v}(\ddot{\theta}_{g,\ell}))\right|, \tag{4.30}$$

$$\text{s.t.} \ \ |[\boldsymbol{\psi}]_n| = 1, \ \forall n \in \{1, \ldots, N\}. \tag{4.31}$$

Note that we are only interested in distinguishing single-bounce paths to estimate the depth correctly [31], $\bar{\theta}_{g,\ell} = \ddot{\theta}_{g,\ell} = \theta_{g,\ell}$. In such case, the optimization problem is reduced to

$$\boldsymbol{\psi}^{\star} = \arg\max_{\boldsymbol{\psi}} \ \left|(\mathbf{v}(\theta_{g,\ell}) \odot \mathbf{g})^T \boldsymbol{\psi}\right|^2, \tag{4.32}$$

$$\text{s.t.} \ \ |[\boldsymbol{\psi}]_n| = 1, \ \forall n \in \{1, \ldots, N\}. \tag{4.33}$$

Assume prior knowledge of (i) the distance vector $\boldsymbol{\delta}$ between the radar feeding antenna and the RIS elements and (ii) the direction specified by $\theta_{g,\ell}$. The RIS interaction vector $\boldsymbol{\psi}^{\star}$ can then be designed using equal-gain conjugate beamforming as

$$\boldsymbol{\psi}^{\star} = (\mathbf{v}(\theta_{g,\ell}) \odot \arg(\mathbf{g}))^* = (\mathbf{v}(\theta_{g,\ell}) \odot \mathrm{e}^{-\jmath 2\pi(\boldsymbol{\delta} - \delta_1)/\lambda})^*. \tag{4.34}$$

Next, we explain how to design the RIS interaction codebook to construct a sensing grid of reflected directions that fits the rectangular grid of a depth camera. More specifically, let $\mathcal{O}$ be the set of spherical coordinate angles representing the grid point directions from the desired rectangular grid, such that $\mathcal{O} = \{\theta_m\}_{m=0}^{M-1}$. We adopt the design of the set $\mathcal{O}$ from [30, Sec. VII, Eq. 19] to eliminate any grid mismatch distortion. The set $\mathcal{O}$ can be completely described using the scene field of view FoV,

the aspect ratio of the depth map A_{R}, and the number of horizontal and vertical grid points $\overline{N}_{\mathrm{H}}, \overline{N}_{\mathrm{V}}$, as detailed in [31]. For the m^{th} grid point pointing towards the far-field direction $\theta_m \in \mathcal{O}$, the RIS interaction vector $\boldsymbol{\psi}_m^{\star}$ can then be designed as

$$\boldsymbol{\psi}_m^{\star} = \arg\max_{\boldsymbol{\psi}_m} \ |\mathsf{h}_{g,\ell}[m]| \tag{4.35}$$

$$\text{s.t.} \ \ |[\boldsymbol{\psi}_m]_n| = 1, \ \forall n \in \{1, \ldots, N\}, \tag{4.36}$$

$$\boldsymbol{\psi}_m^{\star} = \left(\mathbf{v}(\theta_m) \ \odot \ \mathrm{e}^{-\jmath 2\pi(\boldsymbol{\delta}-\delta_1)/\lambda}\right)^*, \ m \in \mathcal{M}, \tag{4.37}$$

where $\mathsf{h}_{g,\ell}[m]$ is defined in (4.27). Finally, given prior knowledge of (i) the distance vector $\boldsymbol{\delta}$ and (ii) the set of codebook angles $\mathcal{O}$, the RIS interaction codebook can be calculated as

$$\boldsymbol{\mathcal{F}} = \left\{\boldsymbol{\psi}_m \in \mathbb{C}^{N\times 1} : \boldsymbol{\psi}_m = (\mathbf{v}(\theta_m) \odot \mathrm{e}^{-\jmath 2\pi(\boldsymbol{\delta}-\delta_1)/\lambda})^*, \theta_m \in \mathcal{O}\right\}. \tag{4.38}$$

where $|\boldsymbol{\mathcal{F}}| = |\mathcal{O}| = M$. Given a pre-designed RIS interaction codebook, we formulate next the scene depth map estimation solution.

4.5.3 Scene Depth Estimation

In this section, we formulate the scene depth map estimation solution, which is outlined in Algorithm 6. First, the RIS interaction codebook $\boldsymbol{\mathcal{F}}$ is designed, as covered in Section 4.5.2. Then, the sensing system sweeps over the RIS codebook and acquires the receive sensing signal for every RIS interaction vector $\boldsymbol{\psi}_m \in \boldsymbol{\mathcal{F}}$, as defined in (4.28). After that, the receive sensing matrix $\mathbf{Z}$ is constructed as in (4.20). The receive sensing matrix is then processed using 1D Fourier transforms along its column dimension, to calculate the scene range estimate for every grid point $m \in \mathcal{M}$. The Fourier-based range profile matrix $\mathbf{Z}^{\mathrm{RP}} \in \mathbb{C}^{M_{\mathrm{sample}}\times M}$ can be formulated as

$$\mathbf{Z}^{\mathrm{RP}} = \mathrm{FFT}_m\left(\mathbf{Z}\right), m \in \mathcal{M}, \tag{4.39}$$

Algorithm 6 RIS-Based Scene Depth Estimation Solution

Inputs: Field of view FoV, aspect ratio A_{R},

number of horizontal/vertical grid points $\overline{N}_{\mathrm{H}}, \overline{N}_{\mathrm{V}}$.

Outputs: Depth map estimate $\widehat{\mathbf{D}}_{\mathrm{map}}$.

1: Design RIS interaction codebook $\boldsymbol{\mathcal{F}}$, as in Section 4.5.2.

2: **for** $m = 1$ **to** M **do** ▷ For each $\boldsymbol{\psi}_m$

3: Acquire receive *sensing* signal $z[s, m]$, $\forall s \in \mathcal{S}$, (4.28).

4: Construct receive *sensing* matrix $\mathbf{Z}$, as in (4.20).

5: Calculate scene range estimate vector $\hat{\mathbf{r}}$, as in (4.40).

6: Construct the range map estimate $\widehat{\mathbf{R}}_{\mathrm{map}}$, as in (4.41).

7: Construct the depth map estimate $\widehat{\mathbf{D}}_{\mathrm{map}}$, as in [31].

where m is the column index of the matrix $\mathbf{Z}$. The scene range estimate vector $\hat{\mathbf{r}} \in \mathbb{R}^M$ can then be calculated as

$$[\hat{\mathbf{r}}]_m = \Delta_{\mathrm{R}} \times \arg\max_{s} \left| \left[\mathbf{Z}^{\mathrm{RP}}\right]_{s,m} \right|, m \in \mathcal{M}, \tag{4.40}$$

where $\Delta_{\mathrm{R}} = \varsigma/(2\mathrm{BW})$ is the range resolution of the Fourier-based estimation solution.

Next, the sensing system constructs the 2D range map estimate matrix $\widehat{\mathbf{R}}_{\mathrm{map}} \in \mathbb{R}^{\overline{N}_{\mathrm{V}} \times \overline{N}_{\mathrm{H}}}$ from the 1D scene range estimate vector $\hat{\mathbf{r}} \in \mathbb{R}^M$. Let the horizontal grid index be denoted by $h_{\mathrm{map}} \in \{1, \ldots, \overline{N}_{\mathrm{H}}\}$ and the vertical grid index be denoted by $v_{\mathrm{map}} \in \{1, \ldots, \overline{N}_{\mathrm{V}}\}$. By converting the linear indices to matrix subscripts, the range map estimate $\widehat{\mathbf{R}}_{\mathrm{map}}$ can be constructed as

$$\left[\widehat{\mathbf{R}}_{\mathrm{map}}\right]_{v_{\mathrm{map}}, h_{\mathrm{map}}} = [\hat{\mathbf{r}}]_m \,,\; m = (v_{\mathrm{map}} - 1)\overline{N}_{\mathrm{H}} + h_{\mathrm{map}}, \tag{4.41}$$

where $m \in \{1, \ldots, M\}$. After that, the depth map estimate $\widehat{\mathbf{D}}_{\mathrm{map}} \in \mathbb{R}^{\overline{N}_{\mathrm{V}} \times \overline{N}_{\mathrm{H}}}$ can then be calculated from the range map estimate $\widehat{\mathbf{R}}_{\mathrm{map}}$ and the set of angles of the grid points' spherical coordinates, as detailed in [31]. Finally, the depth map estimate is

mapped from the codebook resolution of of $\overline{N}_{\mathrm{V}} \times \overline{N}_{\mathrm{H}}$ pixels to the desired up-scaled depth map resolution of $M_h \times M_w$ pixels, using 2D signal interpolation [31].

4.6 Simulation Results

In this section, we evaluate the performance of our proposed RIS based depth map estimation solution.

4.6.1 Simulation Framework

We follow the simulation framework adopted in [31] to evaluate the performance of the proposed solution with realistic channels. We first build a detailed floor plan with sufficient number of facets using a high-fidelity 3D graphics design engine, Blender [112]. this floor plan is then exported to an accurate 3D ray-tracing simulator, Wireless Insite [1]. Using the ray-tracing output data, we use MATLAB to construct the receive signal models and implement the proposed solution. For comparison, we generate the ground truth depth map by placing a depth camera in Blender at the same position of the RIS reference element, and adjusting the camera scene parameters to follow the same scene parameters of the proposed RIS sensing codebook grid of reflected directions.

System Model: The RIS-aided sensing system parameters are summarized in Table 4.1. The adopted FMCW radar configuration can be achieved by current commerical FMCW radar systems. The adopted RIS architectures are 30×30 and 40×40 UPAs, i.e. $N_{\mathrm{H}} = N_{\mathrm{V}} = N_{\mathrm{RIS}} \in \{30, 40\}$. It is worth noting that the far field assumption of the channel between the RIS and the targets may not be valid if the size of the RIS architecture is increased beyond certain limits. It is interesting to characterize these limits and analyze the near-field operation in the future work. For simplicity, assume the radar cross section gain of the RIS elements is an isotropic gain with half-

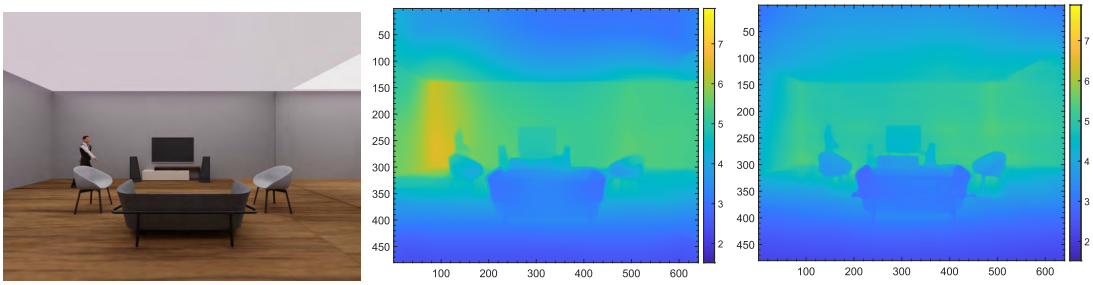
Table 4.1: The Adopted RIS-Aided Sensing System Parameters

System Configuration	
RIS architecture, $N_{\mathrm{H}} \times N_{\mathrm{V}}$	$\{30 \times 30; 40 \times 40\}$
Starting frequency, f_0	$60\,\mathrm{GHz}$
Chirp slope, S	$300\,\mathrm{MHz}\,\mu\mathrm{s}^{-1}$
ADC Sampling frequency, F_{S}	$38\,\mathrm{MS/s}$
Samples per chirp, M_{sample}	512
Chirp repetition interval, T_{PRI}	$13.47\,\mu\mathrm{s}$
Derived Parameters	
Chirp duration, T_{active}	$13.47\,\mu\mathrm{s}$
Transmission bandwidth, BW	$4.04\,\mathrm{GHz}$
Range resolution, Δ_{R}	$3.71\,\mathrm{cm}$
Maximum range, $\mathrm{R}_{\mathrm{max}}$	$18.95\,\mathrm{m}$
Chirp rate, F_{chirp}	$74.2\,\mathrm{kHz}$
RIS codebook size, $\lvert\boldsymbol{\mathcal{F}}\rvert = M$	$\{14,400;\ 25,600\}$
Depth map sensing rate, F_{DM}	$\{5.15, 2.90\}\,\mathrm{Hz}$

wavelength RIS element spacing, $\mathsf{d} = \lambda/2$. The transmit power of the radar system is set to 20 dBm and 15 dBm for the 30×30 and 40×40 RIS architectures, respectively.

The transmit/receive gain of the feeding antenna is assumed to reach a maximum of 25 dBi in the direction of the RIS elements. The maximum effective isotropic radiated power (EIRP) is then 45 dBm and 40 dBm for the adopted RIS architectures, respectively. The radar system transmits M repeated chirps to sense the environment using M RIS interaction vectors out of the RIS interaction codebook.

Receive Signal Generation: The receive radar signals are generated in two steps. The



(a) RGB Scene Image

(b) RGB-Based Depth Map [8]

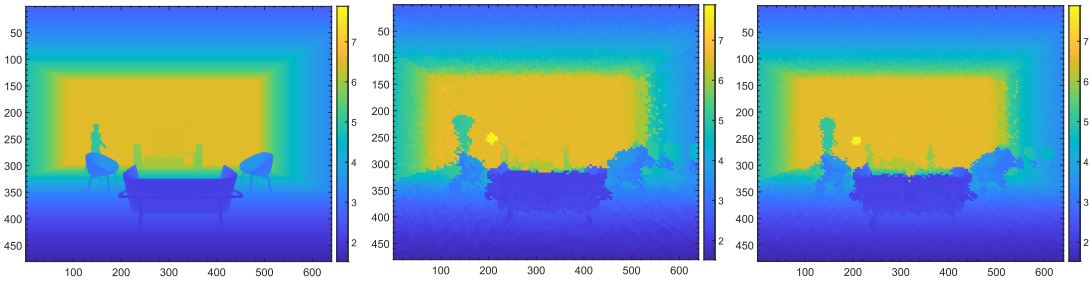
$\Delta_{\text{RMSE}} = 94.5\,\text{cm}$

$\Delta_{\text{MAE}} = 82.9\,\text{cm}$

(c) RGB-Based Depth Map [9]

$\Delta_{\text{RMSE}} = 98.3\,\text{cm}$

$\Delta_{\text{MAE}} = 86.6\,\text{cm}$



(d) Ground Truth Depth Map

(e) Proposed Sol. (30×30 RIS)

$\Delta_{\text{RMSE}} = 37.5\,\text{cm}$

$\Delta_{\text{MAE}} = 14.5\,\text{cm}$

(f) Proposed Sol. (40×40 RIS)

$\Delta_{\text{RMSE}} = 31.9\,\text{cm}$

$\Delta_{\text{MAE}} = 11.6\,\text{cm}$

Figure 4.2: For the Living Room Scenario, the Proposed RIS-Based Depth Estimation Solution Is Compared Against Two RGB-Based Depth Estimation Solutions [8, 9] and the Ground Truth Depth Map. The RIS Is Equipped with 30×30 or 40×40 UPA Elements and Codebook Oversampling Factors of Four Are Employed. (a) The Scene under Study; (b, c) The Estimated Maps from Monocular RGB Images Using RGB-Based Solutions [8, 9]; (d) The Ground Truth Depth Map; (e, f) The Estimated Depth Maps Using Our Proposed RIS-based Solution.

first step is generating the parameters of the channel paths using Wireless InSite [1]. The adopted propagation model and diffuse scattering parameters are the same as the ones in [31]. The second step is using the generated channel data to construct the receive baseband digital signals (4.28). The thermal noise floor is calculated based on a noise figure of 10 dB and the transmission bandwidth of 4.04 GHz.

RIS-Aided Depth Map Estimation Parameters: The RIS interaction matrix is de-

signed based on a 100° field of view centered on the RIS boresight, a 4/3 scene aspect ratio, and horizontal and vertical oversampling factors of $F_{\mathrm{H}}^{\mathrm{OS}} = F_{\mathrm{V}}^{\mathrm{OS}} = 4$. The RIS codebook size is calculated using $|\boldsymbol{\mathcal{F}}| = M = N_{\mathrm{H}} F_{\mathrm{H}}^{\mathrm{OS}} N_{\mathrm{V}} F_{\mathrm{V}}^{\mathrm{OS}}$. Correspondingly, the ground truth depth maps are generated by Blender with a depth camera of 100° field of view and a sensor width of 32 mm. The image quality of the ground truth depth maps and the up-scaled estimated depth maps is set to 480p resolution, i.e. 640×480 pixels. Next, we evaluate the performance of our proposed RIS aided depth map estimation solution in an indoor living room scenario.

4.6.2 Results for A Living Room Scenario

In this scenario, we consider a $15.6\,\mathrm{m} \times 6.5\,\mathrm{m} \times 3.8\,\mathrm{m}$ indoor space, with a glass wall dividing the space into two rooms. The room under study is a $9.6\,\mathrm{m} \times 6.5\,\mathrm{m} \times 3.8\,\mathrm{m}$ living room, where a 1.8 m tall person is moving from left to right. The adopted materials of the inanimate objects/surfaces follow the ITU default parameter values at 60GHz. The adopted materials are as follows: concrete for the walls, floorboard for the floor, ceiling board for the ceiling, glass for the glass wall and the TV, wood for the entertainment center, the speakers, the arm chairs, and the sofa. The RIS is assumed to be placed on the wall behind the sofa. The number of facets ranges between $\approx$ 2k and 30k for the inanimate objects/surfaces. $20,542$ facets are used for the person model. The number of facets are as follows: $9,946$ facets for the wall in front, $4,222$ for the left wall, $10,494$ for the glass wall, $8,391$ facets for the floorboard, $10,494$ facets for the ceiling, $2,926$ facets for the TV, $1,152$ facets for the entertainment center, $1,152$ facets for the speakers, $6,238$ facets for the arm chairs, $30,660$ facets for the sofa, and $20,542$ facets for the person model. We compare the proposed solution against two RGB-based solutions [8, 9] to demonstrate the capability of the RIS-aided sensing in (i) detecting transparent surfaces and (ii) achieving higher trueness of the

estimated depth values. We follow the official implementation of these RGB-based solutions and utilize the well-trained models on the NYU depth V2 dataset [99].

Fig. 4.2 compares the estimated depth maps from the RIS-based solution against the ones from the RGB-based solutions, which uses monocular RGB images to estimate the depth maps. As shown, the RGB-based solutions can construct the shape of the objects/person more clearly than the proposed solution; i.e they achieve a higher depth precision. These RGB-based solutions, however, do not achieve high depth accuracy due to their low level of trueness, especially when misdetecting the transparent glass wall. As for the proposed RIS-based solution, even though the glass has the lowest scattering factor among all the other materials, the depth of the glass wall can be better perceived with a lower estimation error; i.e. the proposed solution can achieve a higher depth trueness. the depth estimation accuracy of the RIS-based solution, however, suffers from inter-path interference at some directions, where the receive powers of undesired paths are higher than the ones of the desired single-bounce paths. Although the RIS based solution offers a higher spatial resolution than the mmWave MIMO based solution [31], the RIS reflected beams are yet relatively wide compared to the ideal pencil beams. For this reason, high estimation errors are observed around the edges of the objects/person. It would be interesting to address these challenge in future work.

4.7 Conclusion

In this chapter, we considered the problem of scene depth map estimation using mmWave wireless sensing systems. For this problem, we proposed to leverage RISs to accurately estimate high-resolution depth maps. To achieve this objective, we formulated the RIS wireless sensing based scene depth estimation problem and proposed a comprehensive framework for building scene depth maps using RIS aided mmWave

sensing systems. The proposed framework includes designing an RIS interaction codebook capable of creating a sensing grid of reflected beams that meets the desirable characteristics of efficient scene depth map construction. Using the designed RIS interaction codebook, a post-processing solution is developed to build high-resolution depth maps. By adopting accurate 3D ray-tracing models, the simulation results showed that the developed solution can achieve depth map estimation errors in the order of 12 cm. This highlights the potential of leveraging this proposed solution in achieving accurate depth perception of the surrounding environment.

Chapter 5

SUMMARY AND FUTURE WORK

5.1 Summary

This dissertation focused on addressing the key challenges in (a) reconfigurable intelligent surface (RIS) aided wireless communication systems and (b) RIS aided wireless sensing systems for scene depth estimation. We considered RIS-aided wireless communication systems and developed efficient solutions that design the RIS interaction (reflection) matrices with negligible training overhead. We first proposed a novel RIS architecture where only a small number of the RIS elements are active (connected to the baseband). By leveraging compressive sensing and deep learning tools, we then developed three solutions that design the RIS reflection matrices for this new architecture with almost no training overhead. The three proposed solutions are extensively evaluated and compared against each others. The three proposed solutions can achieve near-optimal data rates with negligible training overhead and with a few active elements. Some interesting insights were also developed on the impact of various system and channel parameters. In addition, given an objective of developing standalone RIS architectures, the third solution exploits a deep reinforcement learning framework for the RIS to learn how to predict, on its own, the optimal interaction matrices directly from the sampled channel knowledge. This solution does not require an initial dataset collection phase, as opposed to the supervised learning based solutions.

We also considered the problem of estimating accurate depth maps for AR/VR devices. For this problem, we proposed leveraging the mmWave communication systems

that are deployed on the AR/VR devices to estimate and build high-resolution depth maps. We formulated the communication-constrained depth map sensing problem and proposed a comprehensive framework for realizing this objective. The proposed framework includes (i) the construction of depth map specific sensing codebooks using practical mmWave antenna arrays and (ii) the development of efficient post-processing solutions for jointly processing the receive signals from the multiple sensing beams and estimating high-resolution depth maps. The simulation results highlight the potential of leveraging this proposed framework to complement RGB-D based depth maps and realize immersive depth perception for wireless virtual/augmented reality systems. This work represents an important step towards developing RIS-aided wireless sensing systems for scene depth estimation.

Last, we investigated the scene depth estimation problem using wireless sensing systems much further; we proposed leveraging RISs to accurately estimate high-resolution depth maps. To achieve this objective, we formulated the RIS sensing based scene depth estimation problem and proposed a comprehensive framework for building scene depth maps using RIS aided mmWave sensing systems. The proposed framework includes designing an RIS interaction codebook capable of creating a sensing grid of reflected beams that meets the desirable characteristics of efficient scene depth map construction. Using the designed RIS interaction codebook, a post-processing solution is developed to build high-resolution depth maps. the simulation results highlight the potential of leveraging this proposed framework in achieving accurate depth perception of the surrounding environment.

5.2 Future Work

In this section, we summarize the possible directions for future research as follows.

RIS-aided wireless communication systems: It would be interesting to address

any additional challenges introduced in highly-dynamic environments and aim at increasing the performance robustness of the proposed solutions under such conditions. Given the hardware connstraints imposed by practical RIS implementations, it would be also interesting to analayze the performance of the proposed solutions under the constraint of discrete RIS induced phase shift values. For the proposed solutions, an interesting extension would be the optimization of the sparse distribution of the active sensors leveraging tools from nested and co-prime arrays. For the deep reinforcement learning based solution, an interesting extension would be the development of a *fully-standalone* RIS operation framework, where the RIS configures itself with no control from the wireless communication infrastructure.

RIS-aided wireless sensing systems for scene depth estimation: to decreases the sensing overheard, it would be interesting to leverage the sparse nature of the mmWave channels and develop a scene depth estimation framework based on compressive sensing for RIS aided mmWave sensing systems. Given the practical hardware constraints of mmWave MIMO architectures, it would be also interesting to explore a phase-only approximation of the proposed side lobe reduction approach. In addition, it would be interesting to consider the case of near-field channels between the RIS and the targets and develop reliable solutions for such a case. It would also be interesting to investigate the coupling between the RIS elements and its effect on the scene depth estimation performance.

REFERENCES

[1] Remcom. Wireless InSite. 2021. http://www.remcom.com/wireless-insite.

[2] Ahmed Alkhateeb. DeepMIMO: A Generic Deep Learning Dataset for Millimeter Wave and Massive MIMO Applications. In *Proc. of Information Theory and Applications Workshop (ITA)*, pages 1–8, San Diego, CA, Feb 2019. URL https://www.deepmimo.net/.

[3] Blender 2.80. 2020. URL http://www.blender.org.

[4] BlenderKit. 2020. URL https://www.blenderkit.com/.

[5] Blender Demo Files. 2020. URL https://www.blender.org/download/demo-files/.

[6] TurboSquid. 2020. URL https://www.turbosquid.com/Search/3D-Models.

[7] Free3D. 2020. URL https://free3d.com/.

[8] Junjie Hu, Mete Ozay, Yan Zhang, and Takayuki Okatani. Revisiting Single Image Depth Estimation: Toward Higher Resolution Maps With Accurate Object Boundaries. In *Proc. of IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1043–1051, 2019.

[9] René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. Vision Transformers for Dense Prediction. In *Proc. of IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 12179–12188, 2021.

[10] A. Puglielli, N. Narevsky, P. Lu, T. Courtade, G. Wright, B. Nikolic, and E. Alon. A Scalable Massive MIMO Array Architecture Based on Common Modules. In *Proc. of IEEE International Conference on Communication Workshop (ICCW)*, pages 1310–1315, June 2015.

[11] S. Hu, F. Rusek, and O. Edfors. Beyond Massive MIMO: The Potential of Data Transmission With Large Intelligent Surfaces. *IEEE Transactions on Signal Processing*, 66(10):2746–2758, May 2018. ISSN 1053-587X.

[12] S. V. Hum and J. Perruisseau-Carrier. Reconfigurable Reflectarrays and Array Lenses for Dynamic Antenna Beam Control: A Review. *IEEE Transactions on Antennas and Propagation*, 62(1):183–198, Jan 2014. ISSN 0018-926X.

[13] C. Huang, A. Zappone, G. C. Alexandropoulos, M. Debbah, and C. Yuen. Reconfigurable Intelligent Surfaces for Energy Efficiency in Wireless Communication. *IEEE Transactions on Wireless Communications*, 2019.

[14] X. Tan, Z. Sun, D. Koutsonikolas, and J. M. Jornet. Enabling Indoor Mobile Millimeter-wave Networks Based on Smart Reflect-arrays. In *Proc. of IEEE International Conference on Computer Communications (INFOCOM)*, pages 270–278, April 2018.

[15] Fangchang Mal and Sertac Karaman. Sparse-to-Dense: Depth Prediction from Sparse Depth Samples and a Single Image. In *Proc. of International Conference on Robotics and Automation (ICRA)*, pages 1–8. IEEE, 2018.

[16] Gernot Riegler, Yiyi Liao, Simon Donne, Vladlen Koltun, and Andreas Geiger. Connecting the Dots: Learning Representations for Active Monocular Depth Estimation. In *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7624–7633, 2019.

[17] Yinda Zhang, Sameh Khamis, Christoph Rhemann, Julien Valentin, Adarsh Kowdle, Vladimir Tankovich, Michael Schoenberg, Shahram Izadi, Thomas Funkhouser, and Sean Fanello. Activestereonet: End-to-end Self-supervised Learning for Active Stereo Systems. In *Proc. of the European Conference on Computer Vision (ECCV)*, pages 784–801, 2018.

[18] C. Cheng, Y. Wang, C. Wei, C. Chen, and L. Lin. IR Pattern Characteristics For Active Stereo Matching, October 2019.

[19] Tobias Gruber, Mariia Kokhova, Werner Ritter, Norbert Haala, and Klaus Dictmayer. Learning Super-resolved Depth from Active gated Imaging. In *Proc. of International Conference on Intelligent Transportation Systems (ITSC)*, pages 3051–3058. IEEE, 2018.

[20] Tobias Gruber, Frank Julca-Aguilar, Mario Bijelic, and Felix Heide. Gated2depth: Real-time Dense Lidar From Gated Images. In *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1506–1516, 2019.

[21] Qurrat-Ul-Ain Nadeem, Abla Kammoun, Anas Chaaban, Mérouane Debbah, and Mohamed-Slim Alouini. Asymptotic Max-Min SINR Analysis of Reconfigurable Intelligent Surface Assisted MISO Systems. *IEEE Transactions on Wireless Communications*, 19(12):7748–7764, 2020. doi: 10.1109/TWC.2020.2986438.

[22] Ertugrul Basar. Reconfigurable Intelligent Surface-Based Index Modulation: A New Beyond MIMO Paradigm for 6G. *IEEE Transactions on Communications*, 68(5):3187–3196, 2020. doi: 10.1109/TCOMM.2020.2971486.

[23] X. Mu, Y. Liu, L. Guo, J. Lin, and N. Al-Dhahir. Exploiting Intelligent Reflecting Surfaces in NOMA Networks: Joint Beamforming Optimization. *IEEE Transactions on Wireless Communications*, 19(10):6884–6898, 2020. doi: 10.1109/TWC.2020.3006915.

[24] Robert W. Heath, Nuria González-Prelcic, Sundeep Rangan, Wonil Roh, and Akbar M. Sayeed. An Overview of Signal Processing Techniques for Millimeter Wave MIMO Systems. *IEEE Journal of Selected Topics in Signal Processing*, 10(3):436–453, April 2016. ISSN 1932-4553.

[25] Ahmed Alkhateeb, Jianhua Mo, Nuria Gonzalez-Prelcic, and Robert W. Heath. MIMO Precoding and Combining Solutions for Millimeter-Wave Systems. *IEEE Communications Magazine*, 52(12):122–131, Dec. 2014. ISSN 0163-6804. doi: 10.1109/MCOM.2014.6979963.

[26] 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 *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8445–8453, 2019.

[27] Yinda Zhang and Thomas Funkhouser. Deep Depth Completion of a Single RGB-D Image. In *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[28] Abdelrahman Taha, Muhammad Alrabeiah, and Ahmed Alkhateeb. Enabling Large Intelligent Surfaces With Compressive Sensing and Deep Learning. *IEEE Access*, 9:44304–44321, Apr 2021. doi: 10.1109/ACCESS.2021.3064073.

[29] Abdelrahman Taha, Muhammad Alrabeiah, and Ahmed Alkhateeb. Deep Learning for Large Intelligent Surfaces in Millimeter Wave and Massive MIMO Systems. In *Proc. of IEEE Global Communications Conference (GLOBECOM)*, pages 1–6, Dec 2019. doi: 10.1109/GLOBECOM38437.2019.9013256.

[30] A. Taha, Y. Zhang, F. B. Mismar, and A. Alkhateeb. Deep Reinforcement Learning for Intelligent Reflecting Surfaces: Towards Standalone Operation. In *Proc. of IEEE International Workshop on Signal Processing Advances in Wireless Communications (SPAWC)*, pages 1–5, May 2020. doi: 10.1109/SPAWC48557.2020.9154301.

[31] Abdelrahman Taha, Qi Qu, Sam Alex, Ping Wang, William L. Abbott, and Ahmed Alkhateeb. Millimeter Wave MIMO-Based Depth Maps for Wireless Virtual and Augmented Reality. *IEEE Access*, 9:48341–48363, 2021. doi: 10.1109/ACCESS.2021.3067839.

[32] Y. Zhang, J. Zhang, Y. Wang, Z. Yu, and B. Zhang. A 4-bit Programmable Metamaterial Based on VO2Mediums. In *Proc. of IEEE/MTT-S International Microwave Symposium (IMS)*, pages 984–986, June 2018.

[33] C. Liaskos, S. Nie, A. Tsioliaridou, A. Pitsillides, S. Ioannidis, and I. Akyildiz. A New Wireless Communication Paradigm through Software-Controlled Metasurfaces. *IEEE Communications Magazine*, 56(9):162–169, Sep. 2018. ISSN 0163-6804.

[34] Minchae Jung, Walid Saad, Youngrok Jang, Gyuyeol Kong, and Sooyong Choi. Performance Analysis of Large Intelligent Surfaces (LISs): Asymptotic Data Rate and Channel Hardening Effects. *IEEE Transactions on Wireless Communications*, 19(3):2052–2065, 2020. doi: 10.1109/TWC.2019.2961990.

[35] Alice Faisal, Hadi Sarieddeen, Hayssam Dahrouj, Tareq Y. Al-Naffouri, and Mohamed-Slim Alouini. Ultramassive MIMO Systems at Terahertz Bands: Prospects and Challenges. *IEEE Vehicular Technology Magazine*, 15(4):33–42, 2020. doi: 10.1109/MVT.2020.3022998.

[36] Elisabeth De Carvalho, Anum Ali, Abolfazl Amiri, Marko Angjelichinoski, and Robert W. Heath. Non-Stationarities in Extra-Large-Scale Massive MIMO. *IEEE Wireless Communications*, 27(4):74–80, 2020. doi: 10.1109/MWC.001.1900157.

[37] Emil Björnson, Luca Sanguinetti, Henk Wymeersch, Jakob Hoydis, and Thomas L. Marzetta. Massive MIMO is a Reality—What is Next?: Five Promising Research Directions for Antenna Arrays. *Digital Signal Processing*, 94: 3–20, 2019. ISSN 1051-2004. URL `https://www.sciencedirect.com/science/article/pii/S1051200419300776`. Special Issue on Source Localization in Massive MIMO.

[38] Luca Sanguinetti, Emil Björnson, and Jakob Hoydis. Toward Massive MIMO 2.0: Understanding Spatial Correlation, Interference Suppression, and Pilot Contamination. *IEEE Transactions on Communications*, 68(1):232–257, 2020. doi: 10.1109/TCOMM.2019.2945792.

[39] H. Wymeersch and B. Denis. Beyond 5G Wireless Localization with Reconfigurable Intelligent Surfaces. In *Proc. of IEEE International Conference on Communications (ICC)*, pages 1–6, June 2020. doi: 10.1109/ICC40277.2020.9148744.

[40] A. Alkhateeb, S. Alex, P. Varkey, Y. Li, Q. Qu, and D. Tujkovic. Deep Learning Coordinated Beamforming for Highly-Mobile Millimeter Wave Systems. *IEEE Access*, 6:37328–37348, 2018. ISSN 2169-3536. doi: 10.1109/ACCESS.2018.2850226.

[41] A. Alkhateeb, I. Beltagy, and S. Alex. Machine Learning for Reliable Mmwave Systems: Blockage Prediction and Proactive Handoff. In *Proc. of IEEE Global Conference on Signal and Information Processing (GlobalSIP)*, pages 1055–1059, Nov 2018.

[42] X. Li, A. Alkhateeb, and C. Tepedelenlioglu. Generative Adversarial Estimation of Channel Covariance in Vehicular Millimeter Wave Systems. In *Proc. of Asilomar Conference on Signals, Systems, and Computers*, pages 1572–1576, Oct 2018.

[43] J. Wang, Z. Lan, C.W. Pyo, T. Baykas, C.S. Sum, M.A. Rahman, J. Gao, R. Funada, F. Kojima, H. Harada, et al. Beam Codebook Based Beamforming

Protocol for Multi-Gbps Millimeter-Wave WPAN Systems. *IEEE Journal on Selected Areas in Communications*, 27(8):1390–1399, Nov. 2009.

[44] S. Hur, T. Kim, D. J. Love, J. V. Krogmeier, T. A. Thomas, and A. Ghosh. Millimeter Wave Beamforming for Wireless Backhaul and Access in Small Cell Networks. *IEEE Transactions on Communications*, 61(10):4391–4403, Oct. 2013. ISSN 0090-6778.

[45] A. Alkhateeb, O. El Ayach, G. Leus, and R. W. Heath. Channel Estimation and Hybrid Precoding for Millimeter Wave Cellular Systems. *IEEE Journal of Selected Topics in Signal Processing*, 8(5):831–846, Oct. 2014. ISSN 1932-4553.

[46] T. S. Rappaport, F. Gutierrez, E. Ben-Dor, J.N. Murdock, Y. Qiao, and J. I. Tamir. Broadband Millimeter-Wave Propagation Measurements and Models Using Adaptive-Beam Antennas for Outdoor Urban Cellular Communications. *IEEE Transactions on Antennas and Propagation*, 61(4):1850–1859, Apr. 2013. ISSN 0018-926X.

[47] T. S. Rappaport, R. W. Heath, R. C. Daniels, and J. N. Murdock. *Millimeter Wave Wireless Communications*. Pearson Education, 2014.

[48] M. K. Samimi and T. S. Rappaport. Ultra-Wideband Statistical Channel Model for Non Line of Sight Millimeter-wave Urban Channels. In *Proc. of IEEE Global Communications Conference (GLOBECOM)*, pages 3483–3489, Dec 2014.

[49] S. Foo. Liquid-crystal Reconfigurable Metasurface Reflectors. In *Proc. of IEEE International Symposium on Antennas and Propagation USNC/URSI*, pages 2069–2070, July 2017. doi: 10.1109/APUSNCURSINRSM.2017.8073077.

[50] Qingqing Wu and Rui Zhang. Intelligent Reflecting Surface Enhanced Wireless Network via Joint Active and Passive Beamforming. *IEEE Transactions on Wireless Communications*, 18(11):5394–5409, 2019.

[51] R. Mendez-Rial, C. Rusu, N. Gonzalez-Prelcic, A. Alkhateeb, and R.W. Heath. Hybrid MIMO Architectures for Millimeter Wave Communications: Phase Shifters or Switches? *IEEE Access*, PP(99):1–1, 2016. ISSN 2169-3536.

[52] J. Lee, G.-T. Gil, and Y. H. Lee. Exploiting Spatial Sparsity for Estimating Channels of Hybrid MIMO Systems in Millimeter Wave Communications. In *Proc. of IEEE Global Communications Conference (GLOBECOM)*, pages 3326–3331, Dec 2014.

[53] T. Cai, W. Liu, and X. Luo. A Constrained l1 Minimization Approach to Sparse Precision Matrix Estimation. *Journal of the American Statistical Association*, 106(494):594–607, 2011.

[54] J. A. Tropp. Greed is Good: Algorithmic Results for Sparse Approximation. *IEEE Transactions on Information Theory*, 50(10):2231–2242, Oct 2004. ISSN 0018-9448.

[55] P. Pal and P. P. Vaidyanathan. Nested Arrays: A Novel Approach to Array Processing With Enhanced Degrees of Freedom. *IEEE Transactions on Signal Processing*, 58(8):4167–4181, Aug 2010. ISSN 1053-587X.

[56] P. P. Vaidyanathan and P. Pal. Sparse Sensing With Co-Prime Samplers and Arrays. *IEEE Transactions on Signal Processing*, 59(2):573–586, Feb 2011. ISSN 1053-587X.

[57] Z. Tan, Y. C. Eldar, and A. Nehorai. Direction of Arrival Estimation Using Co-Prime Arrays: A Super Resolution Viewpoint. *IEEE Transactions on Signal Processing*, 62(21):5565–5576, Nov 2014. ISSN 1053-587X. doi: **10.1109/TSP.2014.2354316**.

[58] C. Rusu, N. Gonzalez-Prelcic, and R. W. Heath. Algorithms for the Construction of Incoherent Frames under Various Design Constraints. *Signal Processing*, 152:363 – 372, 2018. ISSN 0165-1684.

[59] P. Xia, S. Zhou, and G. B. Giannakis. Achieving the Welch bound with difference sets. *IEEE Transactions on Information Theory*, 51(5):1900–1907, May 2005. ISSN 0018-9448.

[60] I. Goodfellow, Y. Bengio, and A. Courville. *Deep Learning*. MIT press, 2016. URL `https://www.deeplearningbook.org/`.

[61] L. Deng, D. Yu, et al. Deep Learning: Methods and Applications. *Foundations and Trends® in Signal Processing*, 7(3–4):197–387, 2014.

[62] Y. A. LeCun, L. Bottou, G. B. Orr, and K.-R. Müller. Efficient Backprop. In *Neural networks: Tricks of the trade*, pages 9–48. Springer, 2012.

[63] K. Hornik, M. Stinchcombe, and H. White. Multilayer Feedforward Networks are Universal Approximators. *Neural Networks*, 2(5):359–366, 1989.

[64] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A. Rusu, Joel Veness, Marc G. Bellemare, Alex Graves, Martin Riedmiller, Andreas K. Fidjeland, Georg Ostrovski, Stig Petersen, Charles Beattie, Amir Sadik, Ioannis Antonoglou, Helen King, Dharshan Kumaran, Daan Wierstra, Shane Legg, and Demis Hassabis. Human-level Control through Deep Reinforcement Learning. *Nature*, 518(7540):529–533, 2015.

[65] Faris B. Mismar, Brian L. Evans, and Ahmed Alkhateeb. Deep Reinforcement Learning for 5G Networks: Joint Beamforming, Power Control, and Interference Coordination. *IEEE Transactions on Communications*, 2019.

[66] Yu Zhang, Muhammad Alrabeiah, and Ahmed Alkhateeb. Deep Learning for Massive MIMO With 1-Bit ADCs: When More Antennas Need Fewer Pilots. *IEEE Wireless Communications Letters*, 9(8):1273–1277, 2020. doi: **10.1109/LWC.2020.2987893**.

[67] Hado Van Hasselt, Arthur Guez, and David Silver. Deep Reinforcement Learning with Double Q-learning. In *Proc. of AAAI conference on artificial intelligence*, 2016.

[68] 2019. URL `https://github.com/Abdelrahman-Taha/LIS-DeepLearning`.

[69] J. Mo, A. Alkhateeb, S. Abu-Surra, and R. W. Heath. Hybrid Architectures with Few-bit ADC Receivers: Achievable Rates and Energy-rate Tradeoffs. *IEEE Transactions on Wireless Communications*, 16(4):2274–2287, 2017.

[70] R. H. Walden. Analog-to-Digital Converter Survey and Analysis. *IEEE Journal on Selected Areas in Communications*, 17(4):539–550, 1999.

[71] B. Murmann. ADC Performance Survey 1997-2019. 2019. URL `https://web.stanford.edu/~murmann/adcsurvey.html`.

[72] Yong Lim and Michael P Flynn. A 100 MS/s, 10.5 Bit, 2.46 mW Comparator-less Pipeline ADC Using Self-biased Ring Amplifiers. *IEEE Journal of Solid-State Circuits*, 50(10):2331–2341, 2015.

[73] VIVE. VIVE Wireless Adapter. 2020. URL `https://www.vive.com/us/wireless-adapter/`.

[74] Intel. Intel RealSense Depth Camera D435. 2020. URL `https://www.intelrealsense.com/depth-camera-d435/`.

[75] J. A. Zhang, A. Cantoni, X. Huang, Y. J. Guo, and R. W. Heath. Joint Communications and Sensing Using Two Steerable Analog Antenna Arrays. In *2017 IEEE 85th Vehicular Technology Conference (VTC Spring)*, pages 1–5, June 2017. doi: 10.1109/VTCSpring.2017.8108565.

[76] Preeti Kumari, Junil Choi, Nuria González-Prelcic, and Robert W. Heath. IEEE 802.11ad-Based Radar: An Approach to Joint Vehicular Communication-Radar System. *IEEE Transactions on Vehicular Technology*, 67(4):3012–3027, 2018.

[77] Jaime Lien, Nicholas Gillian, M. Emre Karagozler, Patrick Amihood, Carsten Schwesig, Erik Olson, Hakim Raja, and Ivan Poupyrev. Soli: Ubiquitous Gesture Sensing with Millimeter Wave Radar. *ACM Trans. Graph.*, 35(4), July 2016. ISSN 0730-0301. URL `https://doi.org/10.1145/2897824.2925953`.

[78] Teng Wei and Xinyu Zhang. MTrack: High-Precision Passive Tracking Using Millimeter Wave Radios. In *Proc. of the Annual International Conference on Mobile Computing and Networking*, pages 117–129, 2015.

[79] Ashutosh Sabharwal, Philip Schniter, Dongning Guo, Daniel W. Bliss, Sampath Rangarajan, and Risto Wichman. In-Band Full-Duplex Wireless: Challenges and Opportunities. *IEEE Journal on Selected Areas in Communications*, 32(9): 1637–1652, 2014.

[80] Nicholas A Estep, Dimitrios L Sounas, Jason Soric, and Andrea Alù. Magnetic-Free Non-Reciprocity and Isolation Based on Parametrically Modulated Coupled-Resonator Loops. *Nature Physics*, 10(12):923–927, 2014.

[81] Tolga Dinc and Harish Krishnaswamy. A 28GHz Magnetic-Free Non-Reciprocal Passive CMOS Circulator Based on Spatio-Temporal Conductance Modulation. In *Proc. of IEEE International Solid-State Circuits Conference (ISSCC)*, pages 294–295. IEEE, 2017.

[82] Jin Zhou, Negar Reiskarimian, and Harish Krishnaswamy. Receiver with Integrated Magnetic-Free N-Path-Filter-Based Non-Reciprocal Circulator and Baseband Self-Interference Cancellation for Full-Duplex Wireless. In *Proc. of IEEE International Solid-State Circuits Conference (ISSCC)*, pages 178–180. Institute of Electrical and Electronics Engineers Inc., 2016.

[83] Aravind Nagulu and Harish Krishnaswamy. Non-Magnetic 60GHz SOI CMOS Circulator Based on Loss/Dispersion-Engineered Switched Bandpass Filters. In *Proc. of IEEE International Solid-State Circuits Conference (ISSCC)*, pages 446–448. IEEE, 2019.

[84] Mark A. Richards, James A. Scheer, and William A. Holm. *Principles of Modern Radar: Basic Principles*. SciTech Publishing, Raleigh, NC, USA, 2010.

[85] Preeti Kumari, Sergiy A. Vorobyov, and Robert W. Heath. Adaptive Virtual Waveform Design for Millimeter-Wave Joint Communication-Radar. *IEEE Transactions on Signal Processing*, 68:715–730, 2020. doi: 10.1109/TSP.2019.2956689.

[86] Q.H. Spencer, B.D. Jeffs, M.A. Jensen, and A.L. Swindlehurst. Modeling The Statistical Time and Angle of Arrival Characteristics of an Indoor Multipath Channel. *IEEE Journal on Selected Areas in Communications*, 18(3):347–360, 2000. doi: 10.1109/49.840194.

[87] Peter F. M. Smulders. Statistical Characterization of 60-GHz Indoor Radio Channels. *IEEE Transactions on Antennas and Propagation*, 57(10):2820–2829, 2009.

[88] O. El Ayach, S. Rajagopal, S. Abu-Surra, Zhouyue Pi, and R.W. Heath. Spatially Sparse Precoding in Millimeter Wave MIMO Systems. *IEEE Transactions on Wireless Communications*, 13(3):1499–1513, Mar. 2014. ISSN 1536-1276. doi: 10.1109/TWC.2014.011714.130846.

[89] Anna Guerra, Francesco Guidi, Davide Dardari, Antonio Clemente, and Raffaele D'Errico. A Millimeter-Wave Indoor Backscattering Channel Model for Environment Mapping. *IEEE Transactions on Antennas and Propagation*, 65 (9):4935–4940, 2017.

[90] Steven M. Kay. *Fundamentals of Statistical Signal Processing: Estimation Theory*. Prentice Hall, 1993.

[91] Patrick Bidigare, Upamanyu Madhow, Raghu Mudumbai, and Dzul Scherber. Attaining Fundamental Bounds on Timing Synchronization. In *Proc. of International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5229–5232, 2012.

[92] Andrew Herschfelt, Hanguang Yu, Shunyao Wu, Hyunseok Lee, and Daniel W. Bliss. Joint Positioning-Communications System Design: Leveraging Phase-Accurate Time-of-Flight Estimation and Distributed Coherence. In *Proc. of Asilomar Conference on Signals, Systems, and Computers (ACSSC)*, pages 433–437, 2018.

[93] Andrew Herschfelt. *Simultaneous Positioning and Communications: Hybrid Radio Architecture, Estimation Techniques, and Experimental Validation*. PhD thesis, Arizona State University, 2019.

[94] Ahmed Alkhateeb and Robert W. Heath. Frequency Selective Hybrid Precoding for Limited Feedback Millimeter Wave Systems. *IEEE Transactions on Communications*, 64(5):1801–1818, 2016.

[95] William L. Melvin and James A. Scheer. *Principles of Modern Radar Vol. II: Advanced Techniques*. SciTech Publishing, Edison, NJ, USA, 2013.

[96] Moawad Ibrahim Dessouky, Hamdy Sharshar, and Yasser Albagory. Efficient Sidelobe Reduction Technique for Small-sized Concentric Circular Arrays. *Progress In Electromagnetics Research*, 65:187–200, 2006.

[97] Yasser Attia Albagory, Moawad Dessouky, and Hamdy Sharshar. An Approach for Low Sidelobe Beamforming in Uniform Concentric Circular Arrays. *Wireless Personal Communications*, 43(4):1363–1368, 2007.

[98] Emanuele Grossi, Marco Lops, and Luca Venturino. Detection and Localization of Multiple Targets in IEEE 802.11ad Networks. In *Proc. of Asilomar Conference on Signals, Systems, and Computers (ACSSC)*, 2019.

[99] Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. Indoor Segmentation and Support Inference from RGBD Images. In *Proc. of European Conference on Computer Vision (ECCV)*, pages 746–760, 2012.

[100] Xiaoyan Ying, Umut Demirhan, and Ahmed Alkhateeb. Relay Aided Intelligent Reconfigurable Surfaces: Achieving the Potential Without So Many Antennas. 2020. arXiv:2006.06644.

[101] Yu Zhang and Ahmed Alkhateeb. Learning Reflection Beamforming Codebooks for Arbitrary RIS and Non-Stationary Channels. 2021. arXiv:2109.14909.

[102] Augusto Aubry, Antonio De Maio, and Massimo Rosamilia. Reconfigurable Intelligent Surfaces for N-LOS Radar Surveillance. *IEEE Transactions on Vehicular Technology*, 70(10):10735–10749, 2021.

[103] Haobo Zhang, Hongliang Zhang, Boya Di, Kaigui Bian, Zhu Han, and Lingyang Song. MetaRadar: Multi-Target Detection for Reconfigurable Intelligent Surface Aided Radar Systems. *IEEE Transactions on Wireless Communications*, 21(9):6994–7010, 2022. ISSN 1558-2248.

[104] Stefano Buzzi, Emanuele Grossi, Marco Lops, and Luca Venturino. Foundations of MIMO Radar Detection Aided by Reconfigurable Intelligent Surfaces. *IEEE Transactions on Signal Processing*, 70:1749–1763, 2022.

[105] Lianlin Li, Hengxin Ruan, Che Liu, Ying Li, Ya Shuang, Andrea Alù, Cheng-Wei Qiu, and Tie Jun Cui. Machine-Learning Reprogrammable Metasurface Imager. *Nature Communications*, 10(1):1082–1089, 2019.

[106] Timothy Sleasman, Mohammadreza F. Imani, Aaron V. Diebold, Michael Boyarsky, Kenneth P. Trofatter, and David R. Smith. Computational Imaging With Dynamic Metasurfaces: A Recipe for Simple and Low-Cost Microwave Imaging. *IEEE Antennas and Propagation Magazine*, 64(4):123–134, 2022.

[107] Ying He, Dongheng Zhang, and Yan Chen. High-Resolution WiFi Imaging with Reconfigurable Intelligent surfaces. *IEEE Internet of Things Journal*, 2022. 10.1109/JIOT.2022.3210686.

[108] Jingzhi Hu, Hongliang Zhang, Kaigui Bian, Zhu Han, H. Vincent Poor, and Lingyang Song. MetaSketch: Wireless Semantic Segmentation by Reconfigurable Intelligent Surfaces. *IEEE Transactions on Wireless Communications*, 21(8):5916–5929, 2022.

[109] Umut Demirhan and Ahmed Alkhateeb. Integrated Sensing and Communication for 6G: Ten Key Machine Learning Roles. *arXiv preprint arXiv:2208.02157*, 2022.

[110] Junfeng Guan, Sohrab Madani, Suraj Jog, Saurabh Gupta, and Haitham Hassanieh. Through Fog High-Resolution Imaging Using Millimeter Wave Radar. In *Proc. of IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11464–11473, 2020.

[111] Anjun Chen, Xiangyu Wang, Shaohao Zhu, Yanxu Li, Jiming Chen, and Qi Ye. mmBody Benchmark: 3D Body Reconstruction Dataset and Analysis for Millimeter Wave Radar. In *Proc. of ACM International Conference on Multimedia*, pages 3501–3510, 2022.

[112] Blender Online Community. *Blender - A 3D Modelling and Rendering Package*. Blender Foundation, Blender Institute, Amsterdam, 2022. URL `http://www.blender.org`.

[113] Karthik Ramasubramanian. Using a Complex-Baseband Architecture in FMCW Radar Systems. Technical report, Dallas, TX, USA, May 2017.

APPENDIX A

PREVIOUSLY PUBLISHED WORK

The publications related to the dissertation are as follows.

1. A. Taha, M. Alrabeiah and A. Alkhateeb, "Enabling Large Intelligent Surfaces With Compressive Sensing and Deep Learning," in IEEE Access, vol. 9, pp. 44304-44321, 2021, doi: 10.1109/ACCESS.2021.3064073.

2. A. Taha, Q. Qu, S. Alex, P. Wang, W. L. Abbott and A. Alkhateeb, "Millimeter Wave MIMO-Based Depth Maps for Wireless Virtual and Augmented Reality," in IEEE Access, vol. 9, pp. 48341-48363, 2021, doi: 10.1109/ACCESS.2021.3067839.

3. A. Taha, H. Luo, and A. Alkhateeb, "Reconfigurable Intelligent Surface Aided Wireless Sensing for Scene Depth Estimation," in arXiv preprint arXiv:2211.08210, Nov. 2022. [Online]. Available: https://arxiv.org/abs/2211.08210.

All co-authors of these publications have granted their permission to use the articles in the dissertation