OUKA
The University of Osaka
Institutional Knowledge Archive

Title	A Study on Wireless Sensing for Indoor Environments: From Open Resource Foundations to Deployment-Aware Methods
Author(s)	加藤, 空知
Citation	大阪大学, 2026, 博士論文
Version Type	VoR
URL	https://doi.org/10.18910/105503
rights	
Note	

The University of Osaka Institutional Knowledge Archive : OUKA

https://ir.library.osaka-u.ac.jp/

The University of Osaka

A Study on Wireless Sensing for Indoor Environments: From Open Resource Foundations to Deployment-Aware Methods

Submitted to
Graduate School of Information Science and Technology
The University of Osaka

January 2026

Sorachi KATO

List of Related Publications

Journal Articles

1. Sorachi Kato, Pu (Perry) Wang, Toshiaki Koike-Akino, Takuya Fujihashi, Hassan Mansour, Petros Boufounos, "Multi-Band Wi-Fi Neural Dynamic Fusion," *IEEE Transactions of Wireless Communications*, vol. 24, no. 5, pp. 3698–3714, 2025.

2. Sorachi Kato, Tomoki Murakami, Takuya Fujihashi, Takashi Watanabe, Shunsuke Saruwatari, "CBR-ACE: Counting Human Exercise using Wi-Fi Beamforming Reports," *Journal of Information Processing*, vol. 30, pp. 66–74, 2022.

3. Sorachi Kato, Takeru Fukushima, Tomoki Murakami, Hirantha Abeysekera, Yusuke Iwasaki, Takuya Fujihashi, Takashi Watanabe, Shunsuke Saruwatari, "CSI2Image: Image Reconstruction from Channel State Information Using Generative Adversarial Networks," *IEEE Access*, vol. 9, pp. 47154–47168, 2021.

International Conference Papers

1. Sorachi Kato, Ryoma Yataka, Pu (Perry) Wang, Pedro Miraldo, Takuya Fujihashi, Petros Boufounos, "RAPTR: Radar-based 3D Pose Estimation using Transformer," in *the 39th Annual Conference on Neural Information Processing Systems (NeurIPS)*, pp. 1–26, 2025.

2. M. Mahbubur Rahman* , Ryoma Yataka*, Sorachi Kato*, Pu (Perry) Wang*, Peizhao Li, Adriano Cardace, Petros Boufounos, "MMVR: Millimeter-wave Multi-View Radar Dataset and Benchmark for Indoor Perception," in *European Conference on Computer Vision (ECCV)*, pp. 306–322, 2024.

3. Sorachi Kato, Pu (Perry) Wang, Toshiaki Koike-Akino, Takuya Fujihashi, Hassan Mansour, Petros Boufounos, "Object Trajectory Estimation with Multi-Band Wi-Fi Neural Dynamic Fusion," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 13261–13265, 2024.

* Equal contribution.

Abstract

Wireless sensing, which leverages radio signals to infer subject states and activities, has emerged as a privacy-preserving and environment-robust alternative to cameras and wearable devices. Unlike vision-based approaches, wireless signals are insensitive to lighting or occlusions, and unlike inertial sensors, they impose no wearing or charging requirements while enabling non-invasive monitoring of entire bodies or multiple subjects simultaneously. These characteristics are particularly appealing in indoor environments, where most human activities occur and privacy-safe perception is highly demanded. In this domain, two major sensing paradigms have emerged: Wi-Fi-based sensing, which leverages existing communication infrastructure, and mmWave radar sensing, which enables fine-grained and reliable measurements. These modalities are particularly attractive in scenarios where non-intrusive, infrastructure-compatible perception is required, including indoor environments that involve daily human activities, assistive monitoring, and autonomous systems.

However, despite significant progress, much prior work remains confined to laboratory settings, and practical deployment is hindered by the limited accessibility and diversity of sensing data. Full channel state information (CSI) has been a central signal representation in Wi-Fi sensing research, as it reflects detailed multipath propagation. However, accessing CSI often requires non-standard hardware configurations or firmware modifications, making large-scale and reproducible data collection difficult. Radar sensing faces a related difficulty: although compact FMCW modules have become increasingly available, data acquisition still depends on dedicated hardware and controlled setups. In this situation, researchers often rely on publicly available datasets for model training and evaluation. However, existing resources are typically limited in scale, environmental diversity, participant variation, and annotations of downstream tasks, often confined to a few controlled environments. Consequently, across both modalities, open and representative data access remains scarce, restricting reproducibility and hindering generalization in everyday indoor settings. These limitations reveal two fundamental and complementary challenges. First, wireless signals must be both openly accessible and representative: collected from commodity devices and across diverse, realistic environments. Second, on top of such data, methodologies must explicitly account for the constraints of realistically obtainable wireless measurements and annotations.

In this dissertation, we address both directions, spanning from data acquisition to advanced

sensing methodologies in Wi-Fi and radar modalities. On the side of data accessibility, we make two contributions. First, in Wi-Fi sensing, we exploit Compressed Beamforming Reports (CBRs), natively supported by the IEEE 802.11ac/ax standards, as a practical alternative to full CSI. We develop WiPiCap, a tool that enables commodity devices to capture and reconstruct CBRs, lowering the barrier to collecting Wi-Fi sensing data. Second, for radar-based perception, we construct MMVR, a large-scale multi-view dataset with synchronized annotations for detection, pose estimation, and segmentation. To our knowledge, MMVR is the largest indoor human perception dataset built on radar signals, providing diversity and serving as a crucial foundation for advancing perception in cluttered, multi-subject environments. Together, WiPiCap and MMVR serve as foundational resources: the former democratizes the collection of standardized Wi-Fi signals, while the latter offers scale and breadth for radar-based indoor perception.

On the methodological side, the dissertation focuses on developing frameworks that effectively leverage accessible wireless data. For Wi-Fi sensing, we take two complementary perspectives. First, to understand what can be achieved with accessible and standardized feedback, we investigate two approaches based on CBRs. CBR-ACE adopts a lightweight signal processing–based framework tailored to the characteristics of CBRs, demonstrating effective human activity estimation even under device-agnostic conditions. In contrast, CSI2Image employs a generative adversarial network to reconstruct physical-space images from CBRs, illustrating that such standardized feedback can support more advanced spatial representations. For radar-based human-centric sensing, we propose RAPTR, a method for 3D human pose estimation with weak supervision. It achieves 3D pose estimation without expensive motion-capture annotations, instead relying on 2D keypoints and 3D bounding boxes as supervision. RAPTR introduces a two-stage decoding process guided by structured loss functions that progressively recover feasible human poses from weak labels, and offers a realistic path toward scalable radar-based pose estimation. We further propose a neural dynamic fusion (NDF) framework that combines sub-7 GHz CSI and 60 GHz SNR measurements introduced in recent IEEE 802.11 standards. NDF models these asynchronous multi-band signals as latent dynamic processes that are aligned via neural ordinary differential equations (ODEs) and enables continuous device trajectory estimation. This approach extends the dissertation's scope, illustrating how multi-band integration can enhance dynamic perception in realistic indoor settings.

Collectively, this dissertation advances wireless sensing from data accessibility to methodology, from Wi-Fi to radar, thereby bringing wireless sensing closer to practical deployment in real-world indoor environments. We further examine the complementary characteristics and constraints of radar- and Wi-Fi-based sensing, and outline several open challenges and research opportunities in wireless sensing.

Contents

1 Introduction **13**

1.1 Background and Motivation . . . 13

1.2 Related Works . . . 15

1.2.1 Wi-Fi Sensing . . . 15

1.2.2 mmWave Radar Sensing . . . 17

1.3 Challenges to Practical Wireless Sensing . . . 18

1.3.1 Open Accessibility of Sensing Data . . . 18

1.3.2 Deployment-aware Methodological Design . . . 20

1.4 Positioning, Scope, and Contributions . . . 21

1.5 Thesis Organization . . . 23

2 Data Foundations for Practical Wireless Sensing **24**

2.1 WiPiCap: A Tool for Wi-Fi CBR Collection . . . 24

2.1.1 Overview of Compressed Beamforming Report . . . 24

2.1.2 Example of Compression Process . . . 27

2.1.3 Design and Implementation of WiPiCap . . . 28

2.1.4 Sampling Rate Benchmark . . . 30

2.1.5 Motion Response Verification . . . 31

2.1.6 Conclusion and Contribution . . . 32

2.2 MMVR: A Large-Scale Multi-View Radar Dataset for Indoor Perception . . . 32

2.2.1 Introduction . . . 32

2.2.2 Existing Indoor Radar Perception Datasets . . . 36

2.2.3 MMVR Sensor Setup . . . 36

2.2.4 Dataset . . . 39

2.2.5 Evaluation and Benchmarks . . . 43

2.2.6 Main Results . . . 44

2.2.7 Ablation Study . . . 46

2.2.8 Conclusion and Contribution . . . 47

3 Wi-Fi Sensing with Standard-compliant Signal Formats **48**
3.1 CBR-ACE: Activity Recognition with Standardized Beamforming Reports 48
3.1.1 Introduction 48
3.1.2 Proposed System: CBR-ACE 50
3.1.3 Experiments 53
3.1.4 Conclusion and Contribution 58
3.2 CSI2Image: Generative Modeling of Spatial Representations using CBR 58
3.2.1 Introduction 58
3.2.2 Preliminary: Generative adversarial networks 60
3.2.3 CSI2Image 61
3.2.4 Evaluation 65
3.2.5 Computational Costs 73
3.2.6 Conclusion and Contribution 74

4 Radar-based 3D Pose Estimation with Annotation Constraints **75**
4.1 RAPTR 75
4.1.1 Introduction 75
4.1.2 Preliminary 77
4.1.3 RAPTR: Radar-based 3D Pose Estimation using Transformer 78
4.1.4 Evaluation 82
4.1.5 Ablation Study 86
4.1.6 Conclusion and Contribution 89

5 Propagation-aware Dynamic Multiband Fusion for Sub-7GHz and 60GHz **90**
5.1 Multi-band Neural Fusion 90
5.1.1 Introduction 90
5.1.2 Problem Formulation 93
5.1.3 Existing Fusion Strategies 94
5.1.4 Multi-Band Neural Dynamic Fusion 96
5.1.5 Performance Evaluation 108
5.1.6 Conclusion and Contribution 119

6 Discussion **121**
6.1 Remaining Limitations 121
6.1.1 Methodological Limitations 121
6.1.2 Community-Level Challenges 123
6.2 Future Prospects for Real-World Deployment 123

6.2.1 Methodological Advancements for Generalization 124
6.2.2 Toward Coexisting and Task-Adaptive Modalities 125
6.2.3 Establishing Community-Unified RF Data Format 125
6.2.4 Grounded Foundation Models for Wireless Sensing 127
6.2.5 Ethical and Privacy Considerations 128

7 Conclusion **130**

List of Figures

1.1 Overview of wireless sensing. 14

2.1 IEEE 802.11ac/ax channel sounding procedure for CBR acquisition. 26
2.2 System setup of WiPiCap. 29
2.3 Temporal response of CBR for human-induced motion. 31
2.4 Comparison of radar signal representations before and after CFAR processing. . . . 33
2.5 Snapshots of indoor radar heatmap datasets. 34
2.6 MMVR sensor configuration. 37
2.7 Comparison between low- and high-resolution radar heatmaps. 38
2.8 Camera-Radar coordinate calibration. 38
2.9 Floorplans of 6 indoor environments in Protocol 1 and Protocol 2. 39
2.10 Visualization of data collection process. 40
2.11 Annotation curation. 41
2.12 Quality of curated annotations. 41
2.13 Multi-view heatmap preprocessing. 42
2.14 AP breakdown over keypoints under 2 protocols and 2 data splits. 45

3.1 Overall structure of CBR-ACE. 49
3.2 Relationship between subcarrier variance and estimation results. 51
3.3 A distribution of estimation values emerged from all subcarriers in a channel. . . . 52
3.4 Experiment environment. 53
3.5 Accuracy improvement with filters and KDE. 54
3.6 Comparison between the variance based method and proposed CBR-ACE. 55
3.7 Application example of CSI2Image with object detection. 59
3.8 DCGAN architecture. 60
3.9 Overview of CSI2Image architecture. 61
3.10 The layer structure of the generator network. 61
3.11 The layer structure of the discriminator network. 62
3.12 Generator-only learning. 64

3.13 GAN-only learning. 64
3.14 Stages in hybrid learning. 64
3.15 Experimental environment for object classification. 65
3.16 Experiment 1: Examples of successful object classification. 66
3.17 Experiment 1: Confusion matrix for object classification based on the image reproduction. 66
3.18 Experimental setting for human localization. 67
3.19 Snapshot of the experiment environment for human localization. 67
3.20 Experiment 2: Successful examples of localization against one user. 68
3.21 Experiment 2: Failed examples of localization against one user. 68
3.22 Experiment 2: Quantitative evaluation of a single-user localization 69
3.23 Experiment 2: Successful examples of localization against up to two users. 70
3.24 Experiment 2: Failed examples of localization against up to two users. 71
3.25 Experiment 2: Quantitative evaluation of the position detection of one or two users. 71
3.26 Experiment 2: Single-user continuous position estimation. 72
3.27 Training duration for device-free localization of one user (32000 epochs). 73

4.1 The concept of RAPTR. 77
4.2 Multi-view radar heatmaps. 77
4.3 The overview of RAPTR architecture. 78
4.4 The pseudo-3D deformable attention mechanism. 80
4.5 Qualitative results. 85
4.6 Pose refinement process in RAPTR. 86

5.1 Wi-Fi fusion from frame-to-frame basis to sequence-to-sequence basis. 91
5.2 The network architecture of the multi-band Wi-Fi neural dynamic fusion. 94
5.3 CSI encoder module. 98
5.4 Unrolling latent dynamic ODEs for latent state alignment and recovery. 99
5.5 Post-ODE fusion schemes considered in this work. 101
5.6 A multi-encoder, multi-decoder NDF structure with multiple integrated decoders. . 103
5.7 In-house multi-band Wi-Fi testbed. 108
5.8 The device connection and data flow within the testbed. 109
5.9 The histogram of number of CSI and beam SNR across non-overlapping sequences. 109
5.10 Hyperparameter search for regularization parameters in the ELBO-based loss function. 111
5.11 Cumulative distribution functions of localization errors for the baseline methods and NDF. 113
5.12 Estimated trajectories for representative baselines and the proposed NDF. 114

5.13 Estimated trajectories in unseen regions. 116
5.14 Two more complex trajectories to assess performance beyond the training distribution. 118
5.15 Cross-trajectory localization results after training on a checkerboard trajectory. . . 118
5.16 Two-dimensional projections of latent representations corresponding to different spatial regions. 119

List of Tables

1.1 Representative Wi-Fi sensing studies. 16
1.2 Representative radar sensing studies. 17

2.1 Effective sampling rates across different AP–client pairs. 30
2.2 Indoor Radar Perception Datasets. 33
2.3 Comparison of multi-view heatmap-based indoor radar datasets. 35
2.4 Baseline benchmarks for 3 tasks under 2 protocols and 2 data splits. 45
2.5 Effect of the number of frames on performance. 46
2.6 Effect of the number of radar views on performance. 47
2.7 Comparison of Detection Backbones. 47
2.8 Baseline benchmarks between HIBER and MMVR. 47

3.1 Micro benchmark on the effect of KDE with fewer samples 56
3.2 MAE performance using various percentile points of the estimated density. 57
3.3 Average CBRs acquisition per second for each device. 57

4.1 Hyper parameters for RAPTR. 83
4.2 3D pose estimation performance on HIBER. 84
4.3 Pose estimation performance on MMVR. 84
4.4 Effect of loss terms for RAPTR. 87
4.5 Effect of deformable attention mechanisms for RAPTR. 87
4.6 Comparison with a 2D-to-3D pose uplifting model. 88
4.7 Effect of view selection patterns on the performance. 88

5.1 Summary of experimental parameters. 110
5.2 Localization errors for all evaluated baselines and the proposed NDF. 113
5.3 Impact of sequence length. 115
5.4 The impact of the three fusion schemes on the estimation error. 115
5.5 Localization error under temporal split. 116

5.6 Quantitative comparison of trajectory similarity at unseen locations using the Fréchet distance. 117
5.7 Fréchet distance between curves in ground truth and estimations. 117

List of Algorithms

1 Generator-only learning 63
2 GAN-only learning 63
3 Hybrid Learning 64

Chapter 1

Introduction

1.1 Background and Motivation

With the widespread integration of digital technologies into everyday life, interactions between physical environments and computational systems have become increasingly intertwined. Emerging infrastructures such as digital twins, smart homes, and embodied AI systems depend on continuous, multimodal sensing to model and respond to human activities [1–4]. In this evolution, sensing is no longer a peripheral function but a foundational layer that anchors the digital world to physical reality.

Over the past decade, diverse modalities have been explored to capture subject presence, motion, and interaction. Vision-based systems have achieved impressive progress in activity recognition, tracking, and pose estimation [5–8]. Wearable and acoustic sensors have likewise extended the range of measurable human states [9,10].

While sensing technologies are often expected to serve as a bridge between the digital and physical worlds by observing real-world activities as comprehensively as possible, such expectations are constrained in practice. Camera-based systems require unobstructed visibility and raise privacy concerns; wearable devices depend on user compliance and sustained power supply; and acoustic sensing is highly susceptible to environmental noise and reverberation. As a consequence, substantial portions of everyday life—such as private, occluded, or poorly lit spaces—remain unseen and unsensed, inherently difficult to sense in a continuous and socially acceptable manner.

Bridging this sensing void is essential for achieving human-centric intelligence that operates seamlessly within real environments. In this context, wireless sensing offers a fundamentally different paradigm: it does not observe surfaces, but rather interprets the invisible imprints that objects and people leave on the surrounding electromagnetic field. Every presence, movement, and interaction perturbs the radio environment, embedding traces of its existence within the fabric of space. By capturing and learning from these perturbations, machines can perceive the world through its physical influence on the shared wireless medium, even in the absence of vision or contact. This

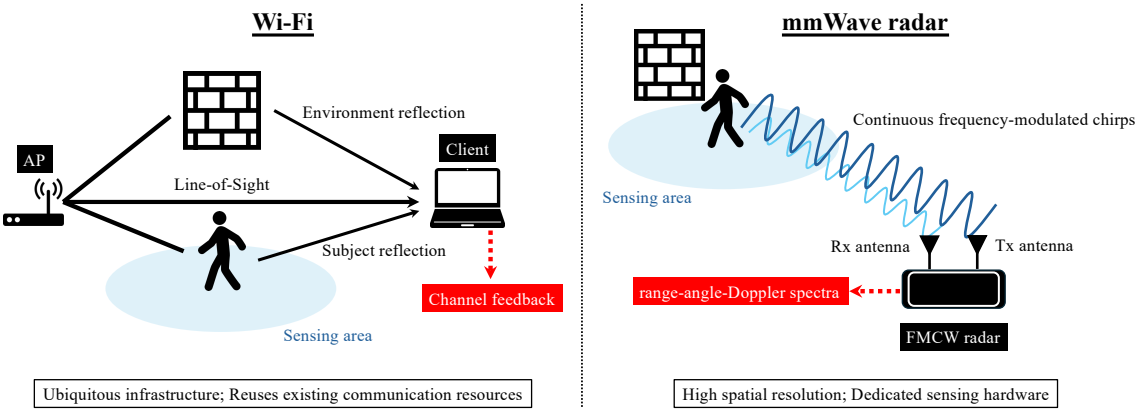


Figure 1.1: Overview of wireless sensing. (left) Wi-Fi sensing utilizes widely deployed communication infrastructure to capture environmental and human-induced multipath reflections, or to infer device states, using channel feedback such as CSI, CBR, or beam SNR, (right) mmWave radar sensing actively transmits frequency-modulated continuous wave chirps and reconstructs range-angle-Doppler spectra via dedicated radar hardware for high-resolution spatial perception.

concept aligns with the broader trajectory of wireless evolution itself. As the global communication infrastructure expands, radio waves, once used purely for data transfer, are being redefined as a dual-purpose medium that simultaneously connects and perceives. The vision of Integrated Sensing and Communication (ISAC), a central pillar of 6G research [11], envisions networks that function as distributed perceptual systems, transforming the radio spectrum into a ubiquitous sensor that records and reveals the dynamics of the physical world. Within this emerging landscape, wireless sensing complements vision and wearables as an indispensable layer that fills the unsensed spaces of everyday environments.

As a practical instantiation of this vision, indoor wireless perception today materializes primarily through two complementary modalities — **Wi-Fi**, which repurposes ubiquitous communication feedback, and **mmWave FMCW radar**, which provides dedicated wide-band measurements — each filling different parts of the sensing void introduced above (Figure 1.1). Wi-Fi sensing has attracted broad interest in both academia and industry because it can piggyback on ubiquitous wireless infrastructure. Early studies in the late 2000s and early 2010s primarily made use of received signal strength indicator (RSSI) measurements with limited granularity. RSSI enables simple tasks such as presence detection or rough localization, but is fundamentally limited by its low resolution. To move beyond these constraints, later research has shifted toward channel state information (CSI), through which exposes subcarrier-wise amplitude and phase components in systems based on orthogonal frequency-division multiplexing (OFDM) systems with multiple-input multiple-output (MIMO) antennas. Formally, CSI on the subcarrier k between a transmit–receive antenna pair can be modeled as

$$H_k = \sum_{\ell} \alpha_\ell e^{-j2\pi f_k \tau_\ell},$$

where α_ℓ represents the complex path gain, τ_ℓ corresponds to the propagation delay of the ℓ-th path, and f_k denotes the center frequency. Because measurements span tens to hundreds of subcarriers and multiple transmit–receive antenna pairs, CSI offers much finer-grained frequency- and space-domain information of the complex multipath structure, enabling applications ranging from gesture recognition to respiration monitoring and fine-grained localization (Section 1.2).

In contrast, mmWave radar sensing is built on frequency-modulated continuous wave (FMCW) technology, whose origins date back to military and automotive applications. These early systems demonstrated precise range and velocity estimation, but their cost and hardware complexity confined them to specialized use. The large bandwidth at tens of gigahertz offers centimeter-level range precision, while Doppler shifts provide accurate velocity estimates. Modern radars employ MIMO antenna arrays, enabling fine-grained angle-of-arrival estimation and effectively synthesizing a large virtual aperture. The received FMCW signal after dechirping f_b yields a beat frequency

$$f_b = \frac{2R}{c} S,$$

where R is the target range, c is the speed of light, and S is the chirp slope. The beat frequency thus directly encodes distance, while Doppler-induced frequency shifts further reveal the target velocity. An antenna array then exploits inter-element phase differences $\Delta\phi$ for a target at azimuth θ given by

$$\Delta\phi = \frac{2\pi d \sin\theta}{\lambda},$$

with antenna spacing d and wavelength λ. Together, range, Doppler, and angular estimates form the basis for 3D spatial perception. Over the past decade, compact industrial- and consumer-grade mmWave modules and ready-to-use evaluation kits have made high-resolution radar more readily available, lowering the barrier to prototyping and data collection. Whereas early civilian use cases centered on outdoor scenarios such as autonomous driving perception [12], recent years have seen a growing shift toward indoor perception tasks such as gesture recognition, respiration monitoring, and full-body pose estimation [13, 14]. These capabilities make radar suitable for a wide spectrum of human-centric applications (Section 1.2).

Despite these advances, both Wi-Fi and radar sensing still face critical challenges. To situate these issues in the broader research landscape, the following section surveys related works on Wi-Fi and radar-based indoor perception, before outlining the key challenges that motivate this dissertation.

1.2 Related Works

1.2.1 Wi-Fi Sensing

Research on Wi-Fi sensing can be broadly organized along the dimensions of signal formats and modeling approaches. In terms of applications, Wi-Fi sensing has enabled a diverse set of indoor

Table 1.1: Representative Wi-Fi sensing studies for indoor perception categorized by signal format.

Format	Key characteristics	Application	Ref.
CSI	Per-subcarrier complex frequency response from OFDM-MIMO; requires specific hardware/driver	Human activity recognition	[15–23]
		Localization / Tracking	[24] (ours), [25–31]
		Human pose / mesh estimation	[32–37]
		Identification	[38–44]
		Vital sensing	[45–51]
		Material sensing	[52–56]
		Indoor scene recovery	[57–63]
		Dataset papers	[21, 64–71]
Beam SNR (IEEE 802.11ad/ay)	Beam-training feedback on mmWave link; show strong directivity but less expressive than CSI	Localization / Tracking	[24] (ours), [72–78]
		Gesture recognition	[79]
		Human posture classification	[77, 80, 81]
CBR (IEEE 802.11ac/ax)	Quantized and compressed channel feedback on IEEE 802.11ac/ax; no hardware modification needed	Human activity recognition	[82] (ours), [83–85]
		Human presence and motion detection	[86–88]
		Localization / Tracking	[89–91]
		Vital sensing	[89, 92]
		Material sensing	[93]
		Indoor scene recovery	[94] (ours), [58, 60]

perception tasks, including human activity monitoring, vital sensing, localization and presence detection, as well as emerging tasks such as user identification and indoor scene reconstruction. Representative applications are summarized in Table 1.1.

Early studies exploited RSSI, a coarse measure of signal attenuation readily available on commodity devices. RSSI supported simple tasks such as coarse device localization and basic human activity recognition [95–97], but its low resolution and sensitivity to environmental noise fundamentally limited its utility. To overcome these limitations, research in the early 2010s turned to Wi-Fi CSI, which exposes per-subcarrier amplitude and phase in OFDM-MIMO systems. CSI provides access to the multipath structure of wireless propagation in far greater detail than RSSI. Toolkits such as the Intel 5300 CSI tool [98] facilitated the collection of CSI on commodity network interface cards (NICs), or USRP (Universal Software Radio Peripheral)-based setups, fueling a surge of work on indoor perception. Beyond widely studied tasks such as activity recognition, localization, and vital sign monitoring, CSI has been extended to more advanced applications. For example, Person-in-WiFi [37] demonstrated body segmentation and 2D pose estimation from CSI, and more recently WiPose [36] and Person-in-WiFi 3D [32] achieved 3D pose estimation. In material sensing, WiMeasure [52] demonstrated millimeter-level subject size measurements without requiring any learning or training. Moreover, recent datasets such as WiMANS [65], a multi-user CSI dataset for activity sensing, broaden the contexts in which CSI-based methods are evaluated. These developments underscore CSI's potential to support increasingly complex human-centric sensing tasks. Still, reliance on specialized NICs, modified drivers, or custom hardware raises

Table 1.2: Representative radar sensing studies for indoor perception categorized by representation format.

Radar Format	Application	Ref.
Point Clouds	Human activity recognition	[99–105]
	Human detection / segmentation	[106,107]
	Human pose / mesh estimation	[99,108–113]
	Localization / Tracking	[114–116]
	Vital sensing	[117]
	Papers with public datasets	[69,105,112,118–124]
Heatmaps	Human activity recognition	[125–129]
	Human detection / segmentation	[130–132]
	Localization / Tracking	[133–135]
	Human pose / mesh estimation	[136] (ours), [118,137–143]
	Vital sensing	[144–147]
	Papers with public datasets	[148] (ours), [131,140,141]

concerns about scalability and reproducibility outside controlled research settings.

More recently, researchers have also considered alternative feedback signals beyond full CSI. For example, beam signal-to-noise ratio (SNR), obtained during beam training in IEEE 802.11ad/ay 60 GHz systems, provides directional but relatively low-dimensional feedback compared to CSI. While it lacks the rich multipath structure of CSI, its directional nature has proven useful for tasks such as indoor localization and gesture recognition. In contrast, Compressed Beamforming Report (CBR), standardized in IEEE 802.11ac/ax, provides quantized channel feedback. Because CBR is standard-compliant information embedded in regular wireless communication frames, it can be readily obtained without any hardware modification. Our early studies [82,94] were among the first to experimentally explore the feasibility of CBRs for sensing, and subsequent works have extended their use to tasks including localization, vital sensing, and indoor scene recovery. This reflects a broader challenge: transitioning from experimental formats to standardized signals that support scalable and reproducible research and system-level validation in everyday environments.

1.2.2 mmWave Radar Sensing

Research on mmWave radar sensing for indoor perception has advanced rapidly in the past decade, driven by the availability of compact FMCW devices and the increasing interest in privacy-preserving alternatives to cameras. Unlike Wi-Fi signals, which are typically repurposed from commodity communication systems, radar provides dedicated wideband measurements, offering centimeter-level range precision and fine-grained velocity estimates. As a result, mmWave radar has been applied to a wide spectrum of tasks summarized in Table 1.2, and research can be broadly categorized by the representation of radar measurements: point cloud (PC) representations and heatmap representations.

PCs are obtained by applying range-Doppler-angle estimation pipelines, where beat frequencies from FMCW chirps yield range, Doppler processing extracts velocity, and MIMO arrays estimate azimuth and elevation. A Constant False Alarm Rate (CFAR) detector is then applied to suppress noise and identify significant reflections, with the surviving peaks reconstructed as sparse 3D points. Representative works include RadHAR [105] and mm-Pose [113], which explored activity classification and joint tracking from sparse PCs. More recent approaches such as mmMesh [112] and Fan et al. [108] further aligned PCs to full-body human pose and meshes with sophisticated deep learning frameworks.

Heatmaps, in contrast, are generated by retaining the range-angle and range-Doppler spectra directly as dense 2D/3D tensors, preserving multipath and Doppler signatures rather than collapsing them into discrete reflections. This representation supports learning-based approaches that treat radar tensors as images or volumes. An early demonstration, RF-Pose [143], achieved through-wall 2D pose estimation using a 3D convolutional neural network (CNN) and orthogonal two-viewed radar heatmaps. Subsequent work extended this paradigm to progress from 2D keypoint estimation to full 3D skeleton recovery and mesh-level reconstruction [139,141,143].

Public datasets have also facilitated progress in both PC- and heatmap-based directions, including RadHAR [105] for PC-based activity recognition and HIBER [131] for heatmap-based 3D human perception. These datasets, while valuable, remain limited in scale and diversity, reflecting a broader challenge also observed in Wi-Fi sensing datasets.

1.3 Challenges to Practical Wireless Sensing

Although wireless sensing has shown substantial potential, several challenges remain before it can become reproducible and robust at scale in everyday indoor environments. This dissertation identifies two challenges: **open accessibility of sensing data** and **deployment-aware methodological design**, which collectively hinder reproducibility, scalability, and generalization – the prerequisites for practical deployment. These two challenges are deeply interdependent. Without accessible and representative sensing data, methodological innovation cannot be validated under realistic conditions. Conversely, without deployment-aware modeling frameworks that explicitly acknowledge the inherent limitations and biases imposed by practical data collection, the data cannot be fully exploited or generalized across environments. Addressing both in tandem is therefore essential for establishing wireless sensing as a reproducible and scalable sensing paradigm.

1.3.1 Open Accessibility of Sensing Data

Open accessibility of sensing data is not merely a matter of convenience or scale, but a prerequisite for reproducibility. A central challenge in wireless sensing is that signal acquisition remains tightly coupled with specific hardware, firmware, and environmental calibration, making it difficult to

reproduce experiments under comparable conditions. This structural dependence prevents cumulative progress — a fundamental contrast to fields such as vision or speech, where standardized data representations and benchmarks have enabled reproducible development.

In Wi-Fi sensing, this limitation manifests as a reliance on full CSI, which requires customized firmware and privileged hardware access. While early tools such as the Intel 5300 CSI Tool [98] and the Atheros CSI Tool [149] made CSI accessible for research, their dependence on specific chipsets restricted portability. More recent efforts, including ESP32-based CSI Toolkits [150], Nexmon [151], and the PicoScenes platform [152], have expanded hardware coverage to newer Wi-Fi chipsets. Nevertheless, the overall ecosystem exposing CSI remains limited compared with the vast diversity of commodity Wi-Fi devices in real-world environments. As a result, most CSI-based studies still rely on a handful of research-oriented chipsets or software-defined radio (SDR) setups, constraining reproducibility and preventing large-scale evaluation on commodity devices. Radar sensing faces a related limitation: although modern FMCW radars have become more compact and affordable, their operation still requires dedicated hardware and specialized signal-processing pipelines. Unlike Wi-Fi, where sensing can often leverage existing communication infrastructure, radar sensing cannot be realized without purpose-built transmitters and receivers configured for controlled chirp generation and synchronization. Commercial modules from vendors such as Texas Instruments and NXP have lowered the entry barrier, but each module family differs in antenna geometry, firmware behavior, and signal pre-processing, leading to variations in the generated range–Doppler or range–angle responses. Consequently, even when researchers employ nominally similar radar models, the resulting data distributions and calibration parameters often differ, complicating cross-lab reproducibility. In addition, radar experiments require precise spatial calibration, synchronized captures, and often dedicated power and networking interfaces, thereby limiting the mobility and scalability of data collection. This requirement not only restricts large-scale deployment in diverse environments, but also hinders the accumulation of comparable datasets, ultimately constraining reproducibility and generalization in radar-based sensing research.

When the data collection process itself poses a barrier, one natural alternative is to rely on publicly available datasets instead of building new sensing setups. However, current datasets in both Wi-Fi and radar sensing still fall short of representing the complexity of everyday indoor environments. Many Wi-Fi CSI datasets, including Widar3.0 [21], UT-HAR [56], and SignFi [71], are collected in a single room or a small number of layouts, often under line-of-sight conditions with a limited pool of participants, which restricts their ability to test generalization beyond controlled settings. Recent efforts toward larger benchmarks like CSI-Bench [64] and WiMANS [65] broaden scale and task coverage, but continue to depend on full CSI and dedicated acquisition toolchains rather than commodity devices, limiting portability. Radar datasets show analogous constraints: although resources such as HIBER [131] and RT-Pose [140] include multi-view FMCW captures

and provide rich annotations (e.g., 2D/3D pose, bounding boxes), they are typically recorded in uncluttered spaces, with few subjects and limited occlusion, and require expensive multi-sensor calibration setups. Because dense ground truth like 3D keypoints or human meshes often relies on motion capture rigs or multi-camera systems, it remains costly to scale radar datasets to diverse rooms, activities, and multi-person scenes. As a result, publicly available datasets — while invaluable for benchmarking — still do not capture the spatial, human, and environmental diversity needed to evaluate whether methods can generalize to everyday indoor settings, and therefore do not fully remove the accessibility barrier: they constrain not only what can be evaluated, but who is able to participate in advancing the field.

1.3.2 Deployment-aware Methodological Design

While open and representative access to wireless signals is essential, a second structural challenge lies in the lack of methodological frameworks that can effectively operate under the practical constraints of real-world wireless sensing. In realistic deployments, both the wireless measurements themselves and the labels available for learning are inherently constrained by hardware capabilities, communication protocols, deployment cost, and annotation effort. Nevertheless, many existing methods implicitly rely on idealized signal availability and dense supervision, leading to a persistent mismatch between methodological assumptions and deployable sensing scenarios.

Traditional signal-processing techniques, such as angle-of-arrival (AoA) or time-of-flight (ToF) estimation, multipath analysis, and Doppler-based filtering, can infer certain spatial states from wireless signals, but rely on idealized assumptions regarding propagation models, antenna configurations, and calibration. In practice, such assumptions rarely hold in multipath-rich indoor environments, cluttered spaces, or deployments based on commodity devices with limited controllability [153, 154]. While these methods provide interpretable results under controlled conditions, their sensitivity to deviations from these assumptions limits robustness and scalability in realistic settings.

Given these limitations, the field has increasingly shifted toward learning-based approaches that extract representations directly from wireless measurements. Early deep learning methods, often based on convolutional and recurrent neural networks (CNNs/RNNs), demonstrated that promising performance could be achieved even from RF observation sequences and established widely used baselines [15, 19, 36, 40–42, 47, 71, 112, 113, 131, 143]. More recent advances have further increased modeling capacity through expressive architectures such as Transformer-style attention mechanisms and graph neural networks (GCNs) [16, 32, 99, 111, 130, 141]. However, despite these architectural advances, most learning-based methods remain adapted from the computer vision or point-cloud domains [155, 156] and implicitly assume sensing and supervision conditions that are difficult to satisfy in realistic deployments. In practice, RF measurements are often only partially

observable, multi-sensor data are not guaranteed to be temporally aligned, and acquiring dense and precise supervisory signals for model training frequently incurs substantial cost or requires specialized equipment. As a result, increasing model expressiveness alone does not resolve the gap between methodological assumptions and deployable sensing conditions, and performance gains achieved under controlled settings often fail to translate into robustness and scalability across diverse environments, subjects, and devices. Beyond these general challenges, Wi-Fi sensing introduces an additional, modality-specific constraint regarding the accessibility of fine-grained CSI. As discussed in the preceding section, detailed CSI is frequently assumed in learning-based Wi-Fi sensing studies while it is often inaccessible on commodity devices due to hardware, firmware, and protocol limitations. This further restricts the form and fidelity of wireless observations available in practice, motivating the exploration of sensing methods that do not rely on idealized CSI access, but instead operate on realistically obtainable feedback.

In summary, these tendencies point to a central methodological gap: wireless sensing still lacks a principled, deployment-aware framework that aligns modeling assumptions with the realities of obtainable measurements, supervision, and sensing conditions. Bridging this gap is essential for developing methods that remain robust, interpretable, and scalable when deployed using the limited and imperfect data available in real-world wireless environments.

1.4 Positioning, Scope, and Contributions

This dissertation addresses the challenges identified in the preceding review, namely **(i) the limited accessibility of reproducible wireless signals** and **(ii) the lack of deployment-aware learning frameworks**, with the explicit aim of bridging physical signal behavior and high-level perception under realistic sensing conditions. It positions wireless sensing as a field that must advance both its data foundation and methodological frameworks in tandem, in ways that enable progress to be reproducible and scalable beyond controlled experimental settings. Accordingly, this dissertation undertakes an integrated effort that combines open and standardized data acquisition with deployment-aware modeling, and validates these efforts through system-level experiments on commodity devices deployed in realistic indoor environments. The central scope is sensing in indoor environments, including human activity recognition, device localization, and full-body pose estimation, using Wi-Fi and mmWave radar as representative modalities.

The specific contributions of this dissertation are as follows:

(i) Establishing open and representative access to wireless signals (Chapter 2)

- **Wi-Fi signal accessibility:** Traditional Wi-Fi sensing has heavily relied on full CSI, which requires specialized hardware and modified drivers. This dissertation demonstrates that

CBR, a standard feedback format readily available on commodity IEEE 802.11ac/ax devices, can serve as a practical alternative for sensing. To promote research adoption, we developed **WiPiCap**, an open-source toolkit that enables easy acquisition and decoding of CBR without custom hardware. This contribution expands the accessibility of Wi-Fi sensing for both academic and applied research, and has already been adopted in follow-up studies.

- **Radar data accessibility:** While radar offers fine-grained measurements, large-scale research has been limited by the lack of publicly available annotated data for practical indoor scenarios. This dissertation contributes the **MMVR dataset**, which provides synchronized high-resolution multi-view mmWave radar data together with ground-truth annotations for human-centric perception tasks. By providing the first dataset of this scale and detail, MMVR establishes a foundational resource for radar-based indoor perception, lowering the entry barrier, and enabling reproducible evaluation across methods.

(ii) Deployment-aware modeling under realistic data constraints (Chapter 3, 4, 5)

- **Establishing CBR as an alternative sensing signal:** We propose **CBR-ACE**, a framework that establishes standard-compliant CBRs as a viable sensing signal for human activity recognition, enabling Wi-Fi sensing beyond the practical limitations of fine-grained CSI access. By characterizing CBR's effective dimensionality and feature structure, we design a lightweight pipeline that estimates activity repetition counts with small errors on commodity devices. This contribution demonstrates that standardized Wi-Fi feedback can underpin wireless perception, laying a methodological foundation beyond the full CSI and validating its reproducible feasibility through a fully operational system built on commodity devices.

- **Exploring spatial representation from CBR:** We further explore the availability of CBR for sensing with **CSI2Image**, a framework that reconstructs indoor images from standard-compliant CBRs using a generative adversarial network (GAN) architecture. Rather than targeting specific recognition tasks, this work examines how consistently spatial information implicitly embedded in CBR can be projected into a spatially meaningful visual form. This contribution shows that CBR supports not only task-specific recognition, but also interpretable generative representations that broaden the scope of wireless sensing.

- **Radar-based pose estimation with low-cost supervision:** We develop **RAPTR**, a transformer-based framework that formulates radar perception as a set prediction problem. Unlike prior approaches, RAPTR achieves 3D human pose estimation under weak supervision by leveraging only 2D keypoints and 3D bounding boxes, guided by a two-stage decoding with hierarchical losses. This contribution shows that realistic radar-based pose estimation is

achievable without dense annotations, aligning with the dissertation's broader aim of practical deployment.

- **Neural fusion across asynchronous Wi-Fi bands:** We propose a **multi-band neural fusion** framework that jointly exploits sub-6GHz CSI and 60GHz beam SNR, two feedback formats already available in commodity IEEE 802.11 systems. To address the inherent asynchrony and frequency-dependent propagation effects, the method encodes each stream in a continuous latent state using neural ordinary differential equations (ODEs). This contribution demonstrates that heterogeneous Wi-Fi measurements can be fused in a sequence-to-sequence manner, considering wireless dynamics, advancing the methodological foundations of practical indoor localization.

Taken together, these contributions directly address the central challenges identified in this dissertation: they lower the barrier to wireless sensing through reproducible data accessibility, expand dataset diversity and realism of experimental evaluations, and strengthen methodological foundations by establishing frameworks explicitly designed to operate under constrained wireless observations and supervision. In the concluding discussion (Chapter 6), we not only reflect on these advances, but also compare the respective strengths and limitations of Wi-Fi and radar sensing, outlining how each modality can evolve toward complementary and scalable roles in future indoor perception systems.

1.5 Thesis Organization

The remainder of this dissertation is organized as follows. Chapter 2 presents the data acquisition platforms, introducing WiPiCap for Wi-Fi CBR collection and MMVR as a large-scale radar dataset. Chapter 3 investigates Wi-Fi sensing with standardized signals through CBR-ACE and CSI2Image. Chapter 4 turns to radar-based perception and proposes RAPTR for 3D human pose estimation. Chapter 5 explores Wi-Fi sensing with richer multi-band information using a neural dynamic fusion framework. Chapter 6 synthesizes insights from Wi-Fi and radar sensing, highlighting their complementary roles, shared limitations, and future directions toward interoperable and scalable wireless perception. Finally, Chapter 7 concludes the dissertation with a summary and outlook.

Chapter 2

Data Foundations for Practical Wireless Sensing

2.1 WiPiCap: A Tool for Wi-Fi CBR Collection

The effectiveness of Wi-Fi sensing has long been demonstrated using CSI collected from commodity wireless NICs. However, the availability of CSI is limited to only a few chipsets and requires non-standard driver modifications. This dependency has severely hampered reproducibility, limited portability, and delayed the path toward practical deployment. To address this issue, we develop WiPiCap, a system that leverages the CBR defined in IEEE 802.11ac/ax as a standard-compliant alternative to CSI. WiPiCap enables Wi-Fi sensing research without relying on vendor-specific extensions, providing a reproducible and hardware-independent data acquisition foundation.

2.1.1 Overview of Compressed Beamforming Report

Modern Wi-Fi standards such as IEEE 802.11n and later revisions rely on multi-antenna transmission as a core mechanism for coping with spatial interference and limited spectrum resources. Rather than increasing raw bandwidth, MIMO exploits spatial structure to trade hardware complexity for robustness and throughput. Indoor wireless propagation is dominated by indirect signal paths created through reflections and diffraction from surrounding structures. As a consequence, the received waveform entangles contributions from multiple propagation mechanisms, making it difficult to associate individual signal components with their originating transmit antennas. The primary task of the receiver is to reconstruct the original signals transmitted from each antenna by processing the received waveform.

IEEE 802.11n addressed this problem by employing full CSI to estimate the multipath environment. Access to full CSI enables classical linear equalization strategies [157] to mitigate inter-stream interference at the receiver.

To improve upon this, recent approaches employ precoding of transmission signals. A typical solution is Eigenspace Spatial Division Multiplexing (E-SDM). From a signal processing perspec-

tive, eigenmode-based transmission can be formulated by decomposing the channel into orthogonal spatial components. While the detailed derivation is well established in the MIMO literature, we briefly outline the formulation here only to clarify the origin of the feedback information used later in this chapter. We consider a point-to-point wireless link consisting of a single transmitter and a single receiver. Let M and N denote the numbers of antennas at the transmitter and the receiver, respectively. Under this setting, the received signal can be expressed as follows:

$$\mathbf{y} = \mathbf{H}\mathbf{x} + \mathbf{n}, \tag{2.1}$$

where $\mathbf{x} \in \mathbb{C}^{1\times M}$ and $\mathbf{y} \in \mathbb{C}^{1\times N}$ correspond to the transmit and receive signal vectors, respectively. The term $\mathbf{n} \in \mathbb{C}^{1\times N}$ represents the Additive White Gaussian Noise (AWGN), and $\mathbf{H} \in \mathbb{C}^{N\times M}$ denotes the complete CSI. The full CSI can be decomposed by SVD as

$$\mathbf{H} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^H, \tag{2.2}$$

where $\mathbf{U}$ and $\mathbf{V}$ denote the left and right singular matrices, respectively, both of which are complex-valued unitary matrices, and $\mathbf{\Sigma}$ is a diagonal matrix whose i-th diagonal entry corresponds to a singular value. Each singular value reflects the amplitude gain of an individual substream. Here, $\mathbf{V}^H$ refers to the Hermitian transpose of $\mathbf{V}$. By inserting Equation (2.2) into Equation (2.1), we obtain

$$\mathbf{y} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^H\mathbf{x} + \mathbf{n}. \tag{2.3}$$

In E-SDM, the transmit signal $\mathbf{x}$ is first precoded using the right singular matrix $\mathbf{V}$, resulting in a spatially orthogonal representation $\hat{\mathbf{x}}$ with suppressed inter-stream interference. The received signal can then be written as

$$\begin{aligned} \mathbf{y} &= \mathbf{H}\hat{\mathbf{x}} + \mathbf{n} \\ &= \mathbf{U}\mathbf{\Sigma}\mathbf{V}^H\mathbf{V}\mathbf{x} + \mathbf{n} \\ &= \mathbf{U}\mathbf{\Sigma}\mathbf{x} + \mathbf{n}. \end{aligned} \tag{2.4}$$

Multiplying the received signal by $\mathbf{U}^H$ yields the reconstructed form $\mathbf{\Sigma}\mathbf{x}+\mathbf{n}$, which shows that eigenbeam-based transmission enables orthogonal substreams whose gains are determined by the singular values. Such orthogonal transmission requires both the transmitter and the receiver to have access to the right singular matrix $\mathbf{V}$ for precoding. As a result, $(\min(M,N))^2$ complex-valued parameters must be conveyed as side information, which can introduce substantial communication overhead and lead to power and throughput degradation.

To reduce this overhead, recent IEEE 802.11 standards adopt a compression scheme based on Givens rotation and quantization [158,159]. Assuming $M \geq N$, let $\hat{\mathbf{V}}$ denote the submatrix formed by the first m rows and n columns of $\mathbf{V}^H$. The matrix $\hat{\mathbf{V}}$ can be factorized as follows:

$$\hat{\mathbf{V}} = \left[\prod_{i=1}^{\min(N,M-1)} \left[D_i \prod_{l=i+1}^{M} G_{li}^T(\Psi_{li})\right]\right] \tilde{\mathbf{I}}_{M\times N}, \tag{2.5}$$

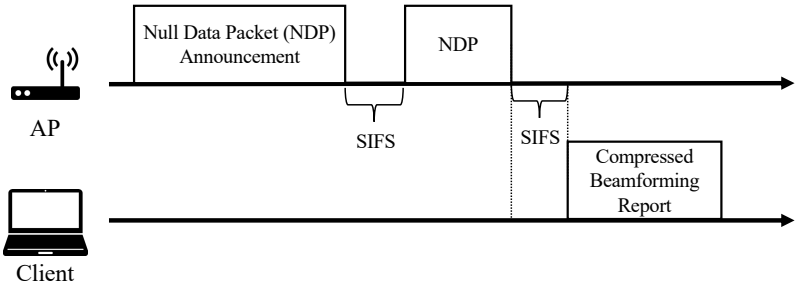


Figure 2.1: IEEE 802.11ac/ax channel sounding procedure for CBR acquisition.

where D_i and G_{li} are derived as follows:

$$D_i = \begin{bmatrix} I_{i-1} & 0 & \dots & \dots & 0 \\ 0 & e^{j\phi_{i,i}} & 0 & \dots & 0 \\ \vdots & 0 & \ddots & 0 & 0 \\ \vdots & \vdots & 0 & e^{j\phi_{M-1,i}} & 0 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix}, \tag{2.6}$$

$$G_{li}(\psi) = \begin{bmatrix} I_{i-1} & 0 & 0 & 0 & 0 \\ 0 & \cos(\psi) & 0 & \sin(\psi) & 0 \\ 0 & 0 & I_{l-i-1} & 0 & 0 \\ 0 & -\sin(\psi) & 0 & \cos(\psi) & 0 \\ 0 & 0 & 0 & 0 & I_{M-l} \end{bmatrix}. \tag{2.7}$$

Here, G_{li} denotes a Givens rotation matrix applied to rows l and i. Each I_k represents a $k \times k$ identity matrix, while $\tilde{\mathbf{I}}_{M\times N}$ corresponds to a zero-padded identity matrix when $M \neq N$. With this formulation, an arbitrary basis matrix can be parameterized by a set of angular variables $\phi_{k,l}$ and $\psi_{k,l}$, which are subsequently quantized into finite-bit values.

By sequentially applying such rotations, the right singular matrix $\mathbf{V}$ is reduced to a parameter set of quantized angles. These angles are serialized across subcarriers and transmitted as the CBR. CBR packets are not broadcast; they are returned by each client device in response to explicit requests from the access point. Moreover, the quantization inherently reduces the amount of available channel information. In practice, CBR packets are generated as part of the beamforming training procedure specified in IEEE 802.11 (Figure 2.1). The access point periodically initiates this process by transmitting a Null Data Packet (NDP) Announcement frame to the client device. The announcement is followed by the transmission of an NDP, which contains pilot symbols used for channel estimation at the client side. Upon receiving the NDP, the client estimates the downlink channel, applies the decomposition defined in the standard, and responds with a beamforming report that encodes the quantized angle information as CBR. This reply occurs after a Short Interframe Space (SIFS), the mandatory minimal waiting interval defined in the standard, which further constrains the achievable temporal resolution. Since this exchange is triggered by the AP, each CBR packet is tightly coupled with a preceding NDP announcement and NDP frame, forming a protocol-level handshake between the AP and the client. Although continuous traffic stimulation

can be used to encourage regular NDP announcements, the achievable sampling rate remains limited to about 10–50Hz in commodity devices. This is significantly lower than CSI acquisition, where low-level access to NIC internals allows much higher sampling frequencies. In contrast, CBR collection passively relies on standard-compliant control exchanges, which inherently constrain the rate at which channel information becomes available.

2.1.2 Example of Compression Process

To illustrate the parameterization process, we consider a 4×2 matrix $\hat{\mathbf{V}}$ corresponding to a MIMO configuration with 4 receiving antennas and 2 transmitting antennas:

$$\hat{\mathbf{V}} = \begin{bmatrix} v_{11}^{(0)} & v_{12}^{(0)} \\ v_{21}^{(0)} & v_{22}^{(0)} \\ v_{31}^{(0)} & v_{32}^{(0)} \\ v_{41}^{(0)} & v_{42}^{(0)} \end{bmatrix}, \tag{2.8}$$

where $v_{ij}^{(t)} \in \mathbb{C}$ denotes the (i,j)-th entry of the right singular matrix after the t-th Givens rotation.

The procedure begins by enforcing real-valued magnitudes on the elements of the bottom row. This is achieved by right-multiplying

$$\hat{\mathbf{V}} = \begin{bmatrix} v_{11}^{(0)} & v_{12}^{(0)} \\ v_{21}^{(0)} & v_{22}^{(0)} \\ v_{31}^{(0)} & v_{32}^{(0)} \\ v_{41}^{(0)} & v_{42}^{(0)} \end{bmatrix} \begin{bmatrix} \frac{v_{41}^{(0)}}{|v_{41}^{(0)}|} & 0 \\ 0 & \frac{v_{42}^{(0)}}{|v_{42}^{(0)}|} \end{bmatrix}^H = \begin{bmatrix} v_{11}^{(1)} & v_{12}^{(1)} \\ v_{21}^{(1)} & v_{22}^{(1)} \\ v_{31}^{(1)} & v_{32}^{(1)} \\ |v_{41}^{(0)}| & |v_{42}^{(0)}| \end{bmatrix}, \tag{2.9}$$

Subsequently, a rotation matrix D_1 is applied to transform the entries of the first column into real-valued components:

$$D_1 = \begin{bmatrix} \exp(j\phi_{11}) & 0 & 0 & 0 \\ 0 & \exp(j\phi_{21}) & 0 & 0 \\ 0 & 0 & \exp(j\phi_{31}) & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}, \tag{2.10}$$

$$\exp(j\phi_{11}) = \frac{v_{11}^{(1)}}{|v_{11}^{(1)}|},\ \exp(j\phi_{21}) = \frac{v_{21}^{(1)}}{|v_{21}^{(1)}|},\ \exp(j\phi_{31}) = \frac{v_{31}^{(1)}}{|v_{31}^{(1)}|}, \tag{2.11}$$

$$D_1^H \begin{bmatrix} v_{11}^{(1)} & v_{12}^{(1)} \\ v_{21}^{(1)} & v_{22}^{(1)} \\ v_{31}^{(1)} & v_{32}^{(1)} \\ |v_{41}^{(0)}| & |v_{42}^{(0)}| \end{bmatrix} = \begin{bmatrix} v_{11}^{(2)} & v_{12}^{(2)} \\ v_{21}^{(2)} & v_{22}^{(2)} \\ v_{31}^{(2)} & v_{32}^{(2)} \\ v_{41}^{(2)} & v_{42}^{(2)} \end{bmatrix} = \begin{bmatrix} |v_{11}^{(1)}| & v_{12}^{(2)} \\ |v_{21}^{(1)}| & v_{22}^{(2)} \\ |v_{31}^{(1)}| & v_{32}^{(2)} \\ |v_{41}^{(0)}| & |v_{42}^{(0)}| \end{bmatrix}. \tag{2.12}$$

For the first and second rows of the matrix, we define the Givens rotation matrix G_{21} as follows:

$$G_{21} = \begin{bmatrix} \cos\psi_{21} & \sin\psi_{21} & 0 & 0 \\ -\sin\psi_{21} & \cos\psi_{21} & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}, \tag{2.13}$$

$$\cos\psi_{21} = \frac{|v_{11}^{(1)}|}{\sqrt{|v_{11}^{(1)}|^2 + |v_{21}^{(1)}|^2}}, \quad \sin\psi_{21} = \frac{|v_{21}^{(1)}|}{\sqrt{|v_{11}^{(1)}|^2 + |v_{21}^{(1)}|^2}}, \tag{2.14}$$

$$G_{21}D_1^H \begin{bmatrix} v_{11}^{(1)} & v_{12}^{(1)} \\ v_{21}^{(1)} & v_{22}^{(1)} \\ v_{31}^{(1)} & v_{32}^{(1)} \\ \left|v_{41}^{(0)}\right| & \left|v_{42}^{(0)}\right| \end{bmatrix} = \begin{bmatrix} \sqrt{|v_{11}^{(1)}|^2 + |v_{21}^{(1)}|^2} & v_{12}^{(3)} \\ 0 & v_{22}^{(3)} \\ |v_{31}^{(1)}| & v_{32}^{(2)} \\ |v_{41}^{(0)}| & |v_{42}^{(0)}| \end{bmatrix}. \tag{2.15}$$

Similarly, we define G_{31} and G_{41} as follows:

$$G_{31} = \begin{bmatrix} \cos\psi_{31} & 0 & \sin\psi_{31} & 0 \\ 0 & 1 & 0 & 0 \\ -\sin\psi_{31} & 0 & \cos\psi_{31} & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}, \tag{2.16}$$

$$\cos\psi_{31} = \frac{\sqrt{|v_{11}^{(1)}|^2 + |v_{21}^{(1)}|^2}}{\sqrt{|v_{11}^{(1)}|^2 + |v_{21}^{(1)}|^2 + |v_{31}^{(1)}|^2}},\ \sin\psi_{31} = \frac{|v_{31}^{(1)}|}{\sqrt{|v_{11}^{(1)}|^2 + |v_{21}^{(1)}|^2 + |v_{31}^{(1)}|^2}}, \tag{2.17}$$

$$G_{41} = \begin{bmatrix} \cos\psi_{41} & 0 & 0 & \sin\psi_{41} \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ -\sin\psi_{41} & 0 & 0 & \cos\psi_{41} \end{bmatrix}, \tag{2.18}$$

$$\cos\psi_{41} = \sqrt{\left|v_{11}^{(1)}\right|^2 + \left|v_{21}^{(1)}\right|^2 + \left|v_{31}^{(1)}\right|^2},\ \sin\psi_{41} = |v_{41}^{(0)}|. \tag{2.19}$$

We further obtain a unit vector at the first column by multiplying G_{31} and G_{41} as follows:

$$G_{41}G_{31}G_{21}D_1^H \begin{bmatrix} v_{11}^{(1)} & v_{12}^{(1)} \\ v_{21}^{(1)} & v_{22}^{(1)} \\ v_{31}^{(1)} & v_{32}^{(1)} \\ \left|v_{41}^{(0)}\right| & \left|v_{42}^{(0)}\right| \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & v_{22}^{(3)} \\ 0 & v_{32}^{(4)} \\ 0 & v_{42}^{(5)} \end{bmatrix}. \tag{2.20}$$

We perform similar procedures for the remaining column, sequentially calculating D_2, G_{32}, G_{42}, and finally obtain an identity matrix $\tilde{\mathbf{I}}_{M\times N}$.

$$G_{42}G_{32}D_2^H G_{41}G_{31}G_{21}D_1^H \begin{bmatrix} v_{11}^{(1)} & v_{12}^{(1)} \\ v_{21}^{(1)} & v_{22}^{(1)} \\ v_{31}^{(1)} & v_{32}^{(1)} \\ \left|v_{41}^{(0)}\right| & \left|v_{42}^{(0)}\right| \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 0 & 0 \\ 0 & 0 \end{bmatrix}. \tag{2.21}$$

Finally, by applying the above procedure, the original matrix $\hat{\mathbf{V}}$ can be parameterized using a finite set of rotation angles. In this illustrative case of a 4×2 system, the decomposition yields specific angular parameters $\phi_{11}, \phi_{21}, \phi_{31}, \psi_{21}, \psi_{31}, \psi_{41}, \phi_{12}, \phi_{22}, \psi_{32}, \psi_{42}$, which fully characterize the matrix. This illustrates how a compact set of angular parameters suffices to reconstruct the matrix, forming the basis of CBR representation.

2.1.3 Design and Implementation of WiPiCap

WiPiCap is a system that operates directly on wireless packet binaries to make CBR by reversing the compression process described above. It captures CBR packets exchanged between access points

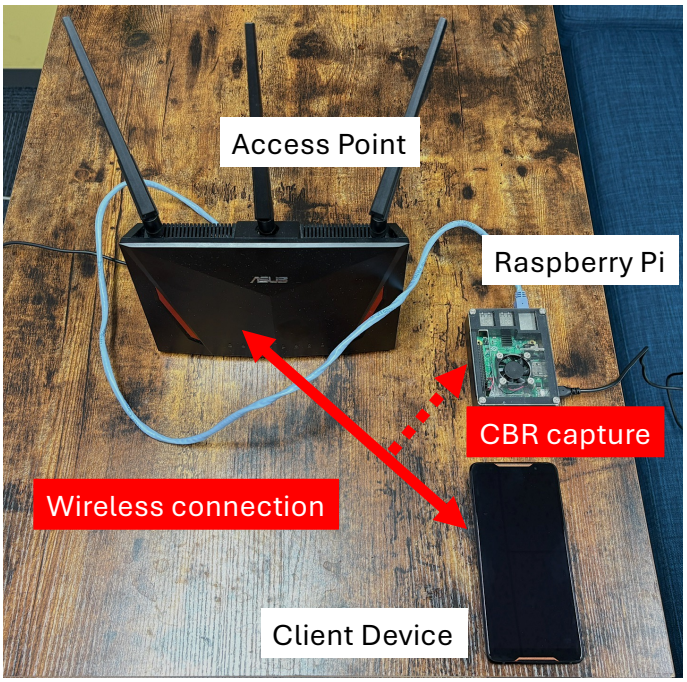


Figure 2.2: System setup of WiPiCap.

and client devices, decodes quantized angle information, and reconstructs the right singular matrix $\mathbf{V}$. WiPiCap is designed to lower the barrier to CBR-based sensing by providing the research community with an open and practical toolset. It consists of an implementation that enables both the collection and decoding of CBR data on commercial off-the-shelf devices, together with a conversion that makes such data directly usable for sensing applications. In this sense, WiPiCap serves not only as a system but also as a community resource for reproducible Wi-Fi sensing using standard-compliant information.

Figure 2.2 illustrates the experimental setup with which WiPiCap collects and decodes CBRs. This setup involves three components: an wireless AP, a client device, and a Raspberry Pi used for packet capture. The AP is connected to the Raspberry Pi via a wired link, while the AP and the client communicate wirelessly. To ensure a steady sampling rate of CBRs, the Raspberry Pi continuously stimulates the client by sending periodic Internet Control Message Protocol (ICMP) pings or streaming User Datagram Protocol (UDP) traffic using iperf, thereby triggering regular NDP announcements and subsequent beamforming reports. During this process, the Raspberry Pi simultaneously performs packet capture and stores the CBR packets for decoding.

WiPiCap supports both Very High Throughput (VHT) beamforming reports defined in IEEE 802.11ac and High Efficiency (HE) beamforming reports defined in IEEE 802.11ax. Once collected, these reports are decoded to reconstruct the quantized angular parameters and, subsequently, the right singular matrix $\hat{\mathbf{V}}$. To correctly handle arbitrary antenna and subcarrier configurations,

Table 2.1: Effective sampling rates across different AP–client pairs.

AP–Client Pair	Antennas (Tx/Rx)	Standard	Rate (Hz)
AX6000			
– iPhone	4/2	11ax	19.40
– MacBook	4/2	11ac	17.07
RT-AC86U			
– iPhone	4/2	11ac	37.90
– MacBook	4/2	11ac	38.46

WiPiCap parses the CBR payload in a comprehensive manner. It should be noted that the reconstruction does not yield the original right singular matrix $\mathbf{V}$, since the standard first applies an irreversible phase adjustment to enforce real-valued elements in the bottom row (Equation (2.9)); thus, WiPiCap reconstructs the post-processed version defined by the specification. To accommodate different experimental needs, WiPiCap provides two independent implementations. The Python implementation is particularly useful when integrating CBR decoding into machine learning pipelines or when analyzing previously collected packet traces as a research dataset. In contrast, the C++ implementation is optimized for performance, capable of running directly on Raspberry Pi hardware without reducing the acquisition rate of CBRs, which also enables real-time visualization and monitoring.

By integrating capture and decoding in a unified, standards-compliant system, WiPiCap enables the research community to collect CBR data and obtain CSI-like representations in a reproducible and hardware-independent manner. This design lowers the barrier to entry for Wi-Fi sensing research and provides a foundation for experiments that were previously infeasible with CSI-only approaches. To further promote reproducibility and accessibility, we release the full WiPiCap implementation as open source at `https://github.com/watalab/WiPiCap.git`, making it readily available for the community to adopt, extend, and deploy.

2.1.4 Sampling Rate Benchmark

To validate the practicality of WiPiCap, we measured the achievable sampling rate of CBR collection across different combinations of commodity APs and client devices. Table 2.1 summarizes the results. These measurements were taken while WiPiCap simultaneously captured packets and executed the C++ decoding pipeline in real time. The decoding procedure was sufficiently lightweight that no packet drops occurred, and the effective acquisition rate was unchanged compared to capture-only runs. Although the achievable sampling rates are lower than those typically available with CSI, they are sufficient for sensing tasks involving relatively large-scale human activities such as HAR and localization. The observed rates varied substantially across AP models, suggesting that the attainable sampling frequency is primarily bounded by the AP's beamforming

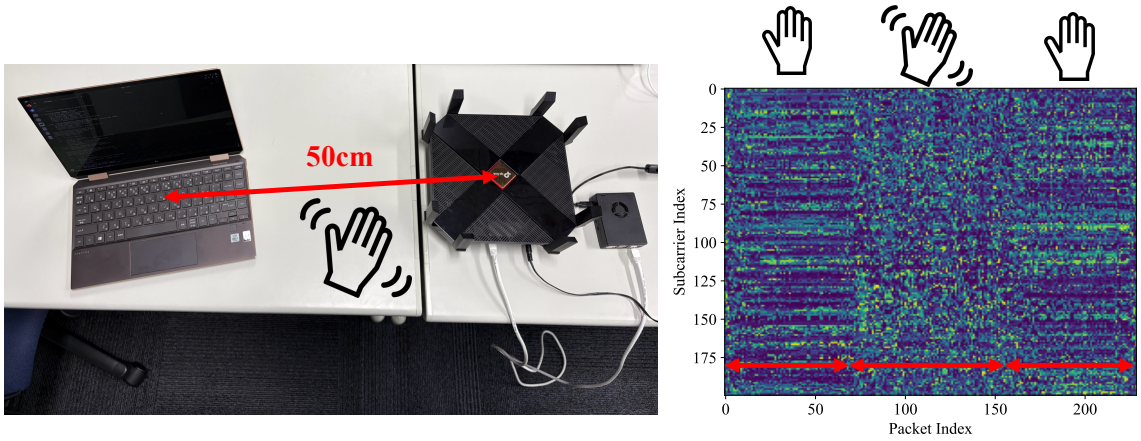


Figure 2.3: Temporal response of CBR for human-induced motion. (Left) A subject periodically moves toward and away from the access point. (Right) The magnitude of the reconstructed V matrix exhibit disturbance with the subject motion, demonstrating that CBR captures motion-induced spatial variation despite its quantization and limited sampling rate.

implementation rather than by client devices or the decoding overhead of WiPiCap.

2.1.5 Motion Response Verification

To evaluate whether the V matrix, which WiPiCap decodes, captures motion-induced channel variations, we conducted a controlled experiment in which a subject repeatedly shakes a hand toward and away from the access point at a constant pace. Although the CBR is a quantized and compressed representation derived from beamforming feedback, the reconstructed V matrix should still reflect changes in the underlying multipath structure caused by human motion.

Figure 2.3 illustrates the response of CBR in a simple hand-movement scenario. In this experiment, a laptop client and an IEEE 802.11ac access point were placed approximately 50 cm apart on a desk and communicated over an 80 MHz channel using a 2×2 MIMO link. Among the two spatial streams, we visualize the reconstructed amplitude of a single stream obtained from the CBR. As the subject shakes a hand between the devices, the amplitude of the reconstructed V matrix exhibits clear disturbances, whereas relatively stable patterns appear when the hand remains still. These modulations correspond to geometric changes in the dominant propagation paths, confirming that CBR retains sufficient sensitivity to motion-induced perturbations despite its lower resolution and limited sampling rate.

This verification demonstrates that, even though CBR is designed primarily for communication-oriented precoding, it encodes meaningful spatial information that reacts to human movement. This result substantiates the feasibility of using CBR as a sensing signal and provides an empirical basis for the applications developed in subsequent sections.

2.1.6 Conclusion and Contribution

WiPiCap serves a dual role within the Wi-Fi sensing ecosystem, functioning both as a practical system and as a shared research resource. By providing a standard-compliant, open-source implementation, it enables researchers to access CSI-like information without chipset-specific modifications, ensuring reproducibility and portability. The real-time decoding experiments confirm that the system introduces negligible overhead, while the benchmark results illustrate that achievable sampling rates depend largely on AP implementations. In summary, as an openly released toolset, WiPiCap establishes a reproducible foundation for next-generation Wi-Fi sensing research, which has already been adopted in follow-up studies [85], and stands as a community resource that lowers the barrier to entry for future studies.

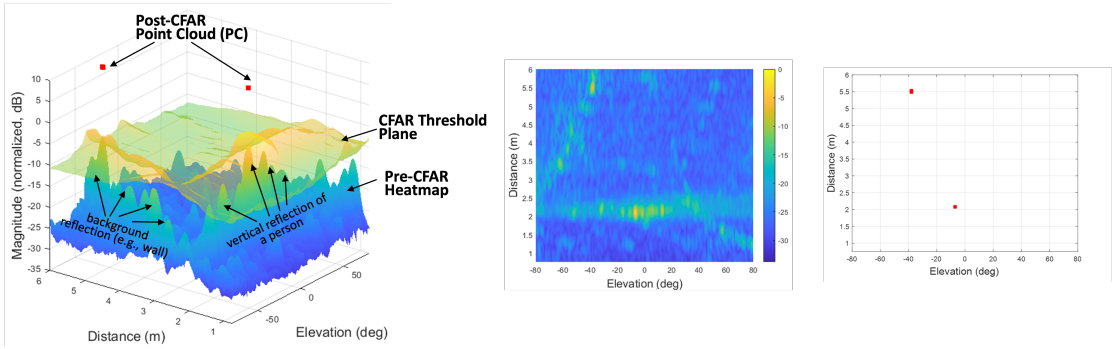
2.2 MMVR: A Large-Scale Multi-View Radar Dataset for Indoor Perception

As discussed in Chapter 1, a key obstacle to advancing practical wireless sensing is the lack of diverse and representative datasets, particularly for radar-based perception. Existing datasets are typically restricted to single-room environments, limited subject pools, or simplified actions, making it difficult to assess generalizability in realistic and cluttered settings [69, 131, 140, 141]. To address this gap, we present MMVR, a large-scale multi-view radar dataset that captures complex indoor environments with multiple subjects, occlusions, and rich scene clutter. MMVR provides synchronized radar views with comprehensive annotations for detection, pose estimation, and segmentation, offering a reproducible and standard benchmark to overcome the dataset diversity bottleneck and foster progress in robust indoor perception.

2.2.1 Introduction

Indoor radar perception has attracted far less attention than automotive radar sensing, despite its relevance to a wide range of applications such as indoor robotic navigation, smart building energy management, and elderly care in visually degraded conditions, e.g., low light, smoke, fire, or dust. Compared to outdoor automotive scenarios [160–174], indoor settings benefit from lower deployment costs and reduced privacy concerns, yet progress has been slower. Consequently, publicly available indoor radar datasets remain scarce and often limited in scale, annotation richness, supported tasks, and standardized benchmarks.

Several early studies have nonetheless investigated the use of radar signals for indoor perception. One of the earliest efforts, RF-Pose [143], employed a T-shaped antenna configuration [175] consisting of 16 horizontal and 4 vertical elements operating in the 5–7GHz band. The method constructs two orthogonal radar heatmaps, fuses them using a convolutional autoencoder, and



(a) **Pre-CFAR heatmap** versus **Post-CFAR PC** (b) Pre-CFAR heatmap (c) Post-CFAR PC

Figure 2.4: Radar signal representations before and after CFAR processing for a sample scene from the MMVR dataset are illustrated. The pre-CFAR heatmap, shown as a 3D view in (a) and a top-down view in (b), captures an extended vertical structure of the subject, manifested as multiple clustered reflections across the elevation (height) dimension. After CFAR processing, the heatmap is compared against a threshold plane (depicted as the semi-transparent surface in (a)), resulting in only a small number of detected points (red squares in (a) and (c)). This process significantly suppresses weaker reflections from the subject, retaining only the strongest responses.

estimates multi-person 2D poses in the image plane. However, the RF-Pose dataset itself is not publicly released. More recently, HuPR [141] adopted a similar sensing setup based on orthogonal low-resolution radar heatmaps, targeting action recognition in empty indoor spaces and pose estimation, and made the resulting dataset publicly available (see Figure 2.5(a)). Owing to their coarse angular resolution, limited to approximately 15° in both azimuth and elevation [141, 143], RF-Pose and HuPR are categorized as *multi-view low-resolution (LR) radar heatmap* datasets in Table 2.2.

Subsequent to RF-Pose, a number of indoor radar studies have explored the use of commercially available millimeter-wave radar platforms rather than custom antenna arrays. In particular, TI's single-chip mmWave sensors, equipped with three transmit and four receive antennas and operating in the 60–64GHz or 76–81GHz bands, have been widely adopted. Owing to the short wavelength,

Table 2.2: Indoor Radar Perception Datasets.

Datasets	Year	Sensor	Views	Data	Tasks	Size	Public
RadHAR [105]	2019	LR (low-resolution)	Single	PC (Point Cloud)	Action	167K	✓
mm-Pose [113]	2020	LR	Multi	PC	Action, Pose	40K	✗
mmMesh [112]	2021	LR	Single	PC	3D Mesh	480K	✗
mRI [122]	2022	LR	Single	PC	Action, Pose	160K	✓
MM-Fi [69]	2023	LR	Single	PC	Action, Pose	320K	✓
RF-Pose [143]	2018	LR	Multi	Heatmap	Pose	-	✗
HuPR [141]	2023	LR	Multi	Heatmap	Pose	141K	✓
HIBER [131]	2023	HR (high-resolution)	Multi	Heatmap	Box, Pose, Seg.	179K	Partial
MMVR(ours)	2024	HR	Multi	Heatmap	Box, Pose, Seg.	345K	✓

LR: low-resolution radar with an angular resolution of 15°.
HR: high-resolution radar with an angular resolution of 1.3°.



(a) HuPR (open space) (b) HIBER (open foreground) (c) MMVR (cluttered space)

Figure 2.5: Example snapshots of indoor radar **heatmap** datasets summarized in Table 2.3. HuPR [141] and HIBER [131] were collected at a single location and, respectively, multiple locations within a single open-space, open-foreground room. In contrast, MMVR was collected across multiple locations spanning multiple environments, including an open-foreground room (d1-d4) and cluttered (d5-d9) rooms.

these systems provide angular resolutions on the order of 15°, which is comparable to that of RF-Pose despite the simpler hardware configuration. While early work relied on multi-view radar heatmaps, many later indoor radar datasets—including RadHAR [105], mm-Pose [113], mRI [122], and MM-Fi [69]—represent observations as low-resolution point clouds obtained from a single viewing direction, typically aligned with the horizontal plane, as summarized in the first block of Table 2.2.

The choice between heatmap-based and point-cloud-based representations has important implications for downstream perception tasks such as image-plane pose estimation and segmentation. Figure 2.4 provides a concrete example using data from the MMVR dataset, where a subject moves at a distance of 2 m from the radar. Prior to CFAR processing [176], the heatmap representations ((a) in 3D and (b) in top-down view) retain a vertically continuous reflection structure across the elevation dimension. When CFAR is applied, the heatmap is evaluated against an adaptively defined threshold surface (shown as a semi-transparent plane in (a)), and only a small number of responses exceeding this threshold are preserved. The resulting point cloud ((a) and (c), red markers) discards many weaker yet informative reflections, leading to the loss of fine-grained spatial cues that are critical for more demanding perception tasks. Consequently, as indicated in Table 2.2, existing single-view low-resolution point cloud datasets have predominantly been used for coarse-level tasks such as action classification [69, 105, 113, 122] rather than detailed spatial understanding.

To capture richer spatial structures from radar signals, recent studies have begun to adopt *multi-view high-resolution (HR) heatmap* representations for indoor radar perception, as exemplified by HIBER [131]. Nevertheless, the HIBER dataset is collected at multiple positions within the *same* room under an open-foreground configuration (see Figure 2.5 (b)), where subjects are not occluded by furniture such as tables or chairs. As a result, the complexity of real indoor environments is

only partially represented.

To facilitate radar-based perception research in more realistic and challenging indoor settings—characterized by cluttered spaces, multiple rooms, and frequent occlusions—and to support more demanding downstream perception tasks, we introduce a large-scale indoor radar dataset named **Millimeter-wave Multi-View Radar (MMVR)**. Its main contributions are summarized as follows:

- MMVR comprises **345K** radar frames collected from **25** participants across **6** distinct indoor environments, including both open and cluttered office and meeting-room scenarios, over **9** different days. To our knowledge, MMVR is the largest publicly available indoor radar dataset that simultaneously spans **multiple days**, **multiple rooms**, and **multiple subjects**.

- The dataset is organized into two complementary subsets. **P1: Open Foreground** contains **107.9K** frames recorded in a single open space with one subject (illustrated in the first row of Figure 2.5(c)), and is intended to establish upper-bound performance under favorable conditions. **P2: Cluttered Space** includes **237.9K** frames captured in **5** cluttered rooms with multiple subjects (shown in the second and third rows of Figure2.5(c)), and is designed to evaluate performance under realistic conditions as well as cross-environment and cross-subject generalization.

- MMVR provides extensive annotations, including approximately **446K** bounding boxes, **7.59 million** keypoints, and **446K** instance segmentation masks (see Table2.3). High-confidence image-plane labels are generated using state-of-the-art RGB-based pipelines—specifically Mask2Former [177] and HRNets [178]—followed by human verification, enabling three perception tasks: object detection, pose estimation, and instance segmentation.

- We further adapt and re-implement representative radar perception baselines, including RF-Pose [143] and RFMask [131], to all three tasks and establish benchmark results on both **P1** and **P2**. Comprehensive ablation studies investigate the influence of key factors such as temporal context length, the number of radar viewpoints, and dataset split strategies.

Table 2.3: Comparison of multi-view heatmap-based indoor radar datasets.

Dataset	**Resolution**			**Annotations**			**Statistics**					
	Range	Azi.	Ele.	BBox	KP	Seg	Rooms	Subjects	Actions	Sequences	Frames	Room Setting
RF-Pose†	10.0 cm	15°	15°	×	✓	×	50	100	free-form	-	-	-
HuPR	4.8 cm	15°	15°	141K	1.97M	×	1	6	static, walking	235	141K	open space
HIBER	12.2 cm	1.3°	1.3°	231K	3.23M	231K	1	10	free-form	152	179K	open foreground
MMVR(ours)	11.5 cm	1.3°	1.3°	**446K**	**7.59M**	**446K**	**6**	**25**	free-form	**395**	**345K**	open & **cluttered**

†RF-Pose is not publicly available.

2.2.2 Existing Indoor Radar Perception Datasets

As summarized in Table 2.2, a number of indoor radar perception datasets have been introduced in recent years. The majority of these datasets primarily investigate *single-view low-resolution (LR) point cloud* representations and focus on relatively constrained tasks such as action recognition or pose estimation. RadHAR [105], for example, constructed a mmWave radar dataset based on point cloud measurements to classify only five types of human activities performed by two subjects. mm-Pose [113] adopted a CNN-based framework to estimate and track 25 body joints under four simple motion patterns—walking, left-arm swing, right-arm swing, and both-arm swing—but the dataset remains small in scale, containing only 39.7K frames. mmMesh [112] further explored the use of single-view LR radar point clouds to directly reconstruct realistic human meshes by incorporating parametric body models, addressing point cloud sparsity through learned priors. This dataset includes 480K frames collected from 20 subjects performing eight daily activities within a limited open-space setting. Both mm-Pose and mmMesh are not publicly released. More recent efforts such as mRI [122] and MM-Fi [69] target indoor rehabilitation scenarios, covering 14 daily actions and 13 rehabilitation exercises. These datasets integrate multi-modal sensing pipelines while incorporating single-view mmWave radar point clouds as one of the sensing modalities.

In contrast to point-cloud-based datasets, RF-Pose [143] and HuPR [141] utilize two orthogonal radar heatmaps with an angular resolution of 15° in both azimuth and elevation. RF-Pose features large-scale data collection with 100 subjects across 50 rooms, but it is not publicly accessible. HuPR, while openly released, provides 141K frames annotated with 2D bounding boxes and keypoints, yet exhibits limited diversity in terms of environments, participants, and action categories, and does not include segmentation labels, as shown in Table 2.3. HIBER partially released 179K frames collected from 10 subjects using two orthogonal high-resolution radar heatmaps in a single open-foreground room. While HIBER represents a step toward higher-resolution radar perception, it remains constrained in environmental diversity. Positioned between the closed RF-Pose dataset and the limited-scope HIBER release, MMVR is designed to substantially expand coverage across rooms, subjects, and tasks. This diversity enables systematic cross-environment and cross-subject evaluation and supports the development of more generalizable radar perception models. In addition, MMVR provides nearly twice the number of annotated labels compared to HIBER.

2.2.3 MMVR Sensor Setup

For MMVR, we construct a mobile sensing platform that integrates two high-resolution radar units together with an RGB-D camera, as illustrated in Figure 2.6(a). Along the red processing path, an Intel RealSense D455 camera captures visual observations of the scene shown in Figure 2.6(b), which are subsequently used to generate image-based annotation labels via pretrained models, as

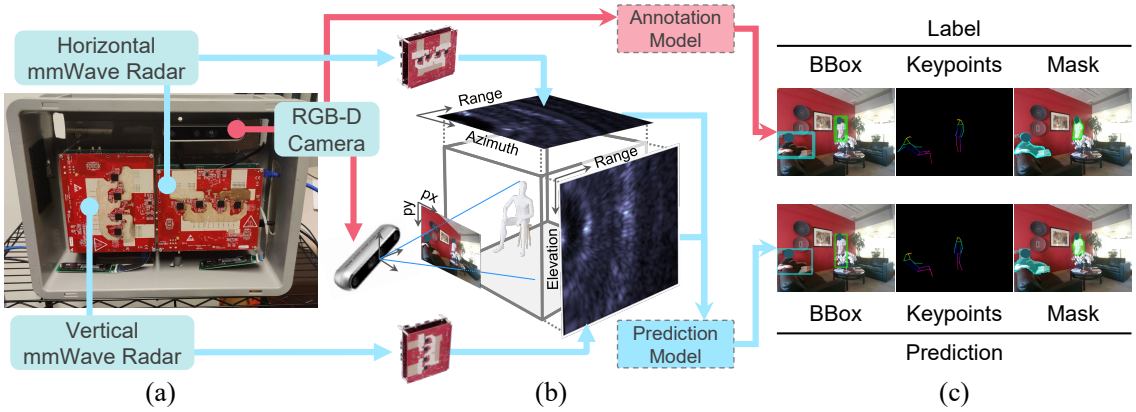


Figure 2.6: MMVR sensor configuration: (a) a setup consisting of two mmWave radars and an RGB-D camera; (b) each frame contains a pair of orthogonal high-resolution radar heatmaps synchronized with an RGB image; (c): the two radar heatmaps are used as inputs to radar perception models for the three perception tasks, supervised by labels derived from the RGB modality.

visualized in Figure 2.6 (c). In parallel, following the light-blue processing path, two TI AWR2243 mmWave cascade radars acquire radar measurements from complementary viewpoints, producing range–azimuth and range–elevation representations of the same scene (Figure 2.6(b)). These two radar views are then provided as inputs to baseline perception models, which generate predictions for the three downstream tasks shown in Figure 2.6 (c).

RGB-D Camera

Owing to the compact size of the RealSense camera, it is mounted on the upper-right section of the testbed enclosure, positioned directly above the horizontally oriented radar sensor. The Intel RealSense D455 supports RGB image resolutions of up to 1280×800 and stereo depth maps of up to 1280×720, with an effective depth sensing range of approximately 6 m, depending on ambient lighting conditions. During data acquisition, we configure the camera to capture RGB images at a resolution of 480×640 and a frame rate of 15 fps. As illustrated in Figure 2.6 (b), the captured RGB-D data can be mapped from the three-dimensional camera coordinate system onto the two-dimensional image plane using a standard pinhole camera model. The timestamps of the RGB-D frames are synchronized with the host desktop system, achieving a temporal alignment accuracy of approximately 50 microseconds.

Multi-View High-Resolution Radar

To obtain fine angular resolution along both azimuth and elevation axes, we deploy two AWR2243 cascade radar units in mutually orthogonal orientations, as depicted in Figure 2.6(a). By coherently integrating four single-chip FMCW radar devices, each cascade radar synthesizes a virtual antenna array comprising 86 elements spaced at half-wavelength intervals, achieving an angular resolution

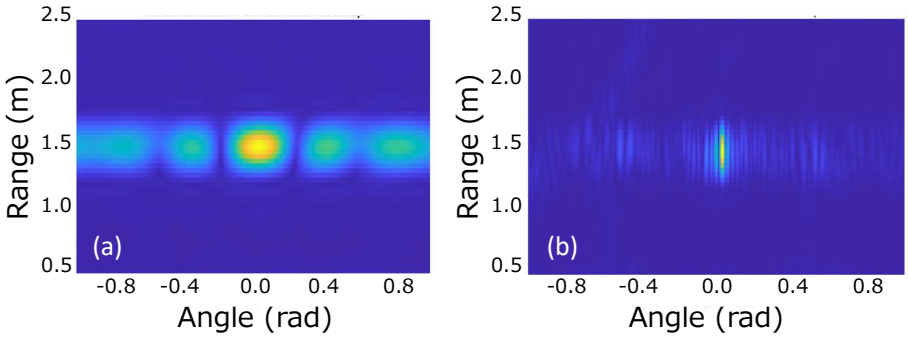
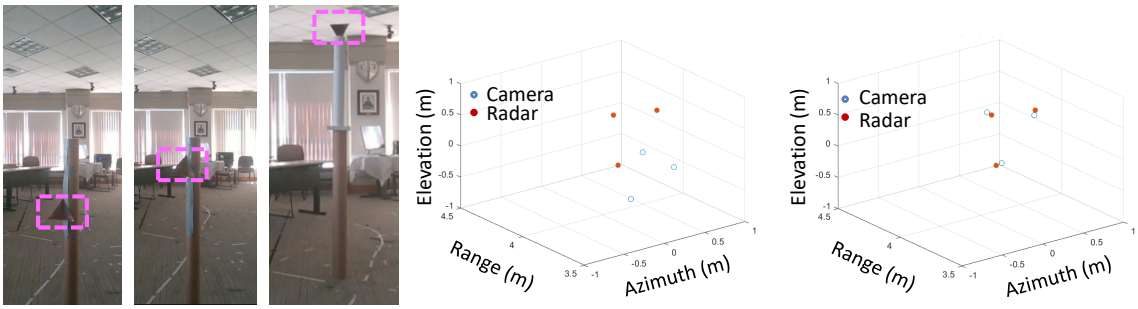


Figure 2.7: Comparison between (a) low-resolution and (b) high-resolution radar heatmaps of a corner reflector.



(a) Corner reflector at selected locations

(b) 3D positions at separate Camera and Radar coordinates

(c) 3D positions after Camera-to-Radar calibration

Figure 2.8: Camera-Radar coordinate calibration.

of approximately 1.3° in the boresight direction[1]. This resolution is substantially higher than the roughly 15° angular resolution provided by commonly used single-chip radar platforms, such as IWR1443 and IWR1843, which are employed in datasets including RadHAR [105], mRI [122], MM-Fi [69], and HuPR [141]. With this perpendicular sensor arrangement, the system simultaneously captures high-resolution radar observations in the range–azimuth and range–elevation domains, as shown in Figure 2.6(b). Figure 2.7 further contrasts the range–azimuth heatmaps of a corner reflector obtained using high-resolution AWR2243 radars versus low-resolution IWR1843 sensors.

Calibration

The spatial alignment between the camera and radar coordinate frames is obtained by recording the three-dimensional locations of a corner reflector placed at multiple positions, measured independently in the camera and radar coordinate systems. Let the corresponding point sets be denoted as $\mathbf{B}\text{camera}, \mathbf{A}\text{radar} \in \mathbb{R}^{3\times N}$, where N represents the total number of calibration samples. The calibration problem is formulated as estimating a rigid transformation—consisting of a rotation

[1]The boresight direction refers to the direction in which the antenna array exhibits maximum gain.

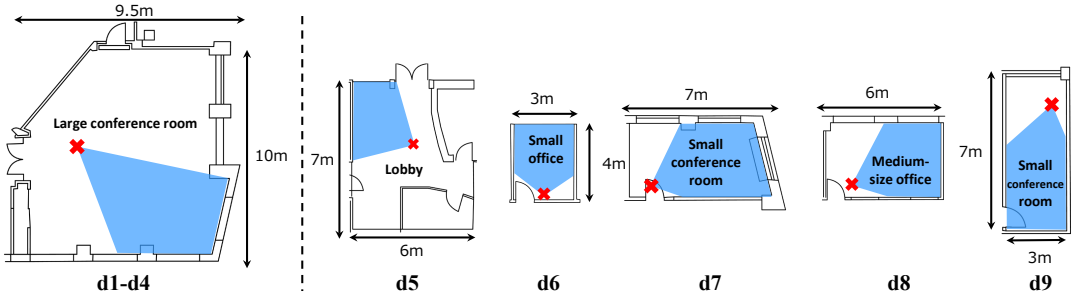
matrix $\mathbf{R}$ and a translation vector $\mathbf{t}$—that minimizes

$$\min_{\mathbf{R},\mathbf{t}} \sum_{i=1}^{N} |\mathbf{R}, \mathbf{B}\text{camera}[:, i] + \mathbf{t} - \mathbf{B}\text{radar}[:, i]|_2^2 . \tag{2.22}$$

This optimization can be efficiently solved via singular value decomposition (SVD), which yields the optimal rigid alignment between two 3D point sets. Figure 2.8 illustrates the corner reflector placements used for calibration (a), along with the relative positions observed in the camera and radar frames before (b) and after (c) applying the estimated transformation.

2.2.4 Dataset

MMVR is designed to support multiple perception objectives - including object detection, pose estimation, and instance segmentation - across a wide range of indoor conditions involving 6 distinct rooms and 25 participants.



(a) **P1**: open foreground in a single room (b) **P2**: cluttered space in multiple rooms

Figure 2.9: Floorplans of 6 indoor environments in Protocol 1 and Protocol 2.

Data Collection

MMVR comprises a total of 345K data frames acquired under 2 data collection protocols:

- **Protocol1 (P1: Open Foreground)**: 107.9K frames captured in a single open-area room with one participant. The corresponding floorplan is shown in Figure 2.9(a), and an example snapshot is provided in the first row of Figure 2.5(c). Data were recorded across four separate days (d1–d4), with one or two sessions conducted per day. Throughout these sessions, the participant performed activities such as walking and jumping without obstruction, remaining fully visible to both the radar sensors and the RGB camera.

- **Protocol2 (P2: Cluttered Space)**: 237.8K frames collected in five cluttered indoor rooms involving single or multiple participants. Figure 2.9(b) illustrates the corresponding floorplans, while example scenes are shown in the second and third rows of Figure 2.5(c). Beginning on Day5 (d5), six data collection sessions were conducted in each room. Participants engaged in a variety of everyday activities, including walking, sitting, stretching, reading, writing on a board, and conversing.

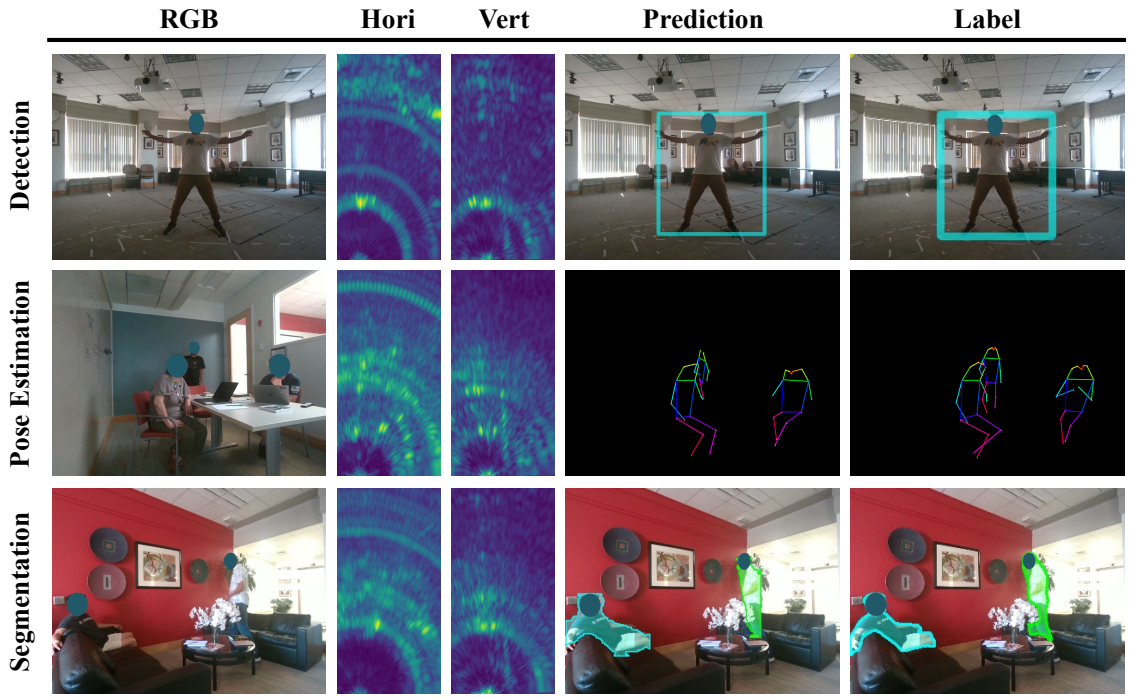


Figure 2.10: Visualization of data collection process: P1 (open-foreground scene in the first row) and P2 (cluttered environments in the second and third rows). The columns show the RGB images (first column), the corresponding multi-view (horizontal and vertical) radar heatmaps (second and third columns), baseline prediction outputs (fourth columns), and the associated annotation labels (fifth columns).

All frames are partitioned into non-overlapping temporal segments of one minute, each containing approximately 900 frames based on a sampling rate of 15 fps. Every frame consists of an RGB image, a horizontal radar heatmap, a vertical radar heatmap, and associated annotations for bounding boxes, instance masks, and keypoints. In total, the dataset includes 122 segments for P1 and 273 segments for P2.

Annotation

We rely on pretrained vision-based deep learning frameworks, Mask2Former [177] and HRNet [178], to automatically generate annotation labels for MMVR.

- **Object Detection:** Each RGB frame is first processed by a pretrained visual backbone, namely the Swin Transformer, to obtain a set of low-resolution feature maps. Mask2Former [177, 179] then refines these representations through its pixel decoder to recover high-resolution features. The decoder operates on a fixed set of queries, attending to the encoded feature maps and mapping them to object categories and bounding box predictions. In MMVR, only detections corresponding to the "person" class are retained. Each bounding box is represented in the form $(n, 5)$, where n denotes the number of detected instances and each row corresponds to $[x_{\min}, y_{\min}, x_{\max}, y_{\max}, \text{confidence}]$. An illustrative example with $n = 1$ is

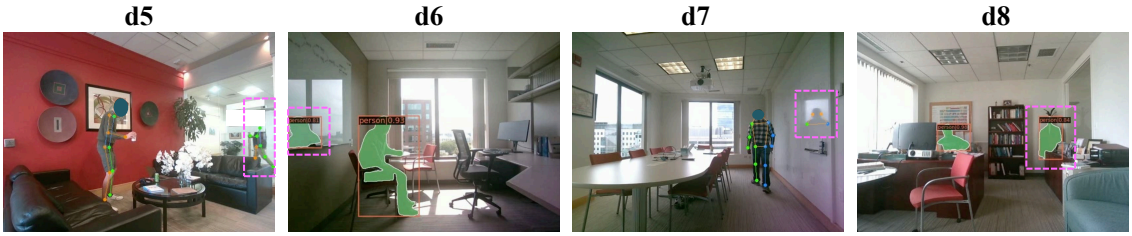


Figure 2.11: Annotation curation: remove spurious labels (magenta boxes) due to the human presence behind the glass, highly reflective whiteboards, and a hanging suit.

shown in the first row of Figure 2.10 under the Label column.

- **Pose Estimation:** For keypoint extraction, we adopt HRNet [178, 180], which preserves high-resolution feature representations by maintaining parallel convolutional branches at multiple scales and repeatedly fusing information across resolutions. This design is well suited for multi-person pose estimation. Each RGB image is processed using an HRNet model pretrained on the COCO Keypoints 2017 dataset to obtain 17 body keypoints per detected person. The resulting annotations are stored in an array of shape $(n, 17, 3)$, where each keypoint is encoded by its image-plane coordinates and an associated confidence score. An example with $n = 3$ subjects is shown in the second row of Figure 2.10.

- **Instance Segmentation:** Mask2Former [177, 179] is further employed to generate instance-level segmentation masks by extending the detection pipeline. The query embeddings are transformed into binary masks and corresponding class predictions, conditioned on the pixel-decoder features. Each segmentation result is represented as a Boolean tensor of size $(n, 480, 640)$, matching the spatial resolution of the RGB input. Figure 2.10 (third row, Label column) provides an example with $n = 2$ segmented individuals.

Annotation Curation

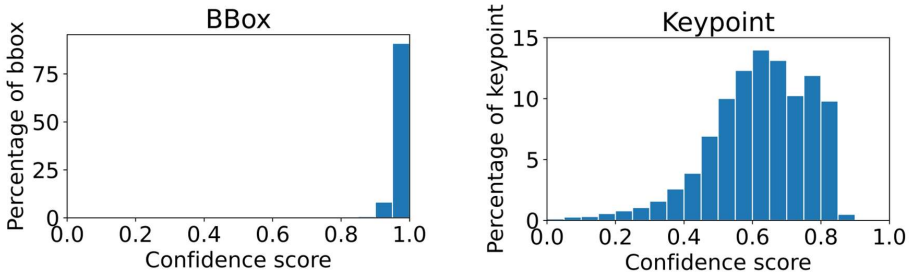


Figure 2.12: Quality of curated annotations.

While the automated annotation pipelines described above enable large-scale label generation, they inevitably introduce incorrect or irrelevant annotations. For example, in the lobby scene recorded on d5 (Figure 2.11), vision-based models produce high-confidence detections for pedestrians visible through glass walls outside the lobby, whereas the 77 GHz radar sensors exhibit limited penetration through glass surfaces. These mismatched annotations are therefore removed from the affected frames. Similar inconsistencies are observed in office environments containing highly reflective whiteboards (d6 and d7 in Figure 2.11) or suspended clothing items such as suits (d8 in Figure 2.11). After manual filtering, Figure 2.12 presents the distributions of confidence scores for the remaining bounding box and keypoint annotations.

Preprocessing for Multi-View Radar Heatmaps

Figure 2.13 summarizes the preprocessing pipeline used to generate multi-view radar heatmaps from the two orthogonally mounted radar units, each forming a virtual antenna array with 86 half-wavelength-spaced elements and transmitting multiple FMCW pulses. The received radar signals are sampled to construct a three-dimensional data cube indexed by the virtual array dimension (horizontal or vertical), fast-time ADC samples, and slow-time pulse indices. A three-dimensional

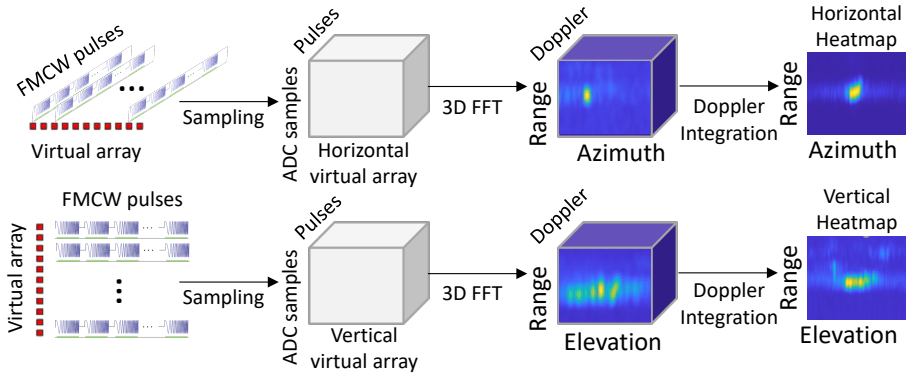


Figure 2.13: Multi-view heatmap preprocessing.

fast Fourier transform (FFT) is then applied to this data cube to obtain spectral representations over angle (azimuth for the horizontal radar and elevation for the vertical radar), range, and Doppler velocity. To improve signal quality, the spectra are integrated along the Doppler dimension, producing range–azimuth and range–elevation heatmaps in polar coordinates. These heatmaps are subsequently mapped into the radar Cartesian frame and, using the camera–radar calibration described in Section 2.2.3, transformed into range–azimuth and range–elevation views expressed in the camera coordinate system.

Synchronization

To temporally align the radar sensors with the camera, we adopt a two-stage synchronization strategy consisting of a coarse alignment followed by a refinement step. In the first stage, a metallic corner reflector is positioned in front of the MMVR sensing platform at the start of each recording session, enabling us to determine a common reference frame across the horizontal radar, vertical radar, and RGB camera streams. Despite this initial synchronization, independent sensor clocks can gradually drift during long recordings. To compensate for this effect, a second alignment stage is applied by manually adjusting the sensor timelines so that salient motion events, such as sitting down or standing up, coincide across all three modalities.

2.2.5 Evaluation and Benchmarks

This section describes the experimental protocol, including the benchmark configuration, evaluation metrics, baseline approaches, and overall results.

Benchmark Setup

To assess model performance under different generalization settings, we adopt two complementary data partitioning strategies:

- **Data Split 1 (S1: Random Split):** All non-overlapping one-minute segments (122 segments for P1 and 273 segments for P2) are randomly divided into training, validation, and test subsets using an 80:10:10 ratio.
- **Data Split 2 (S2: Cross-Session and Unseen Split):** Data segments collected on d5, d6, d7, and d9 are first separated into training, validation, and test sets. In addition, all segments from d8 are reserved exclusively for testing, enabling explicit evaluation of model generalization to a previously unseen environment.

Metrics and Methods

- **Object Detection:** Detection performance is measured using average precision based on the intersection-over-union (IoU) criterion [181]. IoU quantifies the spatial agreement between a predicted bounding box A and its ground-truth counterpart B as

$$\mathrm{IoU}(A, B) = \frac{|A \cap B|}{|A \cup B|}. \tag{2.23}$$

  We report three standard variants: AP50 and AP75, which correspond to relaxed and stringent IoU thresholds, respectively, as well as AP, defined as the mean precision averaged over ten IoU thresholds ranging from 0.5 to 0.95 with an increment of 0.05. RFMask [131] is adopted as the baseline detector, with modifications that allow bounding-box regression

directly in the image plane using annotated BBoxes, rather than computing losses on radar-view planes as in HIBER.

- **Pose Estimation:** For pose estimation, we employ average precision based on object keypoint similarity (OKS) [182]. OKS provides a scale-normalized measure of localization accuracy and is defined for the i-th keypoint as

$$\mathrm{OKS}_i = \exp\left(-\frac{d_i^2}{2s^2k_i^2}\right), \quad i = 1, \ldots, 17, \tag{2.24}$$

  where d_i denotes the Euclidean distance between the predicted and ground-truth keypoints, k_i is a joint-specific normalization constant, and s represents the target subject's scale. As with detection, AP_{50}, AP_{75}, and averaged AP scores are computed across all 17 keypoints. The RF-Pose framework [143] is reimplemented from the ground up as the baseline, and further extended to jointly predict keypoint heatmaps and part affinity fields (PAFs) [183] to improve multi-person association.

- **Instance Segmentation:** Instance segmentation results are generated using RFMask, and evaluation is performed using average precision based on IoU. In this case, IoU is computed at the pixel level rather than over bounding boxes, considering the overlap between predicted and ground-truth person masks, following the standard definition in [181].

2.2.6 Main Results

Object Detection

Table 2.4 summarizes the object detection performance of the RFMask baseline when four consecutive radar frames are used as input. Under Protocol1, the method attains AP_{50} values exceeding 65 for both data partitioning strategies. These results are in line with detection performance reported in prior work. As a reference, RF-Pose [143] achieved an AP of 30.8 using six radar frames, which is comparable to the AP scores of 25.53 and 24.46 obtained on MMVR under Protocol 1. From a qualitative perspective, detection errors are more frequently observed when subjects are positioned at distances of approximately 1 m or less from the radar sensors.

Under Protocol 2, a clear performance gap emerges between the two split strategies: models evaluated under S1 significantly outperform those tested under S2, with notable degradation across all AP metrics in the latter case. This discrepancy highlights the limited ability of the baseline approach to generalize, especially when confronted with previously unseen and cluttered indoor environments.

Pose Estimation

Table 2.4 further reports the baseline results for pose estimation obtained with RF-Pose using sequences of four consecutive radar frames. For Protocol1, the method reaches AP_{50} values of

Table 2.4: Baseline benchmarks for 3 tasks under 2 protocols and 2 data splits.

Protocol	Split	Object Detection			Pose Estimation			Instance Seg.
		AP	AP_{50}	AP_{75}	AP	AP_{50}	AP_{75}	IoU
P1	S1	25.53	67.30	15.86	46.24	62.88	47.45	71.98
	S2	24.46	66.82	11.22	29.82	43.03	30.29	67.03
P2	S1	31.37	61.50	27.48	32.13	44.22	32.58	65.30
	S2	6.03	22.77	0.88	7.11	11.98	6.76	56.07

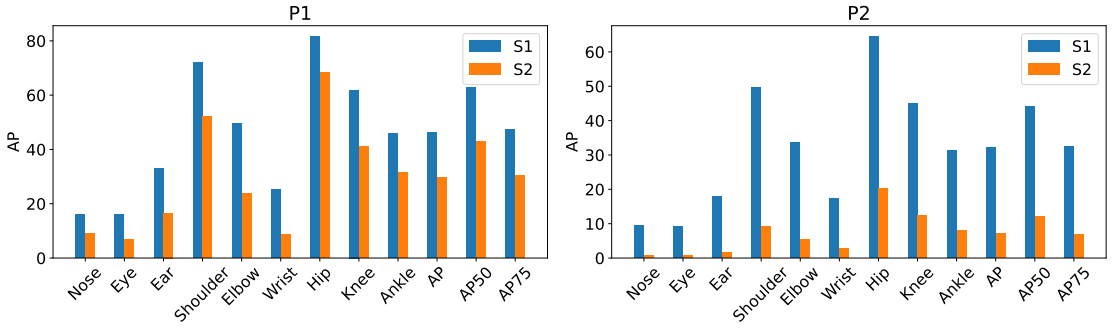


Figure 2.14: AP breakdown over keypoints under 2 protocols and 2 data splits.

approximately 63 under S1 and 43 under S2. In contrast, when evaluated on Protocol2, RF-Pose delivers moderate performance for the random split (S1), achieving an AP_{50} score of 44.22, but its accuracy deteriorates markedly under the cross-session and unseen split (S2), mirroring the trend observed earlier for object detection. Figure 2.14 provides a more detailed analysis by presenting AP scores for individual keypoints. The results indicate that body extremities and facial landmarks tend to exhibit lower precision compared with more stable torso-related keypoints.

Instance Segmentation

The final column of Table 2.4 reports the instance segmentation results produced by RFMask, which remain comparatively stable across the four experimental settings formed by the 2 protocols and 2 data splits. Within the same protocol (either P1 or P2), a modest decrease in performance is observed when moving from S1 to S2, indicating a relatively stronger robustness to unseen environments. By contrast, the performance variation across different protocols is slightly more pronounced than that observed across data splits, suggesting that changes in scene conditions have a greater impact than the choice of split strategy.

Visualization

Figure 2.10 presents qualitative visualization results for all three perception tasks in the Prediction column. These examples demonstrate that the baseline approaches can effectively separate radar responses associated with human subjects from background reflections.

Table 2.5: **# of Frames**: Using longer time horizon improves the performance.

# of frames	Detection			Pose Estimation			Segmentation
	AP	AP_{50}	AP_{75}	AP	AP_{50}	AP_{75}	IoU
4	25.53	67.30	15.86	46.24	62.88	47.45	71.98
12	**33.12**	71.17	**26.84**	**46.86**	**63.22**	**48.14**	71.81
18	32.78	**72.03**	25.44	46.70	63.07	47.95	**73.28**

2.2.7 Ablation Study

We report on several ablation studies using S1 under P1.

Number of Frames

As summarized in Table 2.5, increasing the temporal length of the input sequence generally improves performance across all tasks. Notably, object detection accuracy benefits substantially when the input is expanded to 12 consecutive radar frames. However, when extending the sequence further to 18 frames—corresponding to approximately 1.2 seconds at a frame rate of 15 fps—the performance shows a slight decline compared with the 12-frame setting.

Single View versus Dual View

Table 2.6 compares the performance obtained when using a single radar view, specifically, the horizontal view alone, against that achieved by jointly exploiting both horizontal and vertical views. The results show consistent gains for both RFMask and RF-Pose when dual-view inputs are employed. This suggests that although the horizontal perspective already captures sufficient information to infer subject position and approximate scale in the image plane, incorporating the vertical view contributes complementary cues in the elevation dimension, enabling richer feature representations and more discriminative characterization of different body parts.

Comparison of Detection Backbones

We further evaluate a more recent DETR-based architecture [184] as an additional baseline, applying it to radar inputs with minimal task-specific modification. Table 2.7 reports a comparison between RFMask and DETR for object detection and instance segmentation. The results indicate that DETR consistently outperforms RFMask on both tasks, suggesting that the use of backbone self-attention together with query–feature cross-attention provides more effective feature modeling for radar-based perception.

Comparison over Different Radar Datasets

A potentially insightful ablation would be to train a radar perception model on one dataset (e.g., MMVR) and test it on another (e.g., HIBER), or vice versa. Conducting such cross-dataset evaluations, however, is nontrivial because it would require dataset-specific calibration between

Table 2.6: Single/Dual: Using the dual views works the best.

Method	View	AP	AP_{50}	AP_{75}	IoU
RFMask	Single	24.64	66.17	15.21	70.64
	Dual	**25.53**	**67.30**	**15.86**	**71.98**
RF-Pose	Single	22.22	30.99	22.64	-
	Dual	**46.24**	**62.88**	**47.45**	-

Table 2.7: Method: DETR can improve the performance.

Method	View	AP	AP_{50}	AP_{75}	IoU
RFMask	Single	24.64	66.17	15.21	70.64
DETR	Single	**35.78**	**78.59**	**28.02**	**72.65**

Table 2.8: Baseline benchmarks between HIBER [131] and MMVR.

Room Setting	Dataset	Object Detection			Instance Seg.
		AP	AP_{50}	AP_{75}	IoU
Open Foreground	HIBER (Walk)	17.77	52.46	6.78	48.47
	MMVR (P1)	24.46	66.82	11.22	67.03

radar and camera coordinates, as well as consistent 3D-to-2D projection into the image plane. These steps are necessary due to differences in sensor placement and camera hardware across datasets. To avoid these complications, we instead perform a direct comparison of identical tasks across the two datasets under closely matched conditions—specifically, scenarios involving a single subject walking in an open-foreground room. This comparison is carried out using the "walk" split from HIBER and the "P1" protocol from MMVR. As reported in Table 2.8, the resulting performance levels are broadly comparable between the two datasets. We note that, for the HIBER evaluation, refined bounding boxes in the image plane are used, as the original annotations were overly loose and encompassed regions significantly larger than the subjects themselves.

2.2.8 Conclusion and Contribution

MMVR expands indoor radar data acquisition by leveraging multi-view, high-resolution heatmap representations collected across multiple days, rooms, and participants, with a particular focus on environmental and subject diversity. The dataset comprises 345K radar heatmap frames gathered from 25 individuals in 6 distinct indoor spaces, supporting benchmark evaluations for three core perception tasks—object detection, pose estimation, and instance segmentation—with 446K annotated bounding boxes and segmentation masks, along with 7.59 million labeled keypoints. We envision MMVR as a catalyst for advancing robust radar perception methods capable of functioning reliably across heterogeneous real-world scenarios, including smart homes, security monitoring, eldercare, and navigation assistance, thereby delivering broader societal benefits.

MMVR directly contributes to the dissertation's first objective to enhance signal accessibility for wireless sensing research, by providing an open high-quality, and easily accessible dataset for the community. Since its public release [185], the dataset has been downloaded more than 2,500 times, indicating its growing influence and utility as a foundation for future radar perception research, a direction we further explore in Chapter 6.

Chapter 3

Wi-Fi Sensing with Standard-compliant Signal Formats

3.1 CBR-ACE: Activity Recognition with Standardized Beamforming Reports

As outlined in Chapter 1, one major challenge for practical wireless sensing is the limited accessibility of channel information, since CSI is only available on a few legacy chipsets with non-standard modifications. Another challenge lies in the need for methodologies that enable reproducible recognition of human activities on commodity hardware. This chapter addresses both issues by developing CBR-ACE, which leverages the standard-compliant CBRs introduced in Sec 2.1 and provides an efficient signal-processing pipeline for activity count estimation. In doing so, CBR-ACE demonstrates how practical Wi-Fi sensing applications can be realized without chipset-specific dependencies or heavy model complexity. In this chapter, we build upon the accessibility provided by WiPiCap and focus on addressing the methodology challenge, demonstrating how CBRs can be practically leveraged for human activity recognition.

3.1.1 Introduction

Human activity recognition has attracted extensive attention across a range of sensing modalities, spanning vision-based techniques and sensor-driven approaches. Camera-based solutions often deliver strong recognition accuracy but inherently depend on unobstructed line-of-sight and introduce privacy-related concerns. In contrast, wearable devices or ambient sensors are typically easier to deploy, yet their performance is sensitive to placement and constrained by battery life. Wireless sensing has therefore gained momentum as an alternative paradigm, leveraging variations in radio channels of commodity Wi-Fi systems to infer human activities, even under non-line-of-sight (NLoS) conditions. However, many existing studies depend on CSI available only from

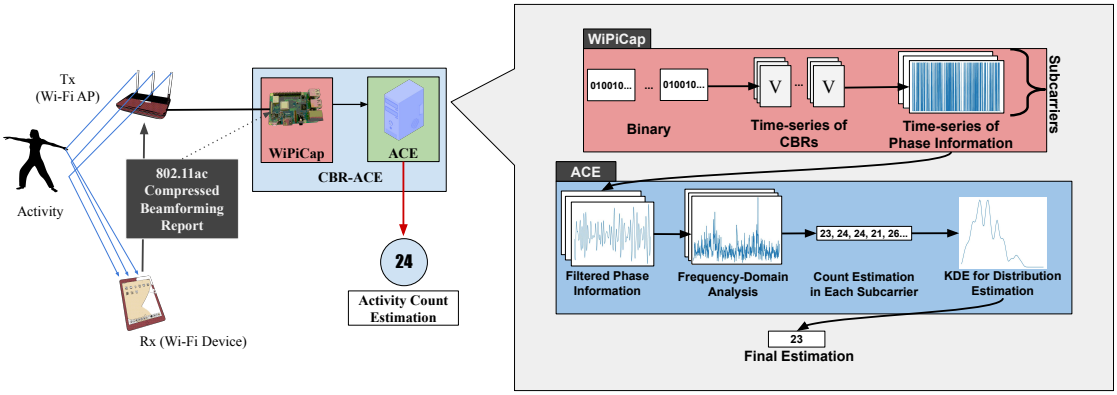


Figure 3.1: Overall structure (left) and procedure (right) of our CBR-ACE. CBR-ACE consists of WiPiCap and ACE for human exercise recognition using the CBRs.

specific legacy network interface cards (e.g., Intel 5300 under IEEE 802.11n), which limits both reproducibility and compatibility with modern hardware.

As detailed in Chapter 2.1, CBRs standardized in IEEE 802.11ac/ax offer a compliant substitute for CSI, making Wi-Fi sensing feasible across a broader set of off-the-shelf devices. Building upon this capability, we explore the direct use of CBRs for human activity recognition. Our focus is on repetitive indoor exercises such as push-ups and squats, where accurately estimating repetition counts serves as a practical indicator of daily physical activity.

To address this task, we introduce a wireless sensing framework termed the CBR-based Activity Count Estimator (**CBR-ACE**). The system consists of two main components: (i) a CBR acquisition pipeline, WiPiCap, described earlier in Chapter 2.1, and (ii) a lightweight signal-processing module, ACE, designed to infer exercise repetitions from CBR time-series measurements.

This work makes four primary contributions. First, to the best of our knowledge, CBR-ACE is the first wireless sensing system that estimates human activity counts using standard-compliant CBR measurements. Second, it presents WiPiCap as an integrated toolchain that combines packet capture, binary parsing, and CBR time-series construction. Third, the proposed ACE method shows that exercise repetition counting can be achieved using classical signal processing techniques, significantly reducing computational demands compared with machine-learning or deep-learning approaches. Finally, experimental results demonstrate that CBR-ACE attains high estimation accuracy, achieving an average counting error of 1.73 and reaching as low as 0.97 under favorable conditions.

3.1.2 Proposed System: CBR-ACE

Overview

The CBR-ACE system is designed to estimate counts of human activities that occur repeatedly at roughly constant intervals, using CBR measurements obtained from IEEE802.11ac–compliant devices. The left panel of Figure3.1 depicts the assumed system configuration. In this setup, a Wi-Fi access point and a Wi-Fi client device are deployed within the same room, both operating under the IEEE 802.11ac protocol. A dedicated sniffer, WiPiCap, is connected to the access point via a wired local area network (LAN). By continuously transmitting UDP packets to the client through the access point, WiPiCap induces repeated CBR feedback requests from the AP to the device, while simultaneously capturing the exchanged CBR packets. As a user performs a repetitive indoor activity at a fixed location, CBR-ACE estimates the number of repetitions based on the collected CBR measurements.

The right panel of Figure 3.1 summarizes the processing pipeline of CBR-ACE. Using WiPiCap, the system first constructs time-series representations of CBR values for each subcarrier. Accurate counting is achieved through three main processing steps. First, phase information extracted from the CBR time series is used as the core feature for activity analysis. Second, the phase signals are denoised by applying a bandpass filter followed by an exponential moving average (EMA) filter to reduce spurious fluctuations. Finally, Kernel Density Estimation (KDE) is applied across all subcarriers to consolidate the intermediate estimates and produce the final activity count.

ACE: Signal Processing-based Activity Count Estimation

ACE derives human activity counts by exploiting phase information extracted from CBR measurements together with classical signal processing techniques. Concretely, the phase time series is transformed into the frequency domain via a Fourier transform, and the repetition count is inferred from the dominant frequency component. Because the method does not rely on any learning-based model, ACE avoids training overhead and remains computationally lightweight.

As a preprocessing step, ACE suppresses noise present in the phase measurements. Such noise primarily originates from hardware imperfections of the access point or Wi-Fi device, as well as reflections from objects unrelated to the target person's motion. Hardware-induced noise is often bursty and exhibits large amplitudes [186]; to mitigate this effect, an EMA filter is applied. The EMA filter uses a window spanning five packets (approximately 250 ms) to attenuate burst noise, followed by a band-pass filter that removes frequency components irrelevant to the target activity. Since the target exercises typically occur at around 1 Hz, the band-pass filter [187] retains frequencies within the range of 0.5 Hz to 1.5 Hz.

When an exercise is performed with an approximately constant repetition interval, the corresponding frequency component becomes prominent in the spectrum. ACE therefore selects, for

each subcarrier, the frequency with the maximum spectral power as a candidate estimate for the activity repetition rate.

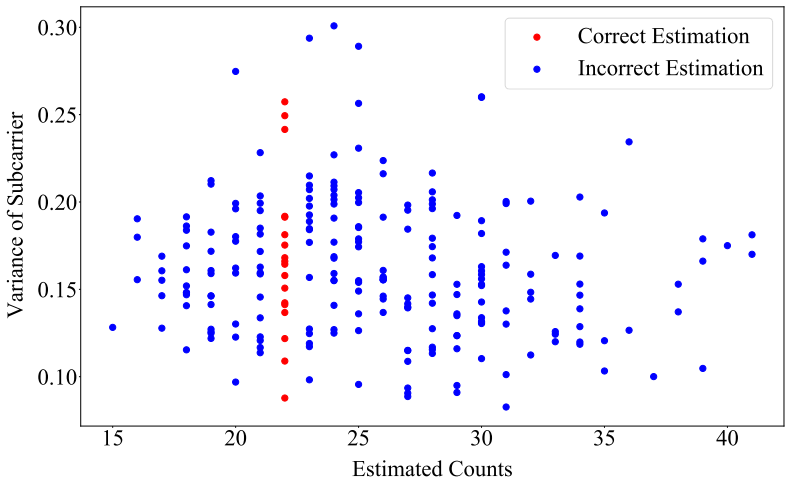


Figure 3.2: Example: relationship between subcarrier variance and correct/incorrect estimation results.

Since each subcarrier undergoes a narrowband fading process influenced by human motion, identifying suitable subcarriers is crucial for reliable count estimation. Prior work based on full CSI typically favors subcarriers exhibiting higher variance [51], under the assumption that variance reflects motion sensitivity. In our experiments with CBRs, however, we do not observe a clear relationship between subcarrier variance and estimation accuracy. Figure 3.2 illustrates this behavior for a representative example in which the true repetition count is 22. Each point corresponds to a single subcarrier, characterized by its variance and the estimated count derived from that subcarrier. Subcarriers yielding exact estimates are highlighted in red. As shown, large variance alone does not guarantee accurate estimation. Nevertheless, the estimates are not arbitrary: Figure 3.3 reveals that the predicted counts form a distribution concentrated around the correct value, indicating a consistent underlying structure despite subcarrier-level variability.

From these observations, we treat the estimate obtained from each subcarrier as an individual candidate, assuming that all candidates are drawn from an underlying probability distribution. ACE aggregates the estimates from all subcarriers by applying Kernel Density Estimation (KDE) to infer the final activity count. Specifically, KDE reconstructs a continuous probability density from a finite set of discrete count estimates. Let $x_1, x_2, \cdots, x_n \in \mathbb{R}$ denote independent and identically distributed (i.i.d) samples drawn from an unknown distribution F with density function f. The KDE of f, also known as the Parzen window estimator, is defined as:

$$\hat{f}(\mathbf{x}) = \frac{1}{n} \sum_{i=1}^{n} K_h \left(\mathbf{x}, x_i \right), \tag{3.1}$$

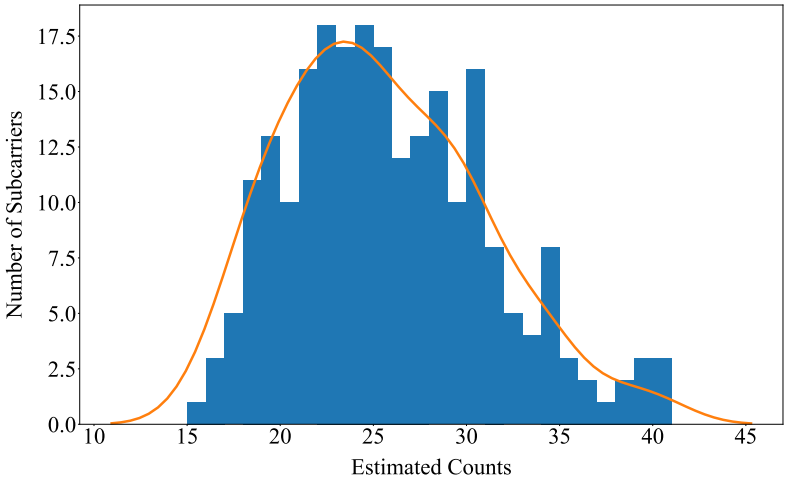


Figure 3.3: Example: a distribution of estimation values emerged from all subcarriers in a channel. The line in orange color indicates the estimated probability distribution function estimated by KDE.

where K_h denotes a kernel function parameterized by bandwidth h. To ensure that $\hat{f}(\mathbf{x})$ forms a valid probability density, the kernel function is required to satisfy:

$$K_h(\cdot,\cdot) \geq 0, \int_{-\infty}^{\infty} K_h(\mathbf{x},\cdot)d\mathbf{x} = 1.$$

In ACE, we adopt a Gaussian function for $K_h(\mathbf{x}, t)$, given by:

$$K_h(\mathbf{x}, x_i) = \frac{1}{\sqrt{2\pi}h} \exp\left(-\frac{||\mathbf{x} - x_i||^2}{2h^2}\right).$$

The kernel bandwidth h plays a critical role in shaping the resulting density estimate. An excessively small or overly large value of h leads to under-smoothing or over-smoothing of the estimated density $\hat{f}$, respectively. Various criteria have been proposed to select an appropriate bandwidth, including the mean integrated squared error (MISE) [188], Silverman's rule of thumb [189], and Scott's rule [190]. In ACE, KDE is employed not to recover the exact underlying distribution, but to identify a reliable count estimate from the inferred density. As such, the choice of h is particularly influential for estimation accuracy. We therefore determine the bandwidth via a grid search strategy. Concretely, KDE is performed using bandwidth values ranging from 1 to 10 with a step size of 0.1, and the optimal bandwidth is selected as the one that minimizes the discrepancy between the estimated continuous density and the original discrete distribution.

Following this procedure, ACE computes the cumulative distribution function (CDF) derived from the estimated density and adopts the 50th percentile as the final activity count estimate. Empirically, we find that this percentile-based estimate provides more stable and accurate results than alternatives such as the raw median, mean, or the mode of the density.

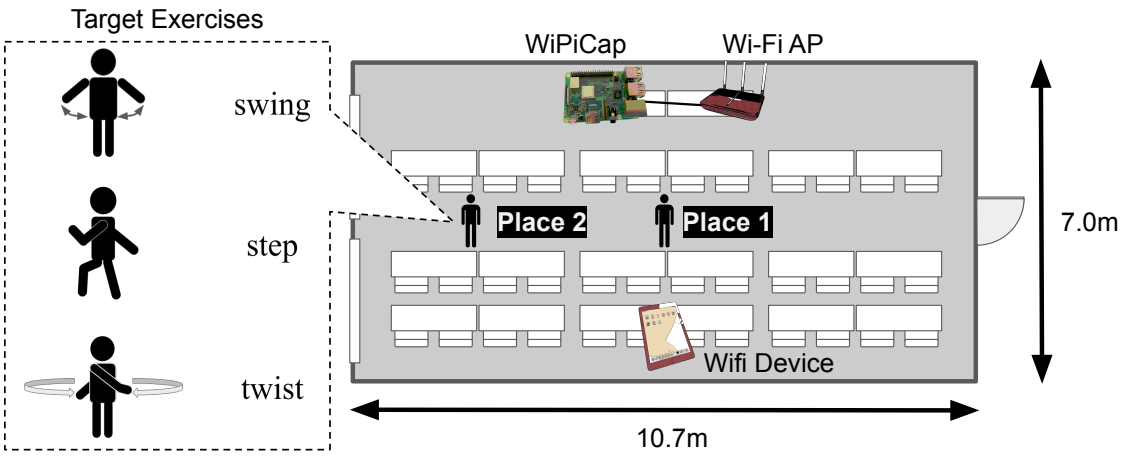


Figure 3.4: Experiment environment. We set one Wi-Fi AP and one Wi-Fi device in a room, and one subject repeats a target exercise at place 1 or place 2 for 30 s.

3.1.3 Experiments

Implementation

CBR-ACE is realized using commercially available off-the-shelf (COTS) wireless hardware. Specifically, a TP-Link Archer C6 served as the Wi-Fi access point, while a MacBook Pro (2020), an iPad (5th Generation), and a Galaxy Note 10 Lite are used as client devices. WiPiCap is deployed on a Raspberry Pi 3 Model B+ equipped with a BCM434355c0 chipset. All experiments are conducted in the 5 GHz band with an 80 MHz channel bandwidth. Although the evaluation focused on IEEE802.11ac, the proposed system is also compatible with IEEE802.11ax, as it relies on standard-compliant CBR feedback. In the experimental setup, only a single client device is associated with the AP at any given time; nevertheless, the system design naturally extends to scenarios with multiple concurrently connected devices. Throughout the measurements, the average CBR packet reception rate is approximately 9.7 Hz. Due to contention and interference from other Wi-Fi devices operating on the same channel, the packet arrival times are not evenly spaced. To address this issue, we applied linear interpolation to obtain uniformly sampled time series. Despite this non-uniformity, the effective sampling rate remained adequate for activity counting, since the target exercise cycles occurred at roughly 1 Hz.

Experiment Environment

Figure 3.4 illustrates the experimental setup and the surrounding environment. All measurements are conducted in a typical office room. During each trial, a single participant performed exercises for 30 seconds while standing either at position1 (P1) or position2 (P2), as indicated in Figure 3.4. PositionP1 lies along the direct line between the Wi-Fi access point and the Wi-Fi device, whereas

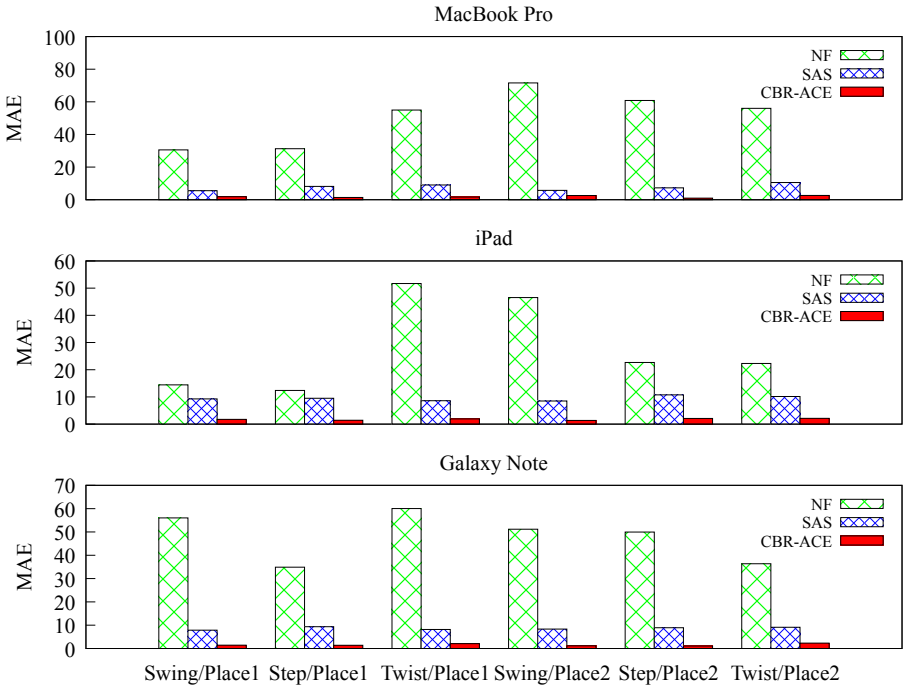


Figure 3.5: Accuracy improvement with filters and KDE: our filters and KDE are effective compared with NF and SAS.

positionP2 is located off this line. In total, the experiment is repeated 890 times by the same participant across a span of 18 days.

Target Activities

Figure 3.4 depicts the three activities considered in our experiments: swing, step, and twist. The swing activity corresponds to repetitive arm swinging, step refers to a stepping motion, and twist involves rotating the upper body left and right while keeping the pelvis relatively fixed. For each trial, we estimate the number of repetitions performed for each activity during a 30 s interval.

Metric

Counting accuracy is quantified using the mean absolute error (MAE), computed as the absolute difference between the ground truth repetition count and the estimated count, averaged over all trials.

Effect of Filters and KDE

CBR-ACE combines an EMA filter, a bandpass filter, and KDE to derive the final activity count estimate. To examine the individual contributions of these components, we evaluated the estimation performance with and without the filtering and KDE steps. Figure 3.5 reports the MAE for

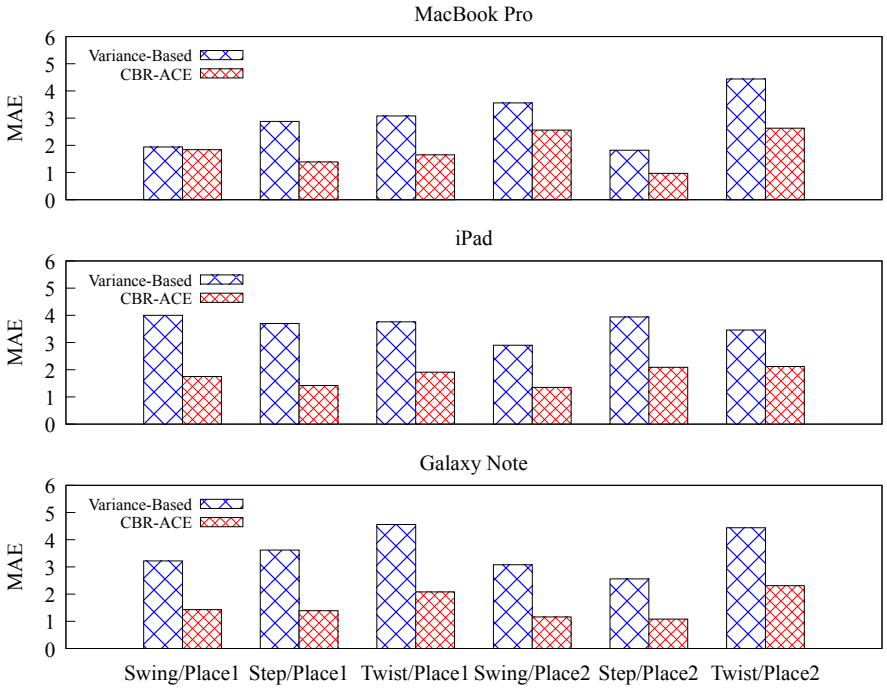


Figure 3.6: Comparison between the variance based method and proposed CBR-ACE.

each activity across different Wi-Fi devices. We consider two baseline variants: (i) **NF**, where no filtering is applied, and (ii) **SAS**, where filtering is applied but estimates from all subcarriers are simply averaged without KDE. Without any filtering, the estimation error is extremely large, with an average MAE of 42.43. Applying the filters and averaging subcarrier-wise estimates reduces the error substantially to 8.6 on average. In contrast, the full CBR-ACE pipeline, which incorporates both filtering and KDE-based aggregation, further lowers the average MAE to 1.73. These results indicate that the filtering stages and KDE-based fusion are both critical for achieving reliable and accurate activity count estimation, regardless of the activity type or device used.

Effect of Criteria for Count Estimation

We have shown that the combination of filtering and KDE enables accurate activity count estimation from CBRs. Building on this result, we next analyze how the choice of the final decision criterion influences estimation accuracy. As a reference, we implement a variance-based baseline inspired by [51], where estimation relies on subcarriers with the second-largest variance. To ensure a fair comparison, this baseline applies the same preprocessing steps as CBR-ACE, including identical filters and parameter settings on the phase time series.

Figure 3.6 compares the MAE of the proposed CBR-ACE and the variance-based approach [51] across different exercises and Wi-Fi devices. Several trends can be observed:

Table 3.1: Micro Benchmark: KDE has an advantage in terms of the estimation accuracy with a smaller number of samples.

Number of samples	Criteria	MAE	Standard Deviation
10	KDE	2.38	1.79
	Median	3.03	2.37
20	KDE	2.04	1.58
	Median	2.80	2.34

- CBR-ACE consistently outperforms the variance-based method in all evaluated scenarios.
- Estimation accuracy varies with the type of activity: motions with fewer degrees of freedom, such as stepping, tend to yield lower errors, whereas more complex motions like twisting result in higher errors.
- The performance of CBR-ACE remains largely stable across different subject locations and device types.

Quantitatively, CBR-ACE achieves a minimum MAE of 0.97, with maximum MAEs of 2.09 and 2.63 at locations 1 and 2, respectively. In contrast, the variance-based baseline attains a best-case MAE of 1.94, while its worst-case errors increase to 4.56 and 4.44 at the two locations.

Micro-benchmark: Effect of KDE with Small Number of Samples

When the number of available subcarriers is sufficiently large, the probability density estimated via KDE closely approximates the underlying discrete distribution, and comparable results can be obtained by taking either the median of the raw discrete samples or that of the KDE-based distribution. In contrast, when only a small number of subcarriers is available, KDE is expected to be more advantageous for identifying an optimal estimate, as is the case for narrower channel bandwidths such as 20 MHz or 40 MHz.

To examine the effectiveness of KDE under limited-sample conditions, we conducted an additional experiment using a reduced number of subcarriers. Specifically, we randomly selected 300 samples from the experimental dataset, and for each sample, randomly drew 10 or 20 subcarriers from the total of 234 subcarriers available in the 80 MHz channel. The density is then estimated using ACE. Table 3.1 reports a comparison between KDE-based estimation and a simple median-based approach under these constrained settings. The results show that KDE consistently achieves lower MAE and smaller standard deviation than the median-based method, demonstrating its robustness when only a limited number of subcarriers is available.

Effect of k-th Percentile Point on Estimation Accuracy

In the proposed ACE framework, the activity count is determined by selecting the 50th percentile of the probability density estimated via KDE. This section examines how the choice of percentile, specifically the 40th, 50th, and 60th percentiles, influences estimation performance. For this analysis, we compare the MAE and the associated standard deviation across all activities and locations for each device.

Table 3.2: MAE performance using the 40th, 50th, and 60th percentile points of the estimated density.

Device	**k-th point**	**MAE**	**Standard Deviation**
MacBook Pro	40%	1.99	3.09
	50%	**1.95**	2.16
	60%	2.36	2.39
iPad	40%	2.43	3.89
	50%	1.78	2.18
	60%	**1.76**	1.72
Galaxy Note	40%	2.09	2.98
	50%	**1.58**	1.62
	60%	1.81	1.99

Table 3.3: Average CBRs acquisition per second for each device.

Device	**MacBook Pro**	**iPad**	**Galaxy Note**
Place 1	10.03	11.00	9.73
Place 2	10.13	8.03	9.60

Table 3.2 summarizes how different percentile selections affect both the MAE and its standard deviation, where the best MAE for each device is highlighted in bold. Overall, the 50th percentile yields the lowest MAE for all devices except the iPad.

We attribute this exception to differences in the CBR sampling rate. As reported in Table 3.3, the iPad exhibits a lower average CBR acquisition rate at place 2 compared with the other devices. Under such conditions, CBR-ACE performs temporal interpolation to obtain uniformly spaced samples. This interpolation tends to underestimate the true repetition count, which in turn degrades accuracy when the median (50th percentile) of the estimated density is used, leading to a higher MAE for the iPad in this case.

3.1.4 Conclusion and Contribution

This section introduces a new wireless sensing framework, termed CBR-ACE, designed for estimating the number of repeated human activities. To our knowledge, CBR-ACE is the first approach that exploits standard-compliant CBRs for activity count estimation, showing that even quantized information in the eigen domain preserves enough temporal variation to support reliable recognition. We also built a practical experimental platform based on off-the-shelf hardware, such as Raspberry Pi and iPhone devices, and verified the performance of CBR-ACE through real-world wireless experiments.

From a thesis-wide perspective, CBR-ACE addressed the methodological challenge outlined in Chapter 1 by enabling activity sensing on IEEE 802.11ac/ax devices with device-agnostic CBRs. This established that standard-compliant wireless signals can serve as a viable foundation for sensing. Although designed as a lightweight proof of concept, it provided empirical evidence that CBRs can substitute for CSI in end-to-end sensing pipelines, and it has since been referenced as an early milestone in the emerging line of CBR-based sensing research [83, 85, 89, 90]. Through this demonstration, CBR-ACE offers a concrete early response to the methodological challenge, complementing the accessibility contribution of WiPiCap and reinforcing the thesis-wide effort to advance standard-compliant wireless sensing.

3.2 CSI2Image: Generative Modeling of Spatial Representations using CBR

This section introduces CSI2Image, which explores a cross-modal representation of wireless sensing by transforming CBRs into image-domain outputs. CSI2Image is motivated by the view that wireless observations encode the state of an entire space in a high-dimensional form, and investigates how consistently such spatial information can be mapped into a more interpretable modality. Using a GAN-based architecture with multiple training strategies and an object-detection-based evaluation protocol, CSI2Image examines the fidelity and stability of wireless-to-image mappings. Through this process, CSI2Image demonstrates that spatial characteristics implicitly embedded in wireless signals can be externally manifested as structured visual representations.

3.2.1 Introduction

This study examines the extent to which wireless signals can reflect information about the surrounding physical space. Wireless sensing has demonstrated the potential to capture various spatial properties through the deployment of access points, and prior work has shown that radio signals can be exploited to infer aspects of physical environments. Among these approaches, channel-based methods using Wi-Fi CSI have attracted attention due to their practicality. In contrast, this study

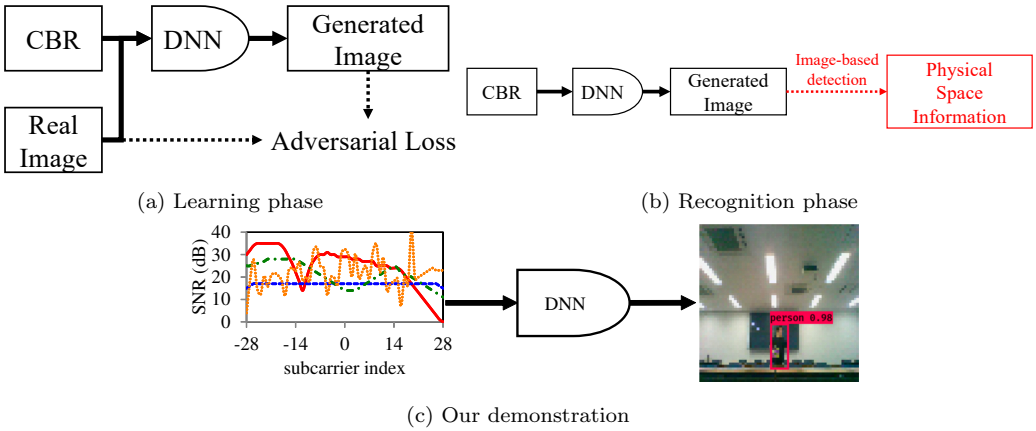


Figure 3.7: Application example of CSI2Image with object detection.

focuses on standard-compliant CBR, which is directly accessible from off-the-shelf Wi-Fi hardware without firmware modification.

This study investigates whether spatial characteristics implicitly embedded in CBR can be consistently mapped into an alternative, more interpretable representation. Specifically, we explore the reconstruction of image-domain outputs from standard-compliant CBR as a means to probe the representational capacity of wireless signals with respect to physical space. Images serve as a human-interpretable modality with high perceptual resolution, and their use further enables the application of mature computer vision techniques, such as object detection, to evaluate the consistency of the wireless-to-image mapping without manual annotation.

Figure 3.7 illustrates a representative application of CSI-to-image translation, namely the automatic construction of wireless sensing models. The overall pipeline is divided into two stages: a training stage and an inference stage. During training (Figure 3.7((a))), CSI measurements and visual observations of the target environment are collected in parallel and used to train a deep neural network. Once trained, the model reconstructs an image from newly captured CSI, and spatial information is subsequently inferred from the reconstructed image, as depicted in Figure 3.7((b)). A concrete example of this automatic wireless sensing model generation process is presented in Figure 3.7((c)) and evaluated in Section 3.2.4.

We introduce **CSI2Image**, a wireless sensing framework that maps radio measurements to image representations of the surrounding space using deep neural networks. To the best of our knowledge, this work is the first to realize channel-to-image translation through a GAN-based formulation. Several prior efforts have explored CSI-to-image reconstruction without GANs [63, 191]. Wi2Vi [63], for instance, targets scenarios in which security cameras become unavailable due to failures or attacks. In normal operation, Wi2Vi extracts background imagery, identifies human presence via frame differencing, and associates these cues with CSI during training. Under adverse conditions, it synthesizes images by overlaying detected users onto the learned background.

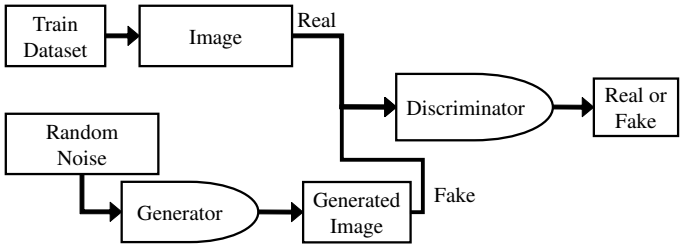


Figure 3.8: DCGAN architecture.

Another approach [187] learns a mapping from CSI to pose images derived from skeleton representations. These methods rely on application-specific priors, such as background images or predefined human skeleton models. In contrast, our approach aims at a more general CBR-to-image translation using GANs, without assuming task-specific auxiliary information.

The main contributions of this work are as follows:

- This work demonstrates, for the first time, that **standard-compliant CBR** can be used to reconstruct human-understandable spatial representations through generative modeling. By translating radio measurements into visual scenes, CSI2Image conceptually extends the wireless sensing framework from signal analysis to spatial reproduction, showing what kinds of spatial information can be inferred from standard Wi-Fi transmissions.

- CSI2Image proposes a **GAN framework** for CBR-to-image translation and investigates three learning paradigms to stabilize the generative process. The framework is further complemented by a quantitative evaluation pipeline using image-based object detection, providing an interpretable and reproducible way to assess spatial reconstruction quality.

- The proposed approach is empirically validated using **off-the-shelf IEEE 802.11ac devices**, confirming that even quantized CBR signals contain sufficient spatial cues for meaningful scene reconstruction. Furthermore, two proof-of-concept applications, object classification and device-free user localization, demonstrate the versatility of this approach and its potential to serve as a foundation for downstream wireless sensing tasks.

3.2.2 Preliminary: Generative adversarial networks

Generative Adversarial Networks (GANs) [192] are a class of generative models designed to synthesize data that statistically resemble the training distribution. They consist of two components trained in opposition: a generator (G), which produces synthetic samples, and a discriminator (D), which evaluates whether an input sample is real or generated. While the discriminator learns to separate authentic data from generated outputs, the generator is optimized to create samples that are indistinguishable from real ones. This adversarial interaction is typically expressed as a

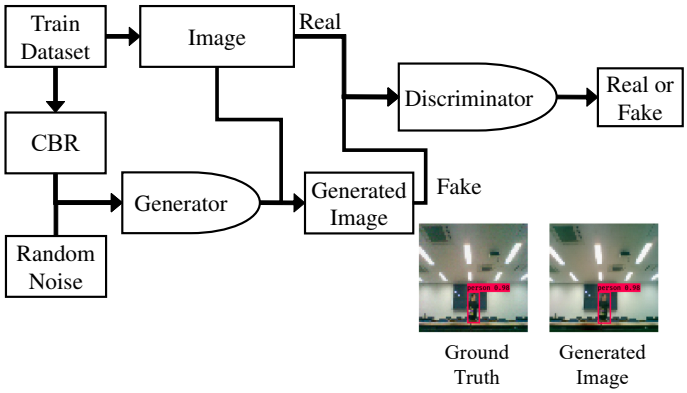


Figure 3.9: Overview of CSI2Image architecture.

min–max optimization problem, where G and D are alternately updated. Such a training paradigm has proven highly effective for producing realistic data and has been adopted across a wide range of applications [193, 194].

Among the numerous GAN architectures, the Deep Convolutional GAN (DCGAN) [195] is especially pertinent to this work. DCGAN improves training stability by replacing fully connected layers with convolutional structures and by incorporating normalization strategies. As illustrated in Figure 3.8, the discriminator is trained to identify genuine images from the dataset and to reject images generated from random latent vectors. Simultaneously, the generator learns to transform random noise into image-like outputs that increasingly resemble real samples. Through repeated adversarial updates, the generator gradually acquires the ability to produce visually plausible images. In this study, DCGAN is adopted as the base architecture for constructing a generative framework that maps standard-compliant wireless observations to spatial image representations.

3.2.3 CSI2Image

Figure 3.9 illustrates the overall architecture of the proposed CSI2Image framework. The system consists of three main components: the training dataset, a generator network, and a discriminator network. Within this framework, we introduce three distinct strategies for training the generator: generator-only training, GAN-based training, and a hybrid training scheme. In the generator-only setting, the discriminator is omitted, as detailed in later of this section.

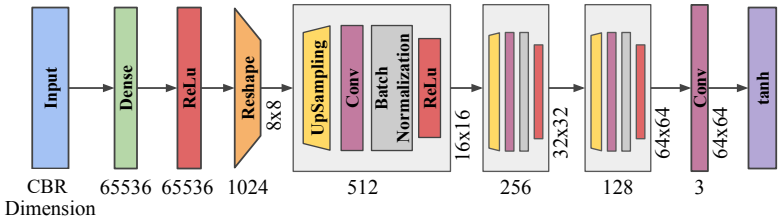


Figure 3.10: The layer structure of the generator network.

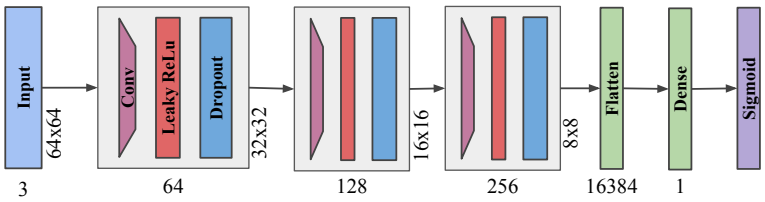


Figure 3.11: The layer structure of the discriminator network.

Training data

The training dataset for CSI2Image is constructed from images and CBRs acquired in a synchronized manner. We use full-color images with a resolution of $64 \times 64 pixels together with the corresponding CBR measuren$

Generator network

Figure 3.10 illustrates the architecture of the generator network. The input CBR is first fed into a fully connected layer with 65,536 units followed by a ReLU activation, after which the output is reshaped into an $8 \times 8 \times 1024$ feature tensor. The dimensionality of the CBR input varies depending on the channel bandwidth. This tensor is then progressively upsampled through three repeated blocks, each consisting of an upsampling operation, a 3×3 convolution, batch normalization, and a ReLU activation. At the final stage, a 3×3 convolution layer with a tanh activation function produces a $64 \times 64 \times 3$ output image. The generator is trained using the Adam optimizer with a learning rate of 0.0002 and a momentum parameter of 0.5, and its objective function is defined by the mean squared error (MSE).

Discriminator network

Figure 3.11 presents the architecture of the discriminator. It takes a 64×64 RGB image as input. The image is processed through four successive blocks, each composed of a 3×3 convolution with stride 2, followed by batch normalization, a LeakyReLU activation ($\alpha = 0.2$), and a dropout layer with a rate of 0.25. The resulting feature maps are flattened and passed through a sigmoid activation to produce a scalar output between 0 and 1, representing the probability of the input being real. The discriminator is optimized using Adam with the same hyperparameter settings as the generator, and it is trained with a binary cross-entropy loss function.

Learning phase

This study explores three training strategies: generator-only learning, GAN-only learning, and a hybrid approach. In generator-only learning, the model focuses on learning the mapping between CBR inputs and corresponding images. GAN-only learning employs both a generator and a discriminator within an adversarial framework. Hybrid learning integrates these two strategies to

leverage their respective advantages.

Figure 3.12 and Algorithm 1 illustrate the architecture and pseudocode of the **generator-only learning** scheme. In this setting, a convolutional neural network–based generator is trained using paired CBR measurements and simultaneously captured images. Because this approach directly fits the observed CBR–image pairs, the learned generator may exhibit limited generalization when presented with previously unseen CBR inputs.

Algorithm 1 Generator-only learning

```
1: N ⇐ number of training iterations
2: for i = 1 to N do
3:     csi_list ⇐ batch of CSI
4:     real_images ⇐ batch of real images
5:     G.csi2image.train(csi_list, real_images)
6: end for
```

Figure 3.13 and Algorithm 2 present the architecture and pseudocode of the **GAN-only learning** approach. In this scheme, the discriminator is trained to assess whether the generated output resembles a real image, which encourages the generator to produce visually sharper and more realistic results compared with generator-only learning. However, because the discriminator evaluates realism rather than correspondence, the generated image is not guaranteed to faithfully reflect the input CBRs. As a consequence, while overall image quality may improve, fine-grained structures that should be specifically induced by the CBR input may not be accurately preserved.

Algorithm 2 GAN-only learning

```
1: N ⇐ number of training iterations
2: for i = 1 to N do
3:     csi_list ⇐ batch of CSI
4:     real_images ⇐ batch of real images
5:     D.train(real_images, REAL)
6:     D.train(G.generate(csi_list), FAKE)
7:     G.generality.train(csi_list, REAL)
8: end for
```

Hybrid learning combines the strengths of generator-only learning and **GAN-only learning** within a unified training procedure. Each training cycle sequentially performs CSI2Image learning (Figure 3.14(a)), image generation, discriminator learning (Figure 3.14(b)), and generality learning (Figure 3.14(c)). Algorithm 3 outlines this process, where N denotes the total number of training iterations and K specifies the interval at which generality learning is applied. Specifically, Lines 4–6 of Algorithm 3 correspond to CSI2Image learning, in which the generator is trained using paired CBR measurements and their associated images, as illustrated in Figure 3.14(a). Lines 7–8 describe discriminator learning: the discriminator is first trained to classify real training images as genuine, and then to label images generated from random noise as fake, following Figure 3.14(b).

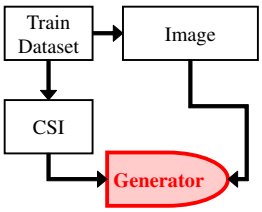


Figure 3.12: Generator-only learning.

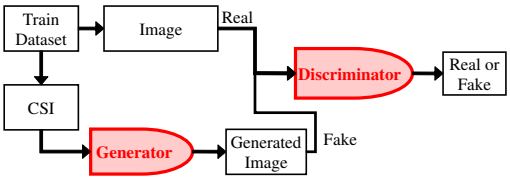
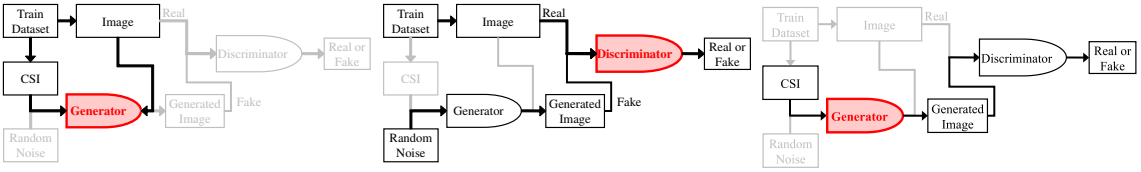


Figure 3.13: GAN-only learning.



(a) CSI2Image learning (b) Discriminator learning (c) Generality learning

Figure 3.14: Stages in hybrid learning.

Lines 9–11 implement generality learning, where the generator is periodically updated every K epochs by feeding CBR inputs and encouraging the discriminator to regard the generated outputs as real, as shown in Figure 3.14(c). A larger value of K tends to enhance generalization but may weaken the influence of CSI information, whereas a smaller K preserves CSI consistency at the cost of reduced visual quality due to limited generalization.

Algorithm 3 Hybrid Learning

```
1: N ⇐ number of training iterations
2: K ⇐ interval of generality training
3: for i = 1 to N do
4:     csi_list ⇐ batch of CSI
5:     real_images ⇐ batch of real images
6:     G.csi2image.train(csi_list, real_images)
7:     D.train(real_images, REAL)
8:     D.train(G.generate(random_noise), FAKE)
9:     if i mod K == 0 then
10:        G.generality.train(csi_list, REAL)
11:    end if
12: end for
```

Image generation phase

During the image generation phase, the trained generator takes measured CBRs as input and produces a 64×64 pixel image that reflects the corresponding scene. This stage serves as the inference phase of the framework: rather than optimizing against real samples, the generator directly converts wireless observations into visual outputs. The generated images preserve coarse spatial cues, including object arrangement and contrast boundaries, thereby translating abstract radio-domain measurements into a form that is easier for humans to interpret. Beyond visualization, these im-

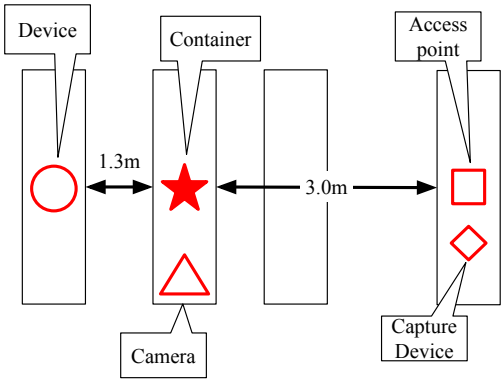


Figure 3.15: Experimental environment for object classification.

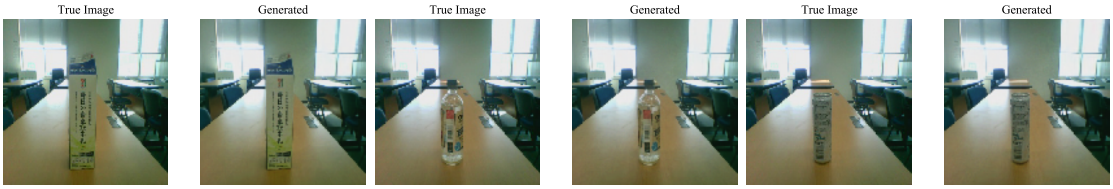
ages can also function as intermediate representations that facilitate downstream tasks such as localization and object-related analysis.

3.2.4 Evaluation

We evaluate the CSI2Image using object classification and device-free user localization tasks as examples of wireless sensing applications.

Experiment 1: Object classification

Evaluation settings: To examine the applicability of CSI2Image to object classification, we conduct experiments using three everyday containers: a milk carton (Milk), a plastic bottle (Bottle), and a beverage can (Can), illustrated in Figure 3.16(a)–(c). The milk carton is a paper-based, rectangular container with a volume of 1000 ml filled with milk. The plastic bottle is cylindrical, made of PET, has a volume of 500 ml, and contains water. The beverage can is an aluminum cylinder with a capacity of 355 ml and is filled with a carbonated drink. The experimental setup for object classification is shown in Figure 3.15, and an example snapshot of the environment is provided in Figure 3.19. All containers are placed between a Wi-Fi access point and a device. No other moving objects are present in the room, and both the container positions and orientations are kept fixed throughout the experiments. The setup consisted of an AP, a camera, a computer, and a capture device used as a CBR sniffer. Specifically, the AP is a TP-Link Archer C6 equipped with two transmit antennas, the camera is a Panasonic CF-SX1GEPDR with a resolution of 1280×720 pixels, the receiver is a MacBook Pro (13-inch, 2020) with two receive antennas, and the sniffer is a Raspberry Pi 3B+ running Raspbian 5.4.79. Experiments are carried out with an 80 MHz channel bandwidth, corresponding to 234 subcarriers. A total of 3506 images are used for training and 500 images for testing. The model is trained using the proposed hybrid learning scheme. Training is performed on an Amazon EC2 instance equipped with an Intel Xeon E5-2686 CPU operating at 2.3 GHz with 18 cores, 64 GB of DDR-SDRAM, a Tesla K80 GPU with 4992 cores running at



(a) Milk package (b) Plastic bottle (c) Drink can

Figure 3.16: Experiment 1: Examples of successful object classification.

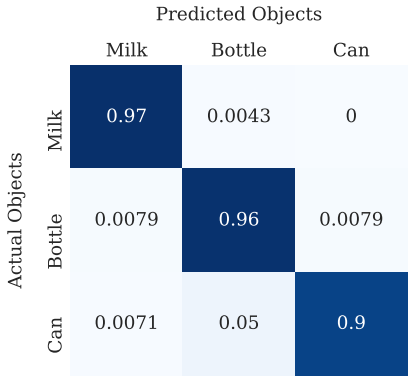


Figure 3.17: Experiment 1: Confusion matrix for object classification based on the image reproduction.

560 MHz, and solid-state storage. The model is trained for 15,000 epochs with a batch size of 16, and the hybrid learning parameter K is set to 8.

Evaluation results: Figure 3.16 presents representative examples of images synthesized by CSI2Image. The method successfully reconstructs visual representations of all 3 container types. The corresponding confusion matrix for the generated images is shown in Figure 3.17. CSI2Image attains classification accuracies of 97 %, 96 %, and 90 % for a milk package, plastic bottle, and drink can, respectively. High accuracy is observed for both the milk package and the plastic bottle. Occasionally, the drink can is misclassified as a plastic bottle, although such cases occur infrequently (around 5 %). This confusion is likely attributable to the similar cylindrical shapes shared by the plastic bottle and the drink can.

Experiment 2: Device-free user localization

As discussed in Section 3.2.4, CSI2Image demonstrates its applicability to object classification. To further illustrate the flexibility of CSI2Image and to assess its behavior in a more quantitative manner, we next examine its use in device-free user localization, progressively increasing task difficulty. Since CSI-to-image translation remains a relatively unexplored research direction, standardized quantitative evaluation protocols are not yet available. To address this gap, we introduce

a new evaluation approach that leverages device-free user localization in conjunction with YOLO, an image-based object detection framework.

Evaluation settings: Figure 3.18 illustrates the experimental setup used for device-free localization, while a representative scene snapshot is shown in Figure 3.19. In these experiments, a Panasonic EA-7HW04AP1ES access point equipped with three transmit antennas is employed. Image acquisition is performed using a Panasonic CF-SX1GEPDR camera with a resolution of 1280×720 pixels. The receiving device is a MacBook Pro (13-inch, 2017) with two receive antennas, and CBR packet capture is handled by a Panasonic CF-B11QW-HBR running CentOS 7.7-1908. Measurements are conducted over a 20 MHz channel, corresponding to 52 subcarriers. The CSI2Image model is trained on a Dell Alienware 13 R3 system featuring an Intel Core i7-7700HQ processor (four cores, 3.8 GHz), 16 GB of DDR-SDRAM, and an NVIDIA GeForce GTX 1060 Mobile GPU operating at 1404 MHz with 1280 CUDA cores, along with solid-state storage. For object detection, we employ YOLO v3 [196], using a confidence threshold of 0.3 to determine valid detections.

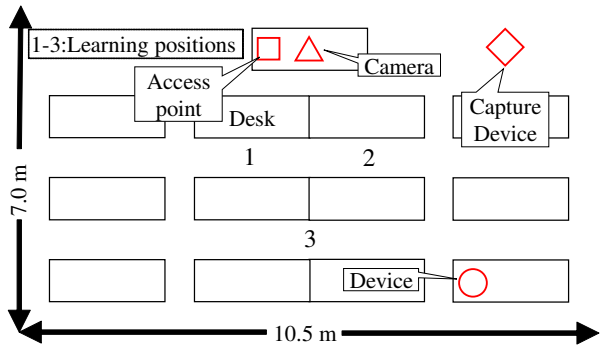


Figure 3.18: Experimental setting for human localization.



Figure 3.19: Snapshot of the experiment environment for human localization.

Single user: In this evaluation, we compare localization performance across three training strategies: generator-only learning (gonly), GAN-only learning (gan), and hybrid learning (hybrid). The following metrics are used for assessment:

- **Object detection success rate:** A higher rate indicates that the synthesized images provide sufficient visual fidelity for reliable object detection.
- **Average confidence score for successful detections:** This score reflects the detection model's certainty when a target is correctly identified.
- **Structural similarity (SSIM)** [197]: SSIM is a widely used image-quality metric that quantifies structural consistency between generated and reference images.



(a) Ground truth (b) Generator-only (c) GAN-only (d) Hybrid

Figure 3.20: Experiment 2: Successful examples of localization against one user.



(a) Ground truth (b) Generator-only (c) GAN-only (d) Hybrid

Figure 3.21: Experiment 2: Failed examples of localization against one user.

- **Localization accuracy:** Defined as the proportion of samples in which the target position is correctly identified based on object detection results.

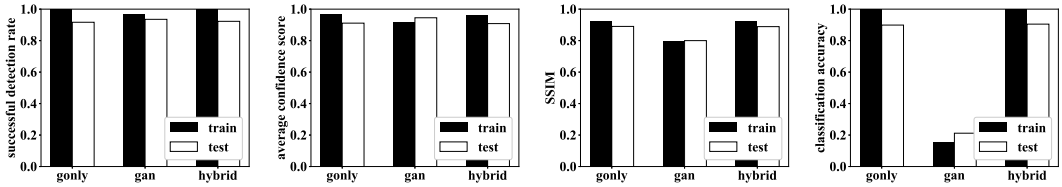
The SSIM metric is defined as shown in Equation (3.2):

$$\text{SSIM} = \frac{(2\mu_x\mu_y + C_1)(2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)}. \tag{3.2}$$

Here, x and y denote vectors composed of pixel intensities from the reference image and the reconstructed image, respectively. The terms μ_x and μ_y correspond to the mean pixel values of x and y, while σ_x and σ_y represent their standard deviations. The covariance between the two images is given by σ_{xy}. The constants C_1 and C_2 are defined as $C_1 = (255K_1)^2$ and $C_2 = (255K_2)^2$, where $K_1 = 0.01$ and $K_2 = 0.03$, following the settings in [197]. SSIM ranges between 0 and 1, with higher values indicating greater similarity between the reconstructed and original images.

The experiment is conducted with a single subject standing at one of the three locations (positions 1–3) shown in Figure 3.18. Accordingly, the generated images fall into three distinct spatial patterns, each corresponding to the subject being located at position 1, 2, or 3. For evaluation, 180 images are used for training and 184 images are reserved for testing. The model is trained for 32,000 epochs with a batch size of 32, and in the hybrid learning scheme the parameter K is set to 8.

Figure 3.20 presents a representative example of successful single-user localization. In each image, the red bounding box indicates the detection result produced by YOLO. When the detected



(a) Successful detection rate (b) Average confidence score (c) SSIM (d) Accuracy

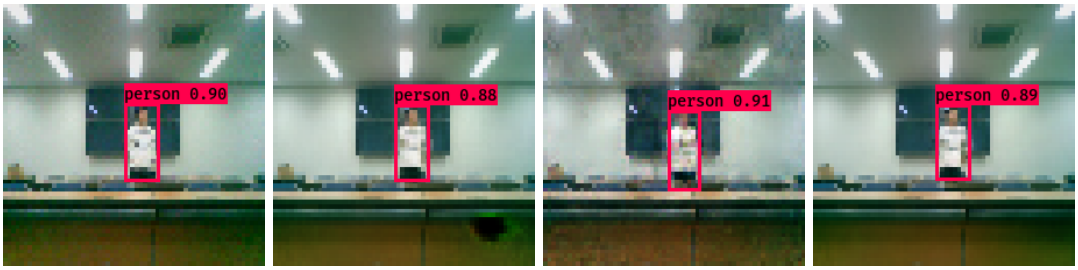
Figure 3.22: Experiment 2: Quantitative evaluation of a single-user localization

person appears on the right-hand side of the image, as in Figure 3.20(a), the localization is regarded as correct. Both the generator-only and hybrid learning approaches, shown in Figures 3.20(b) and 3.20(d), correctly estimate the user location and produce a plausible human silhouette. In contrast, although GAN-only learning (Figure 3.20(c)) correctly identifies the user location, it also generates an additional shadow-like artifact near the center, corresponding to an incorrect location.

Figure 3.21 illustrates a representative failure case in single-user position detection. When a person is detected on the right-hand side of the image, as in Figure 3.21(a), the localization is considered correct. However, as shown in Figures 3.21(b) through 3.21(d), the generated images contain faint, ghost-like artifacts appearing near the center and right regions. In contrast, the result obtained with GAN-only learning in Figure 3.21(c) exhibits a visually cleaner image compared with generator-only and hybrid learning, despite producing an incorrect position estimate. We attribute this behavior to the stronger generalization capability of GAN-only learning, which enables the generation of realistic human appearances. At the same time, this increased generalization weakens the coupling between CSI and spatial position, leading to cases where the reconstructed person is placed at an incorrect location.

Figure 3.22(a) reports the human detection success rates. The black bars correspond to performance on the training set, while the white bars indicate results on the test set. Among the three learning strategies, GAN-only learning attains the highest detection success rate. On the test data, the success rates are approximately 92.7% for generator-only learning, 93.5% for GAN-only learning, and 92.3% for hybrid learning.

Figure 3.22(b) illustrates the average confidence scores for cases in which object detection is successful. Consistent with the detection success rates discussed above, GAN-only learning yields the highest confidence. On the test set, the average confidence scores are approximately 91.1% for generator-only learning, 94.5% for GAN-only learning, and 90.8% for hybrid learning. Overall, GAN-only learning outperforms the other methods in both detection success rate and confidence score. In particular, it produces images that lead to higher confidence scores than those observed on the training data, likely due to the GAN's strong capability for synthesizing visually realistic images. At the same time, this strong generalization tendency explains the relatively lower



(a) Ground truth (b) Generator-only (c) GAN-only (d) Hybrid

Figure 3.23: Experiment 2: Successful examples of localization against up to two users.

positioning accuracy of GAN-only learning, as discussed earlier.

Figure 3.22(c) reports the SSIM scores for the compared methods. Unlike the trends observed for detection success rate and confidence score, GAN-only learning performs the worst in terms of SSIM. On the test set, the SSIM values are approximately 0.890 for generator-only learning, 0.800 for GAN-only learning, and 0.889 for hybrid learning.

To further analyze the degraded SSIM of GAN-only learning, we examined the position detection accuracy. The results indicate that GAN-only learning also exhibits the lowest localization accuracy among the three methods. Specifically, the test-set accuracies are about 89.9% for generator-only learning, 21.2% for GAN-only learning, and 90.5% for hybrid learning.

Consequently, while GAN-only learning achieves the highest detection success rate and average confidence score, its SSIM remains low due to frequent misplacement of the user in the generated images. Notably, GAN-only learning shows poor position detection accuracy even on the training data. This behavior arises because the discriminator in GAN-only learning focuses solely on judging the realism of the generated images, rather than enforcing consistency between the CBR input and the corresponding spatial position, as illustrated in Figure 3.22(d).

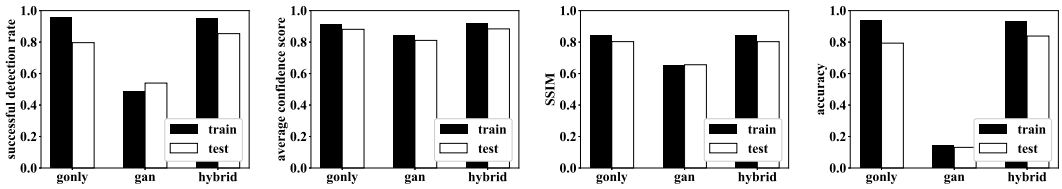
Single or two users: To assess more challenging scenarios, we further evaluated position detection performance in settings involving up to two users. Concretely, six classification cases are considered based on the occupancy of positions 1 to 3 in Figure 3.18: a single person at position 1, 2, or 3, and two persons simultaneously at positions (1,2), (1,3), or (2,3). For this evaluation, 720 images are used for training and 330 images for testing. All other experimental settings followed those described in Section 3.2.4.

Figure 3.23 presents representative examples of successful position detection in scenarios with either one or two users. When a person is correctly detected near the center of the image, as illustrated in Figure 3.23(a), the localization result is regarded as accurate. As shown in Figures 3.23(b) and 3.23(d), both generator-only and hybrid learning recover the correct positions, while also producing clear and well-defined human silhouettes. In contrast, Figure 3.23(c) indicates



(a) Ground truth (b) Generator-only (c) GAN-only (d) Hybrid

Figure 3.24: Experiment 2: Failed examples of localization against up to two users.



(a) Successful detection rate (b) Average confidence score (c) SSIM (d) Accuracy

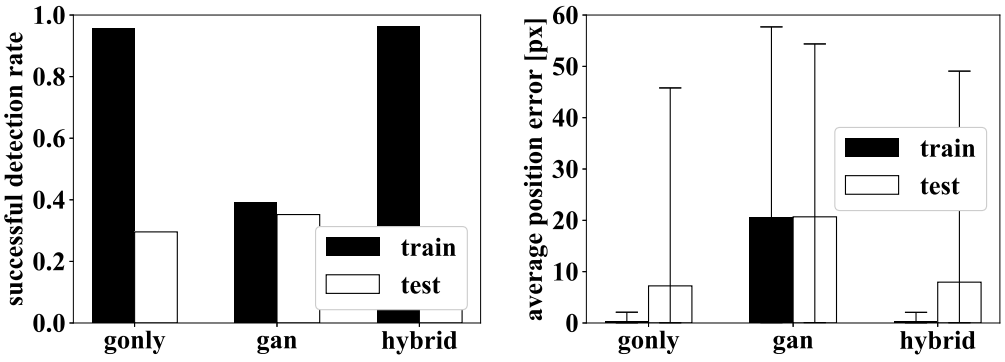
Figure 3.25: Experiment 2: Quantitative evaluation of the position detection of one or two users.

that GAN-only learning identifies the correct position but simultaneously generates an additional shadow-like human figure at an incorrect location. Consistent with the single-user localization results, this behavior suggests that GAN-only learning favors position-agnostic generalization, which can lead to the appearance of spurious users even when the true position is correctly reconstructed.

Figure 3.24 shows examples of unsuccessful position detection under one- or two-user conditions. If a person is detected on the right side of the image, as in Figure 3.24(a), the localization is considered correct. However, in Figures 3.24(b) and 3.24(d), YOLO fails to detect a person on the right despite the presence of human-like visual patterns produced by generator-only and hybrid learning. Meanwhile, GAN-only learning, shown in Figure 3.24(c), generates human figures at incorrect central and left positions. In such cases, the degraded performance of generator-only and hybrid learning is likely due to insufficient positional cues in the input CSI, resulting in blurred or ambiguous reconstructions.

Figure 3.25(a) summarizes the detection success rates for the compared learning strategies. Among them, hybrid learning achieves the best performance: on the test set, the success rates are about 79.6% for generator-only learning, 54.0% for GAN-only learning, and 85.4% for hybrid learning. Notably, GAN-only learning exhibits a low detection rate even on the training data.

Figure 3.25(b) reports the corresponding average confidence scores. A similar trend is observed, where hybrid learning again yields the highest confidence. On the test data, the average confidence scores are approximately 88.1% for generator-only learning, 81.1% for GAN-only learning, and 88.4% for hybrid learning.



(a) Detection success rate

(b) Average position error

Figure 3.26: Experiment 2: Single-user continuous position estimation.

Figure 3.25(c) presents the SSIM results for each method. Consistent with the single-user case, GAN-only learning performs the worst in terms of structural similarity. For the test data, the SSIM values are 0.803 for generator-only learning, 0.656 for GAN-only learning, and 0.803 for hybrid learning.

Finally, Figure 3.25(d) compares the position detection accuracy. Using the test set, hybrid learning attains the highest accuracy, whereas GAN-only learning shows a substantial degradation. The accuracies are approximately 79.3% for generator-only learning, 13.1% for GAN-only learning, and 83.8% for hybrid learning.

Continuous device-free localization for a single user: To examine a more challenging scenario, we conduct experiments in which a single person moved continuously along an oval path connecting positions 1 to 3, as illustrated in Figure 3.18. For this evaluation, 515 images are used for training and 498 images for testing. All other experimental conditions follow those of Experiments 1 and 2. Since the qualitative observations are consistent with the position detection results discussed in Section 3.2.4, we focus here solely on quantitative analysis.

Figure 3.26(a) reports the detection success rates for the compared methods. Overall, the success rates are noticeably lower than those obtained in Sections 3.2.4 and 3.2.4, even though the models achieved high detection rates on the training data. Specifically, the training success rates are 95.7% for generator-only learning, 39.0% for GAN-only learning, and 96.3% for hybrid learning. In contrast, the corresponding test success rates drop to 29.6%, 35.2%, and 27.8%, respectively. This performance gap suggests that the available training data are insufficient for the increased complexity of the continuous motion setting.

Figure 3.26(b) depicts the pixel-wise distance between the left coordinate of the bounding box in the training images and that in the generated images. Smaller distances indicate more accurate tracking of the user's position by CSI2Image. This analysis considers only the generated images

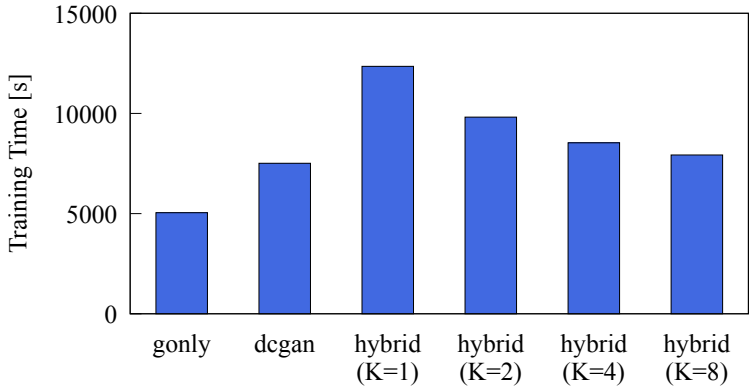


Figure 3.27: Training duration for device-free localization of one user (32000 epochs).

in which a user is successfully detected. The error bars represent the minimum and maximum observed values. The results indicate that GAN-only learning is unsuitable for continuous single-user localization. Although generator-only and hybrid learning outperform GAN-only learning, their performance still requires improvement, as evidenced by the large maximum deviations. On the test set, the maximum positional differences reached 46 px for generator-only learning, 54 px for GAN-only learning, and 49 px for hybrid learning.

3.2.5 Computational Costs

We next assess the computational overhead of CSI2Image by examining both the model training time and the runtime required for CSI-to-image conversion.

Computational Costs of Model Training

We measure and compare the training durations of several learning strategies: generator-only (gonly), DCGAN-based learning (dcgan), and hybrid learning with different generality intervals ($K = 1$, $K = 2$, $K = 4$, and $K = 8$). All methods are trained using the same dataset and the same number of training epochs as in Experiment 2, ensuring a fair comparison across configurations.

Figure 3.27 summarizes the training times of the evaluated methods: 5045.0 s for gonly, 7512.8 s for dcgan, and 12 350.1 s, 9815.8 s, 8535.0 s, and 7614.1 s for hybrid learning with $K = 1$, $K = 2$, $K = 4$, and $K = 8$, respectively. Among these approaches, gonly requires the least training time, whereas hybrid learning is the most time-consuming. Due to space limitations, detailed results are omitted; however, a similar trend in training time is consistently observed across other experimental settings.

Computational Costs of CSI-to-image Conversion

We evaluate the inference cost by measuring the time required to transform CSI into images using a pre-trained model. The training and test datasets, as well as the trained models, are identical to those used in the object classification experiments. Image generation is carried out on an iMac equipped with an Intel Core i5-8500 CPU (six cores, up to 4.1 GHz), 16 GB of DDR-SDRAM, a Radeon Pro 570X GPU (1792 cores at 1105 MHz), and solid-state storage. A total of 1169 test samples are processed by the trained model, resulting in a cumulative generation time of 246.65 s, which corresponds to an average of approximately 0.21 s per image.

3.2.6 Conclusion and Contribution

This chapter introduces CSI2Image, a generative framework that examines the expressive capacity of wireless sensing by translating CBR measurements into image representations. CSI2Image explores whether spatial information embedded in high-dimensional wireless observations can be consistently mapped to an interpretable representation. We study three learning schemes—generator-only, GAN-only, and hybrid learning—to trade off reconstruction fidelity and visual plausibility. Experimental results show that CSI2Image is capable of producing semantically coherent visual cues from CBR observations, enabling accurate object classification and device-free localization in static or single-user scenarios. Moreover, the hybrid approach provides improved robustness in multi-user settings, whereas the performance degradation observed under continuous motion underscores the difficulty of capturing temporal coherence when relying solely on frame-wise CSI inputs.

From a thesis-wide perspective, CSI2Image addresses a central methodological objective outlined in Chapter 1. By transforming standard-compliant CBRs into spatially meaningful visual forms, it demonstrates that Wi-Fi signals inherently contain rich geometric and environmental structure, and that this structure can be revealed by treating the signal as a spatial field rather than a task-specific input. This visualization-oriented approach provides both qualitative and quantitative evidence of the latent spatial information embedded in wireless measurements, offering an empirical route toward interpretable, data-driven sensing. This contribution complements the standard-compliant frameworks developed in Chapters 2.1 and 3.1, which focus on accessibility and practicality, and has since been referenced as an early step toward spatial reconstruction using Wi-Fi signals [57–62].

Chapter 4

Radar-based 3D Pose Estimation with Annotation Constraints

4.1 RAPTR

Building upon the data resources established in the previous chapters, including our MMVR, this chapter focuses on methodological advancement in radar-based indoor human perception, where the challenge lies in achieving detailed 3D understanding under realistic supervision constraints. We aim to develop a learning framework that bridges the gap between fine-grained 3D pose estimation and practical label availability. To this end, we propose RAPTR, a transformer-based framework for 3D human pose estimation under weak supervision. Unlike prior fully supervised approaches that depend on costly fine-grained 3D keypoint labels, RAPTR learns 3D body structures from only 2D keypoints and 3D bounding boxes (BBoxes), guided by structured losses and a two-stage decoding process that progressively refined pose estimations. This design demonstrates that accurate 3D pose recovery is achievable even with limited supervision, offering a scalable and label-efficient pathway toward practical radar-based perception.

4.1.1 Introduction

Accurate human perception is essential for indoor applications, including elderly monitoring, smart building management, and robotic navigation. Although vision sensors offer high spatial resolution, they raise privacy concerns and perform poorly under low light, occlusions, and hazardous conditions (fire or smoke). In contrast, radar provides penetration capability, robustness to adverse conditions, and low deployment cost, ideal for privacy-preserving indoor sensing [124,173,198–200].

By processing 4D radar tensors, RF-Pose 3D [143] demonstrated through-the-wall 3D pose estimation with a CNN, while HRRadarPose [140] employed an hourglass neural network HR-Net [201]. mRI [122] is a multi-modal 3D human pose estimation dataset that integrates mmWave radar, RGB-D cameras, and inertial sensors to facilitate research in human pose estimation and action detection. QRFPose [138] is a novel approach that adopts a DETR [184]-style query mech-

anism for end-to-end 3D regression using multi-view radar perceptions. Existing pipelines often rely on expensive fine-grained 3D keypoint labels [131], typically collected using non-portable 3D motion capture systems such as VICON, or using LiDAR, which can still suffer from occlusions and incomplete observations.

Collecting cheaper, lower-cost labels, such as fine-grained 2D keypoints in the image plane and/or coarse-grained 3D BBoxes, is considerably easier and more scalable particularly in complex indoor settings (e.g., cluttered, occlusion, multi-person), compared with acquiring dense 3D keypoints labels. Examples include RF-Pose [202], HuPR [141], and, more recently, MMVR [148] datasets. To the best of our knowledge, the use of 2D keypoints and 3D BBoxes, as a substitute for costly 3D keypoints, for radar-based 3D human pose estimation has not been systematically investigated in the literature before.

To address this gap, we propose **RAPTR** (RAdar Pose esTimation using tRansformer) in Figure 4.1, a radar-based pipeline designed to take multi-view radar heatmaps as inputs and estimate 3D human poses under weak supervision using only 3D BBox and 2D keypoint labels. RAPTR builds on the two-stage (pose and joint) decoder architecture of the state-of-the-art RGB-based 2D pose estimation PETR framework [203] and introduces a structural loss function that is designed to utilize weak supervision labels to mitigate the depth ambiguity. RAPTR also lifts the 2D deformable attention in PETR to a pseudo-3D deformable attention, wherein reference points (dots in Figure 4.1) and offsets (arrows in Figure 4.1) are proposed in the 3D radar coordinate system and projected onto multiple radar views (dots on the radar heatmaps in Figure 4.1) to eliminate redundant per-view offset estimation and offer better scalability as the number of radar views increases. Our model outperforms a list of radar-based 3D pose baselines over two indoor radar datasets: HIBER [131] and MMVR [148]. The main contributions of this work are:

- To the best of our knowledge, RAPTR is the first radar-based 3D human pose estimation framework to explicitly utilize low-cost weak supervision in the form of 3D BBoxes and 2D keypoints, rather than relying on fine-grained 3D keypoint labels.
- We introduce a structured loss function that tightly couples the two-stage decoder architecture to enable 3D pose estimation under weak supervision. Specifically, we design a 3D Template Loss, which utilizes the 3D BBox labels at the pose decoder, and a combined 3D Gravity and 2D Keypoint Loss at the pose decoder, allowing RAPTR to effectively learn geometrically consistent 3D poses from weak supervision.
- We further introduce a pseudo-3D deformable attention mechanism to bridge the 3D spatial domain and 2D radar views, enabling scalable view association while preserving pose estimation performance.

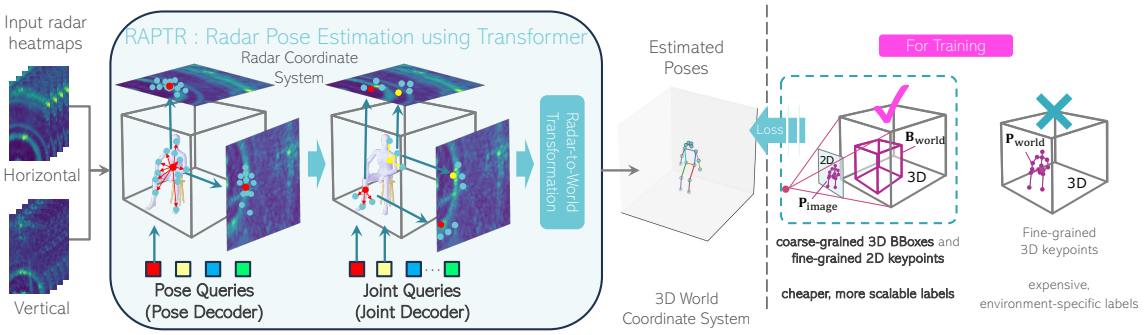


Figure 4.1: RAPTR takes multi-view radar heatmaps as inputs and performs a novel Pseudo-3D deformable attention between (pose and joint) queries and multi-view radar features in a two-stage decoder to estimate 3D human poses in a 3D coordinate system. Rather than relying on expensive, environment-specific fine-grained 3D keypoint labels, RAPTR makes use of cheaper, more scalable labels such as coarse-grained 3D BBoxes and fine-grained 2D keypoints to train the model.

4.1.2 Preliminary

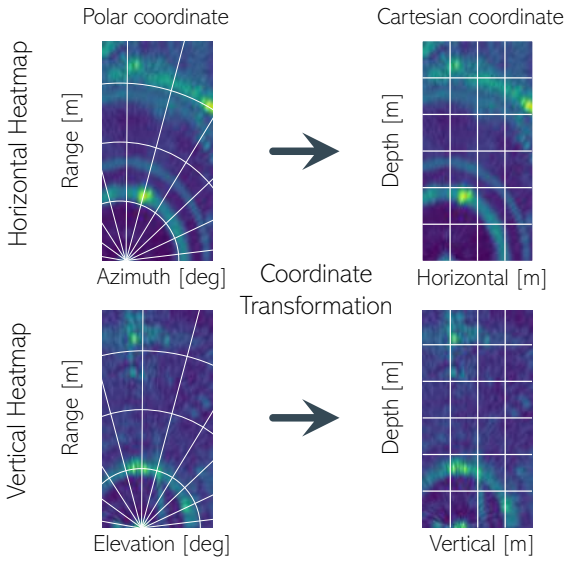


Figure 4.2: Multi-view radar heatmaps.

Multi-View Radar Heatmaps: As shown in Figure 4.2, two synchronized radar arrays (horizontal and vertical) collect reflected pulses that form a 3D data cube per array (ADC samples × pulses × elements). A 3D FFT converts each cube into a range–Doppler–angle spectrum, whose angle dimension is azimuth for the horizontal array and elevation for the vertical array. After Doppler-axis integration to boost SNR (signal-to-noise ratio), we obtain two polar 2D heatmaps (range-azimuth and range-elevation). These are mapped to the Cartesian space: $\mathbf{Y}_{\mathtt{hor}}(t) \in \mathbb{R}^{W \times D}$ for horizontal-depth and $\mathbf{Y}_{\mathtt{ver}}(t) \in \mathbb{R}^{H \times D}$ for vertical-depth at frame t. The temporal context is captured by stacking T consecutive frames, giving $\mathbf{Y}_{\mathtt{hor}} \in \mathbb{R}^{T \times W \times D}$ and $\mathbf{Y}_{\mathtt{ver}} \in \mathbb{R}^{T \times H \times D}$.

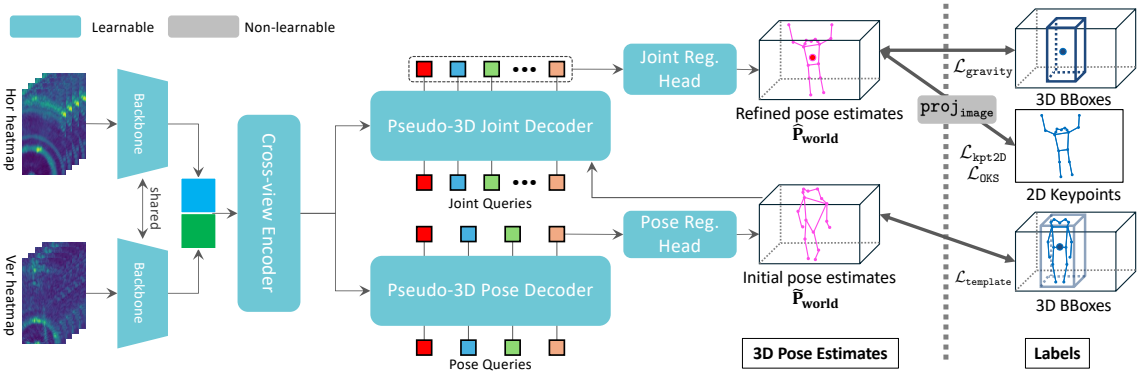


Figure 4.3: The RAPTR architecture consists of: 1) **Cross-view Encoder** that extracts multi-scale radar features; 2) **Pseudo-3D Pose Decoder** that enhances pose queries via a pseudo-3D deformable attention and predicts initial 3D poses; and 3) **Pseudo-3D Joint Decoder** that further refines joint queries and outputs final 3D poses. In terms of **loss function**, RAPTR leverages 3D BBox and 2D keypoint labels through coarse-grained 3D loss (gravity and template) and 2D keypoint loss.

Problem Formulation: The 3D pose estimation task takes T consecutive radar frames, $\mathbf{Y}_{\texttt{hor}}$ and $\mathbf{Y}_{\texttt{ver}}$, as input and estimates poses $\hat{\mathbf{P}}_{\texttt{world}}$ in the 3D world coordinate system,

$$\hat{\mathbf{P}}_{\texttt{world}} = \mathcal{T}_{\texttt{r2w}}(\hat{\mathbf{P}}_{\texttt{radar}}) = \mathcal{T}_{\texttt{r2w}}(f(\mathbf{Y}_{\texttt{hor}}, \mathbf{Y}_{\texttt{ver}})), \tag{4.1}$$

where f represents the 3D pose estimation pipeline in the 3D radar coordinate system, and $\mathcal{T}_{\texttt{r2w}}$ is a known radar-to-world coordinate transformation that converts the estimated 3D poses into the 3D world coordinate system. Rather than relying on costly, non-scalable fine-grained 3D keypoint labels $\mathbf{P}_{\texttt{world}}$, we consider cheaper, more scalable labels such as coarse-grained 3D BBoxes $\mathbf{B}_{\texttt{world}}$ and fine-grained 2D keypoints $\mathbf{P}_{\texttt{image}}$ for supervision, as shown in Figure 4.1.

4.1.3 RAPTR: Radar-based 3D Pose Estimation using Transformer

We present the RAPTR architecture in Figure 4.3, following a left-to-right order, and highlight radar-specific modifications.

Backbone

Given $\mathbf{Y}_{\texttt{hor}} \in \mathbb{R}^{T\times W\times D}$ and $\mathbf{Y}_{\texttt{ver}} \in \mathbb{R}^{T\times H\times D}$, a shared backbone network (e.g., ResNet [204]) generates separate multi-scale horizontal-view and vertical-view radar feature maps: $\mathbf{Z}_{\texttt{hor}} = \{\mathbf{Z}_{\texttt{hor},i}\}_{i=1}^{S} = \texttt{backbone}\,(\mathbf{Y}_{\texttt{hor}})$ and $\mathbf{Z}_{\texttt{ver}} = \{\mathbf{Z}_{\texttt{ver},i}\}_{i=1}^{S} = \texttt{backbone}\,(\mathbf{Y}_{\texttt{ver}})$, where the i-th scale feature maps $\mathbf{Z}_{\texttt{hor},i} \in \mathbb{R}^{W_i\times D_i\times d}$ and $\mathbf{Z}_{\texttt{ver},i} \in \mathbb{R}^{H_i\times D_i\times d}$ have a spatial dimension of $W_i \times D_i$ or $H_i \times D_i$ and a feature dimension of d, and S is the number of scales.

Cross-View Encoder

A cross-view encoder is a Transformer encoder with $L_{\texttt{enc}}$ layers that fuses the horizontal- and vertical-view radar features. Each layer runs a shared cross-attention twice: first with $\mathbf{Z}_{\texttt{hor}}$ as

key/value and $\mathbf{Z}_{\texttt{ver}}$ as query, then vice versa. This bidirectional exchange embeds complementary cues, while residual connections keep view-specific details, producing refined features $\mathbf{F}_{\texttt{enc}}^{(i)}$, $i = 1, \cdots, L_{\texttt{enc}}$,

$$\mathbf{F}_a^{(i)} = \mathbf{F}_a^{(i-1)} + \texttt{CrossAttn}(\mathbf{F}_a^{(i-1)}, \mathbf{F}_b^{(i-1)}), \quad (a, b) \in \{(\texttt{hor}, \texttt{ver}), (\texttt{ver}, \texttt{hor})\}, \tag{4.2}$$

where $\mathbf{F}_{\texttt{hor}}^{(0)} = \mathbf{Z}_{\texttt{hor}}$, $\mathbf{F}_{\texttt{ver}}^{(0)} = \mathbf{Z}_{\texttt{ver}}$, and $\texttt{CrossAttn}(\cdot, \cdot)$ denotes the deformable cross-attention [205] following [203] with fixed positional embeddings added beforehand for efficiency. After $L_{\texttt{enc}}$ iterations, the encoded features $\mathbf{F}_{\texttt{hor}}$ and $\mathbf{F}_{\texttt{ver}}$ are obtained at the output of the cross-view encoder.

Pseudo-3D Pose Decoder

A pose decoder associates N pose queries $\mathbf{Q}_{\texttt{pose}} \in \mathbb{R}^{N \times d}$ (embedding dimension d) with encoded radar features $(\mathbf{F}_{\texttt{hor}}, \mathbf{F}_{\texttt{ver}})$, where each query corresponds to a reference pose refined through pseudo-3D deformable attention over $L_{\texttt{pose}}$ layers. We define the l-th decoder layer as a function $\mathcal{D}_{\texttt{pose}}^{(l)}$ that updates both the pose queries and reference poses in the 3D radar space:

$$(\mathbf{Q}_{\texttt{pose}}^{(l)}, \tilde{\mathbf{P}}_{\texttt{radar}}^{(l)}) = \mathcal{D}_{\texttt{pose}}^{(l)}(\mathbf{Q}_{\texttt{pose}}^{(l-1)}, \mathbf{F}_{\texttt{hor}}, \mathbf{F}_{\texttt{ver}}, \tilde{\mathbf{P}}_{\texttt{radar}}^{(l-1)}), \tag{4.3}$$

where $\tilde{\mathbf{P}}_{\texttt{radar}}^{(0)}$ is initialized by passing $\mathbf{Q}_{\texttt{pose}}$ through an MLP. Reference poses are iteratively refined by applying predicted coordinate offsets $\Delta\tilde{\mathbf{P}}_{\texttt{radar}}^{(l)}$ in the normalized scale:

$$\tilde{\mathbf{P}}_{\texttt{radar}}^{(l)} \in \mathbb{R}^{N \times 3K} = \sigma(\sigma^{-1}(\tilde{\mathbf{P}}_{\texttt{radar}}^{(l-1)}) + \Delta\tilde{\mathbf{P}}_{\texttt{radar}}^{(l-1)}), \quad l = 1, \ldots, L_{\texttt{pose}}, \tag{4.4}$$

where σ and σ^{-1} denote the Sigmoid function and its inverse. The predicted offsets $\Delta\tilde{\mathbf{P}}_{\texttt{radar}}^{(l)} = H_{\texttt{pose}}(\mathbf{Q}_{\texttt{pose}}^{(l)})$ are obtained by passing pose queries at each layer to a shared regression head $H_{\texttt{pose}}$. We convert the initial pose estimates $\tilde{\mathbf{P}}_{\texttt{radar}} = \tilde{\mathbf{P}}_{\texttt{radar}}^{(L_{\texttt{pose}})}$ from the radar coordinate system to the world coordinate system via $\tilde{\mathbf{P}}_{\texttt{world}} = \mathcal{T}_{\texttt{r2w}}(\tilde{\mathbf{P}}_{\texttt{radar}})$, along with the corresponding confidence scores $\tilde{\mathbf{c}}$.

Pseudo-3D Joint Decoder

A following joint decoder associates K joint queries $\mathbf{Q}_{\texttt{joint}} \in \mathbb{R}^{K \times d}$ with encoded radar features $(\mathbf{F}_{\texttt{hor}}, \mathbf{F}_{\texttt{ver}})$, where each query corresponds to a single joint refined by pseudo-3D deformable attention over $L_{\texttt{joint}}$ layers. Here, K joint queries correspond to the same subject. We define the l-th decoder layer as a function $\mathcal{D}_{\texttt{joint}}^{(l)}$ that updates both the joint queries and corresponding joints:

$$(\mathbf{Q}_{\texttt{joint}}^{(l)}, \tilde{\mathbf{p}}_{i,\texttt{radar}}^{(l)}) = \mathcal{D}_{\texttt{joint}}^{(l)}(\mathbf{Q}_{\texttt{joint}}^{(l-1)}, \mathbf{F}_{\texttt{hor}}, \mathbf{F}_{\texttt{ver}}, \tilde{\mathbf{p}}_{i,\texttt{radar}}^{(l-1)}), \tag{4.5}$$

where $\tilde{\mathbf{p}}_{i,\texttt{radar}}^{(l)} \in \mathbb{R}^{K \times 3}, i = 1, \cdots, N$ is one specific pose in the N poses, and $\tilde{\mathbf{p}}_{i,\texttt{radar}}^{(0)}$ is i-th pose prediction from the pose decoder. Joints in the reference pose are iteratively refined by applying predicted coordinate offsets $\Delta\tilde{\mathbf{p}}_{i,\texttt{radar}}^{(l)}$:

$$\tilde{\mathbf{p}}_{i,\texttt{radar}}^{(l)} \in \mathbb{R}^{K \times 3} = \sigma(\sigma^{-1}(\tilde{\mathbf{p}}_{i,\texttt{radar}}^{(l-1)}) + \Delta\tilde{\mathbf{p}}_{i,\texttt{radar}}^{(l-1)}), \quad l = 1, \cdots, L_{\texttt{joint}}, \tag{4.6}$$

where the predicted offsets are given as $\Delta\tilde{\mathbf{p}}^{(l)}_{i,\texttt{radar}} = H_{\texttt{joint}}(\mathbf{Q}^{(l)}_{\texttt{joint}})$ with a shared regression head. We collect refined reference poses from the joint decoder as $\hat{\mathbf{P}}_{\texttt{radar}} = \{\tilde{\mathbf{p}}^{(L_{\texttt{joint}})}_{i,\texttt{radar}}\}^{N}_{i=1}$ and convert them into the 3D world coordinate system as $\hat{\mathbf{P}}_{\texttt{world}} = \mathcal{T}_{\texttt{r2w}}(\hat{\mathbf{P}}_{\texttt{radar}})$.

Pseudo-3D Deformable Attention

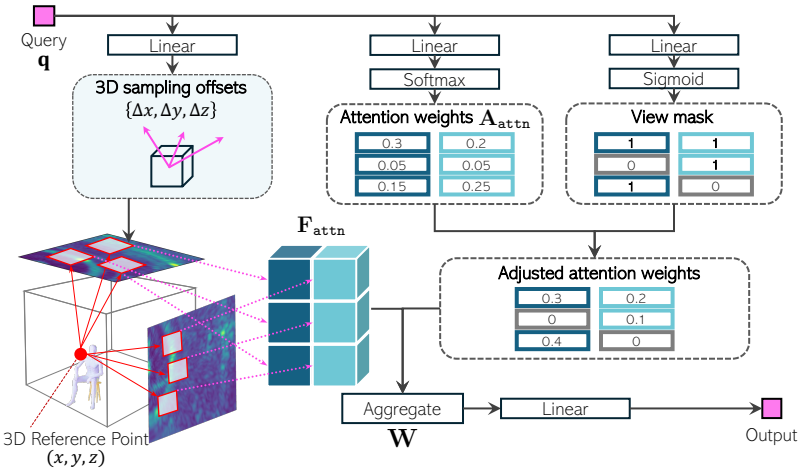


Figure 4.4: The pseudo-3D deformable attention operates on a 3D reference point and 3D sampling offsets that are projected to different radar views for pseudo-3D attention between multi-view radar features and the query.

Our two-stage decoder incorporates a pseudo-3D deformable attention module, where "pseudo" highlights that reference points and sampling offsets are defined in 3D space, while feature sampling occurs on the 2D radar views, as illustrated in Figure 4.4.

Consider a 3D reference point (x, y, z) in the 3D radar space with a corresponding query $\mathbf{q} \in \mathbb{R}^d$ (from either pose queries in the pose decoder or joint queries in the joint decoder). We first feed $\mathbf{q}$ into a linear projection layer to predict a set of 3D sampling offsets $\{\Delta x_i, \Delta y_i, \Delta z_i\}^{N_{\texttt{offset}}}_{i=1}$. Given the 3D reference point and sampling offsets, we can locate the 3D sampling coordinates and project them onto the two radar views, extracting deformable multi-view radar features:

$$\mathbf{f}^{(i)}_{\texttt{hor}} = \mathbf{F}_{\texttt{hor}}(x + \Delta x_i, z + \Delta z_i), \quad \mathbf{f}^{(i)}_{\texttt{ver}} = \mathbf{F}_{\texttt{ver}}(y + \Delta y_i, z + \Delta z_i) \quad i = 1, \cdots, N_{\texttt{offset}}. \tag{4.7}$$

We group deformable multi-view radar features as $\mathbf{F}_{\texttt{attn}} = \{\mathbf{f}^{(1)}_{\texttt{hor}}, \mathbf{f}^{(1)}_{\texttt{ver}}, \cdots, \mathbf{f}^{(N_{\texttt{offset}})}_{\texttt{hor}}, \mathbf{f}^{(N_{\texttt{offset}})}_{\texttt{ver}}\}$. Meanwhile, multi-view attention weights $\mathbf{A}_{\texttt{attn}} \in \mathbb{R}^{N_{\texttt{offset}} \times 2}$ (where 2 corresponds to the two radar views) are proposed by linearly projecting the query and applying a softmax normalization. These weights capture the relative importance of radar features across the two views in a unified manner. Given the deformable multi-view radar features $\mathbf{F}_{\texttt{attn}}$ and the multi-view attention weights $\mathbf{A}_{\texttt{attn}}$, deformable multi-view attention features can be calculated as

$$\bar{\mathbf{F}}_{\texttt{attn}} = \sum_{i=1}^{N_{\texttt{offset}}} (A_{i,0}\mathbf{W}\mathbf{F}^{(2i-1)}_{\texttt{attn}} + A_{i,1}\mathbf{W}\mathbf{F}^{(2i)}_{\texttt{attn}}), \tag{4.8}$$

where $A_{i,0}$ and $A_{i,1}$ are the attention weights in $\mathbf{A}_{\texttt{attn}}$ for the i-th deformable radar feature in the horizontal and, respectively, vertical radar views, and $\mathbf{W} \in \mathbb{R}^{C \times C}$ is a learnable weight matrix. We denote the overall pseudo-3D deformable attention as $\bar{\mathbf{F}}_{\texttt{attn}} = \texttt{DeformableAttn}(\mathbf{F}_{\texttt{ver}}, \mathbf{F}_{\texttt{hor}}, (x, y, z), \mathbf{q})$.

We also propose the view mask module (top right of Figure 4.4) that aims to put more attention weights on the feature on the more important view with a hard-thresholding approach. The view mask module first computes a view selection mask $\mathbf{M}_{\texttt{attn}}$ from a query $\mathbf{q}$ corresponding to reference points of interest, as:

$$\mathbf{M}_{\texttt{attn}} = \sigma(\lambda \cdot \texttt{FFN}(\mathbf{q})) \in \mathbb{R}^{N_{\text{offset}} \times 2}, \tag{4.9}$$

where σ is the Sigmoid function, and $\lambda (\approx 1e5) \in \mathbb{R}$ makes the sigmoid output very close to 0 or 1 while preserving gradients. The element $m_{i,j} (\approx 1)$ signals that the jth view ($j = 0$ to the horizontal and $j = 1$ to the vertical) should retain its share of attention at the ith sampling point, whereas $m_{i,j} (\approx 0)$ marks it for suppression. For each ith sampling point, we can consider three patterns to adjust the attention weights $\mathbf{A}_{\texttt{attn}}$ for that point $A_{i,0}, A_{i,1}$ and, potentially, other sampling points according to the corresponding row in the view selection mask $[m_{i,0}, m_{i,1}]$.

- $[m_{i,0}, m_{i,1}] = [1, 1]$: use both views, weights unchanged;
- $[m_{i,0}, m_{i,1}] = [0, 1]$ or $[m_{i,0}, m_{i,1}] = [1, 0]$: ignore one view and transfer its weight to the other, e.g., when $[m_{i,0}, m_{i,1}] = [0, 1]$, the adjusted attention weights are $\hat{A}_{i,0} = A_{i,0} + A_{i,1}, \hat{A}_{i,1} = 0$, to to ensure that the view selection decision at one sampling point does not influence the other sampling points;
- $[m_{i,0}, m_{i,1}] = [0, 0]$: ignore this sampling point in both views. Its weight is evenly redistributed so that $\sum_{i,j} \hat{A}_{i,j} = 1$ still holds.

In this way, the view selection mask adaptively changes through training so that the attention mechanism associates with the view that is more informative for each sampling point.

Structural Loss Function

As illustrated in Figure 4.3, RAPTR utilizes weak supervision labels: coarse-grained BBox labels $\mathbf{B}_{\texttt{world}}$ in the 3D world coordinate system and fine-grained 2D keypoint labels $\mathbf{P}_{\texttt{image}}$ in the image plane. The loss function is calculated between these labels $\{\mathbf{B}_{\texttt{world}}, \mathbf{P}_{\texttt{image}}\}$ and the initial and refined 3D pose estimates $\{\tilde{\mathbf{P}}_{\texttt{world}}, \hat{\mathbf{P}}_{\texttt{world}}\}$.

3D Template (T3D) Loss at Pose Decoder utilizes coarse-grained 3D BBox labels $\mathbf{B}_{\texttt{world}}$. For each $\mathbf{B}_{\texttt{world}}$, we construct a 3D keypoint template by computing the centroid of the corresponding 3D BBox, which serves as the 3D gravity center label $\mathbf{g}_{\texttt{world}} \in \mathbb{R}^{1 \times 3}$. Then, given a keypoint template defined at the coordinate origin $\mathbf{K}_{\texttt{world}} \in \mathbb{R}^{K \times 3}$, the corresponding template pose $\mathbf{T}_{\texttt{world}}$ is computed as $\mathbf{T}_{\texttt{world}} = \mathbf{K}_{\texttt{world}} + \mathbf{1}^{\top} \mathbf{g}_{\texttt{world}}$. As illustrated in the lower right of Figure 4.3, the

T3D loss $\mathcal{L}_{\texttt{template}}$ is defined as the Euclidean distance between the template poses $\mathbf{T}_{\texttt{world}}$ and the initial 3D pose estimates $\tilde{\mathbf{P}}_{\texttt{world}}$ at the pose decoder.

Combined 3D Gravity (G3D) Loss and 2D Keypoint (K2D) Loss at Joint Decoder utilizes both the coarse-grained 3D BBox labels $\mathbf{B}_{\texttt{world}}$ and the fine-grained 2D keypoint labels $\mathbf{P}_{\texttt{image}}$ in the image plane, as illustrated in the upper right of Figure 4.3.

For the G3D loss, the refined 3D pose estimate $\hat{\mathbf{P}}_{\texttt{world}}$ is collapsed into its centroid as $\hat{\mathbf{g}}_{\texttt{world}} \in \mathbb{R}^{1\times 3}$ by averaging the keypoint coordinates along each spatial axis. The resulting G3D loss $\mathcal{L}_{\texttt{gravity}}$ is then defined as the Euclidean distance between the predicted and ground-truth 3D gravity centers, $\hat{\mathbf{g}}_{\texttt{world}}$ and $\mathbf{g}_{\texttt{world}}$.

For the K2D loss, the refined 3D pose estimate $\hat{\mathbf{P}}_{\texttt{radar}}$ in the radar coordinate system are first transformed into the 3D camera coordinate system via a calibrated coordinate transformation: $\hat{\mathbf{P}}_{\texttt{camera}} = \mathbf{R}\hat{\mathbf{P}}_{\texttt{radar}} + \mathbf{1}^\top \mathbf{t}$ where $\mathbf{R}$ and $\mathbf{t}$ denote the calibrated 3D rotation matrix and the translation vector, respectively. The resulting 3D camera-space pose estimates are then projected onto the 2D image plane via a known 3D-to-2D projection: $\hat{\mathbf{P}}_{\texttt{image}} = \texttt{proj}_{\texttt{image}}(\hat{\mathbf{P}}_{\texttt{camera}})$. Finally, the fine-grained 2D loss combines the image-plane Euclidean error $\mathcal{L}_{\texttt{kpt2D}}$ and the object keypoint similarity (OKS) loss $\mathcal{L}_{\texttt{OKS}}$ [203] between $\mathbf{P}_{\texttt{image}}$ and $\hat{\mathbf{P}}_{\texttt{image}}$.

Structural Loss Function: Following the set-based loss in [184], we employ bipartite matching to associate predictions $\{\hat{\mathbf{g}}_{\texttt{world}}, \tilde{\mathbf{P}}_{\texttt{world}}, \hat{\mathbf{P}}_{\texttt{world}}\}$ with their ground-truth labels $\{\mathbf{g}_{\texttt{world}}, \mathbf{T}_{\texttt{world}}, \mathbf{P}_{\texttt{world}}\}$. Based on these associations, we define the structural loss function as

$$\mathcal{L} = \frac{1}{N'}\sum_{i=1}^{N'}(\lambda_1\mathcal{L}_{\texttt{template}} + \lambda_2\mathcal{L}_{\texttt{gravity}} + \lambda_3\mathcal{L}_{\texttt{kpt2D}} + \lambda_4\mathcal{L}_{\texttt{OKS}}) + \lambda_5\mathcal{L}_{\texttt{cls}}, \tag{4.10}$$

where N' is the number of matched pairs, λ_i is the corresponding weighting factor for each loss term, and $\mathcal{L}_{\texttt{cls}}$ is the classification loss of the focal loss [206] with the confidence scores of the matched estimates.

4.1.4 Evaluation

Settings

- **Datasets:** We assess the performance of RAPTR and baseline models on the HIBER dataset* [131] and the MMVR dataset† [148], both of which are publicly available multi-view mmWave radar datasets designed for indoor human perception tasks. The HIBER dataset includes two-view radar heatmaps from 10 different viewpoints, the corresponding 3D keypoint labels, and the 3D BBox labels. We use data protocols "MULTI" and "WALK", and use views 2 through 10 for training, validation, and testing. The MMVR dataset includes

*https://github.com/Intelligent-Perception-Lab/HIBER
†https://zenodo.org/records/12611978

Table 4.1: Hyper parameters for RAPTR.

Name	Notation	Value	
		HIBER	**MMVR**
Data			
Radar image resolution	W, H, D	160, 160, 200	128, 128, 256
# of keypoints	K	14	17
Model params			
Backbone	-	ResNet 18	
# of feature scale	S	3	
Feedforward dimension in Transformer	-	1024	
# of encoder layers	$L_{\texttt{enc}}$	3	
# of pose decoder layers	$L_{\texttt{pose}}$	2	
# of joint decoder layers	$L_{\texttt{joint}}$	3	
# of deformable sampling offsets	$N_{\texttt{offset}}$	K / 4 (pose / joint decoder)	
# of heads in multi-head attention	-	8	
Feature dimension	d	128	
# of input frames	T	4	
# of pose query	N	10	
Training params			
Optimizer	-	AdamW	
Base learning rate	-	2e-4	
Weight decay	-	1e-4	
LR scheduler	-	Cosine Decay	
Batch size	-	32	
Epochs	-	50	
Gradient clip norm	-	0.1	
Early stopping patience	-	5 epochs	
Loss weights			
3D template loss	λ_1	1.0	
3D gravity loss	λ_2	1.0	
2D Keypoint loss	λ_3	5.0	
2D OKS loss	λ_4	1.0	
Class loss	λ_5	1.0	

two-view radar heatmaps in various indoor scenarios, the corresponding 2D keypoint labels, and the 3D BBoxes. We use a data split "P1S1", a single-person case in an open space.

- **Parameter Settings for RAPTR:** We use $T = 4$ consecutive frames as input to our RAPTR network. For the point decoder, the number of pose queries N is 10. For the joint decoder, the number of joint queries K depends on the dataset to be evaluated: $K = 14$ for HIBER and $K = 17$ for MMVR. The parameters relating to model training are summarized in Table 4.1.

- **Baselines:** We consider the following competitive radar/RF-based 3D pose estimation baselines: **Person-in-WiFi 3D** [32], **HRRadarPose** [140], and **QRFPose** [138]. We evaluate Person-in-WiFi 3D and HRRadarPose using their open-source implementations. As QRF-Pose has no public code, we reimplement it from scratches and verify similar performance to the original report [138] using 3D keypoint labels. For fair comparison, we adopt a loss function, similar to RAPTR, combining 2D keypoint loss and 3D gravity loss.

- **Metrics:** We employ **Mean Per Joint Pose Error (MPJPE)** with the unit of centimeters

Table 4.2: 3D pose estimation performance on HIBER (MPJPE: cm).

Env	Method	Head	Neck	Shoulder	Elbow	Wrist	Hip	Knee	Ankle	Overall	(h)	(v)	(d)
WALK	Person-in-WiFi 3D	54.28	57.01	54.18	54.81	59.98	53.98	60.32	68.84	58.25	25.60	23.94	36.20
	QRFPose	42.23	34.21	37.37	38.05	41.25	31.24	34.39	46.87	38.20	14.78	13.40	26.76
	HRRadarPose	30.23	25.44	33.70	34.15	42.33	27.71	31.55	40.46	33.96	15.14	13.13	19.85
	RAPTR (ours)	21.75	17.41	20.72	23.23	26.55	18.97	21.06	26.10	**22.32**	**8.41**	**4.85**	**17.73**
MULTI	Person-in-WiFi 3D	88.48	85.14	89.44	84.33	84.29	88.69	81.70	81.53	85.25	34.06	28.57	58.93
	QRFPose	49.49	44.48	45.54	46.77	49.06	40.99	41.87	51.57	46.11	18.20	14.13	34.39
	HRRadarPose	30.24	24.24	30.14	35.17	44.34	28.76	31.38	35.31	33.19	16.77	10.75	21.84
	RAPTR (ours)	18.39	13.13	16.44	20.12	24.62	15.01	17.76	23.22	**18.99**	**7.80**	**4.38**	**14.54**

Table 4.3: Pose estimation performance on MMVR (P1S1).

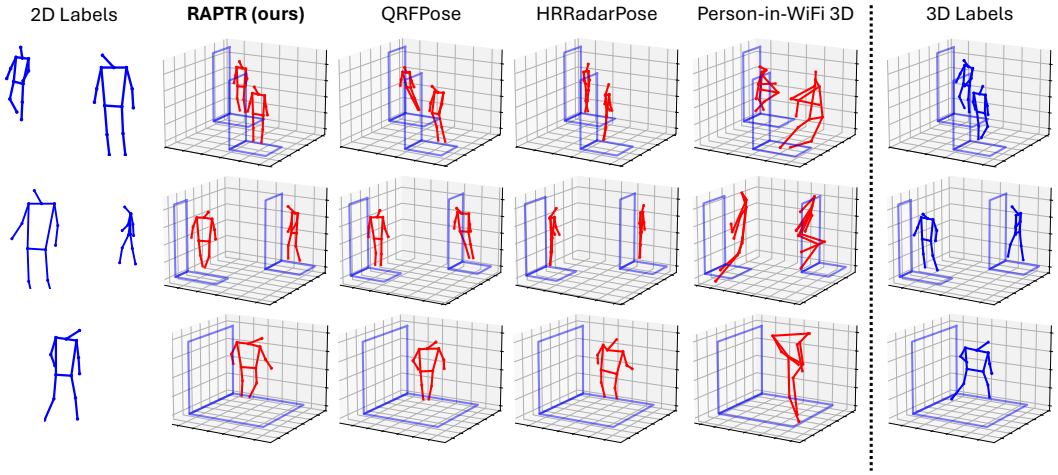
Method	Center distance (cm)	Edge length error (cm)		
		(h)	(v)	(d)
Person-in-WiFi 3D	136.14	33.18	95.43	242.86
QRFPose	210.75	38.12	73.69	409.38
HRRadarPose	164.46	37.84	74.00	313.81
RAPTR (ours)	**31.41**	**22.90**	**10.66**	**50.56**

in the world coordinate. In addition, we evaluate this MPJPE for each body joint and along each 3D axis, horizontal (h), vertical (v), and depth (d), independently. For MMVR, since 3D keypoint labels are not available, we construct a 3D BBox that encloses the estimated 3D keypoints and then use **the distance between the center points** of this BBox and the 3D BBox labels, as well as **the absolute error in the lengths of the edges** along each axis of the box, as metrics to approximate the 3D pose estimation performance.

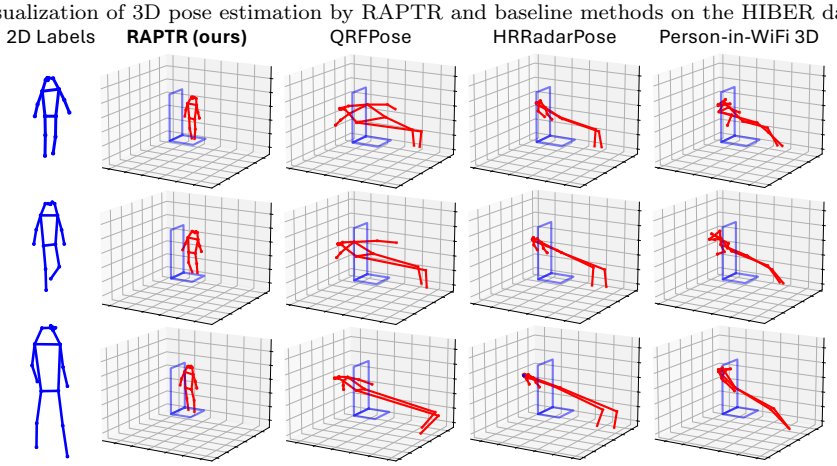
Main results

- **HIBER:** Table 4.2 shows the performance of 3D pose estimation for HIBER, using 2D and coarse 3D labels for baselines and our RAPTR. The qualitative results are provided in Figure 4.5(a).

  For WALK, RAPTR achieves a significantly lower overall MPJPE of 22.32 cm and outperforms all other baselines in the metric. More specifically, RAPTR reduces the overall error by 61.7%, 41.6%, and 34.3% compared to Person-in-WiFi 3D, QRFPose, and HRRadarPose, respectively. The per-joint breakdown demonstrates that RAPTR maintains its performance on relatively challenging joints, such as the wrist and ankle, where other baselines exhibit significant degradation. For example, HRRadarPose reports a wrist error of 42.33 cm, whereas RAPTR reports an error of 26.55 cm. Moreover, RAPTR maintains the error gap between the best- and worst-estimated joints within 10 cm, showing a consistent level of accuracy throughout the body. In terms of directional components, RAPTR shows much lower errors in the horizontal and vertical dimensions than baselines, indicating that RAPTR estimates well-proportioned 3D poses across all axes.



(a) Visualization of 3D pose estimation by RAPTR and baseline methods on the HIBER dataset.



(b) Visualization of 3D pose estimation by RAPTR and baseline methods on the MMVR dataset.

Figure 4.5: Qualitative results. Blue lines indicate the keypoint labels, blue rectangles indicate the 3D BBox labels, and red lines indicate the predictions.

For MULTI, the more challenging multi-person scenario, RAPTR continues to outperform with an overall MPJPE of 18.99 cm and shows a substantial margin compared to the second-best HRRadarPose at 33.19 cm. RAPTR reduces the overall error by 77.7%, 58.8%, and 42.7% compared to Person-in-WiFi 3D, QRFPose, and HRRadarPose, respectively. Although the overall accuracy of Person-in-WiFi 3D and QRFPose, noticeably degrades on the MULTI split compared to WALK, likely due to the increased complexity of handling multiple objects, RAPTR maintains a nearly consistent level of performance.

Referring to the qualitative results provided in Figure 4.5(a), RAPTR estimates structurally consistent 3D poses that match the 3D labels in both position and orientation, while baselines often suffer from misaligned limbs and implausible joint configurations. While baselines often fail to maintain human-like pose structure in the MULTI setting despite performing well in WALK, RAPTR consistently produces plausible estimates in both scenarios, indicating its

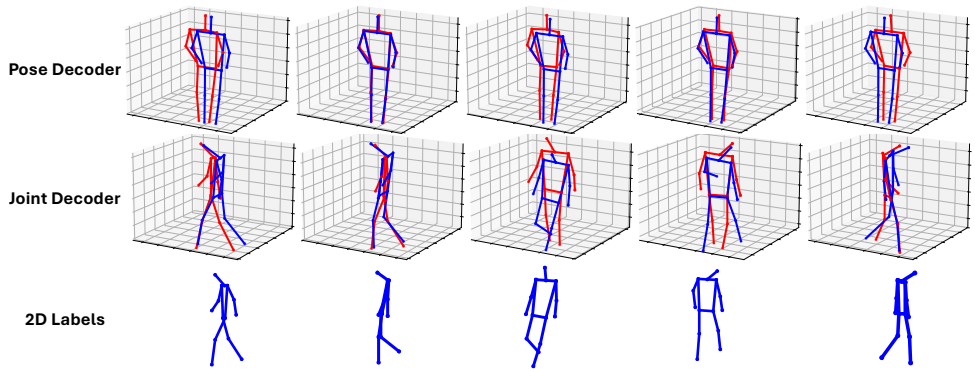


Figure 4.6: Pose Refinement Process: the pose prediction is first constrained by the 3D template at the pose decoder, and subsequently refined at the joint decoder.

robustness to multi-person scenes.

- **MMVR:** Table 4.3 shows the performance comparison for baselines and RAPTR with MMVR, and Figure 4.5(b) provides qualitative results. Although we cannot directly evaluate the precise 3D pose estimation performance for MMVR due to the absence of 3D pose labels, the results demonstrate that RAPTR effectively preserves reasonable human pose and location accuracy in the 3D space. Specifically, RAPTR shows improvements in center distance by 76.9%, 85.1%, and 80.9% compared to Person-in-WiFi 3D, QRFPose, and HRRadarPose, respectively. As shown in Figure 4.5(b), other baselines exhibit degraded performance due to structural collapse in 3D space, caused by overfitting to 2D alignment when projected onto the image plane. We assume that RAPTR effectively avoids this issue by not directly predicting the keypoints, but instead refining the final output through 2D keypoint supervision applied to each joint of a template pose that is placed in the 3D space.

4.1.5 Ablation Study

In this section, we present ablation studies of our RAPTR on the HIBER dataset. Unless otherwise stated, all reported evaluation results are reported as the mean ± standard deviation, computed over three random seeds.

Visualization of Pose Refinement Process: Figure 4.6 illustrates the refinement process of a 3D prediction through the two-stage decoder architecture. The pose decoder first establishes coarse 3D structures of the human body under the constraint of the 3D template loss (1st row). Subsequently, the joint decoder fine-tunes the keypoints to better capture the subject orientation and limb configuration (2nd row), while preserving the structure consistency.

Effect of Loss Terms: Table 4.4 provides an ablation study on the effect of different combinations of loss terms in the two-stage decoder. Note that the loss term `K2D`, `T3D`, and `G3D` in the

Table 4.4: Effect of loss terms for RAPTR (MPJPE: cm).

Loss		WALK	MULTI	Notes
Pose Dec.	Joint Dec.			
–	K2D	381.18 ± 0.28	375.73 ± 6.31	2D keypoint loss only at joint decoder
–	K2D+G3D	28.54 ± 4.57	57.90 ± 9.81	without 3D template loss at pose decoder
K2D+G3D	K2D+G3D	27.49 ± 3.40	23.43 ± 3.44	with 2D keypoint + 3D gravity loss at both decoders
T3D	K2D	25.96 ± 4.95	25.83 ± 3.87	without 3D gravity loss at joint decoder
T3D	K2D+G3D	**22.32 ± 0.06**	**18.99 ± 0.16**	proposed structural loss

Table 4.5: Effect of deformable attention mechanisms for RAPTR (MPJPE: cm).

Attn.	WALK	MULTI	Notes
2D	23.25 ± 1.38	19.47 ± 0.95	RAPTR with decoupled 2D deformable attention
3D	**22.32 ± 0.06**	**18.99 ± 0.16**	RAPTR with pseudo-3D deformable attention

table represent the 2D keypoint loss, the 3D template loss, and the 3D gravity loss, respectively. When only the K2D loss is applied at the joint decoder (row 1), the 3D pose estimation suffers from depth ambiguity due to the absence of any 3D constraint, resulting in a substantial increase in MPJPE to 381.18 cm and 375.73 cm on the WALK and MULTI splits, respectively. From rows 2 to 4 in Table 4.4, we remove or modify one loss term at a time from the proposed structural loss. Removing the T3D loss at the pose decoder (row 2), replacing the T3D with the K2D+G3D loss at the pose decoder (row 3), or removing the G3D loss at the joint decoder (row 4) leads to a noticeable degradation in 3D pose estimation.

Effect of Deformable Attention Mechanisms: Table 4.5 presents an ablation study on the effect of the deformable attention mechanism for RAPTR. In this study, the pseudo-3D deformable attention is replaced with the decoupled 2D deformable attention used in QRFPose [138], while keeping the cross-view encoder, two-stage decoder architecture, and the proposed structural loss unchanged. The results show that the pseudo-3D deformable attention yields marginal performance improvements, approximately 4% and 2.5% on the WALK and MULTI splits, respectively.

Comparison with a 2D-to-3D Pose Uplifting Model: We further compare RAPTR with a baseline that first estimates 2D keypoints in the image plane and subsequently lifts them to 3D space using a pre-trained 2D-to-3D pose uplifting model [207] trained on vision-based datasets such as Human3.6M [208]. To ensure a fair comparison, this baseline adopts the same network architecture as RAPTR, but both the pose and joint decoders are supervised only by the 2D keypoint loss. Because the 3D poses predicted by the uplifting model are defined in a pelvis-centered coordinate system, we additionally estimate a translation offset to align the estimated poses with their correct position in the world coordinate system. As shown in Table 4.6, the pose uplifting baseline performs significantly worse than RAPTR, with MPJPEs of 43.43 cm and

Table 4.6: Comparison with a 2D-to-3D pose uplifting model (MPJPE: cm).

	Loss		WALK	MULTI
	Pose Dec.	Joint Dec.		
Pose Lifting [207]	K2D	K2D	43.43 ± 2.66	41.76 ± 6.85
RAPTR (ours)	T3D	K2D+G3D	**22.32 ± 0.06**	**18.99 ± 0.16**

Table 4.7: Effect of view selection patterns on the performance.

	MPJPE	(h)	(v)	(d)
Both Views	20.31 ± 0.34	8.44 ± 0.34	5.02 ± 0.05	15.21 ± 0.67
Horizontal View Only	20.55 ± 2.05	8.74 ± 0.67	5.42 ± 0.82	14.97 ± 1.99
Vertical View Only	23.78 ± 0.77	12.24 ± 1.45	4.66 ± 0.14	16.61 ± 1.15
Random Mask	21.18 ± 0.33	8.75 ± 0.78	4.96 ± 0.72	16.04 ± 0.10
Adaptive View Selection (ours)	18.99 ± 0.16	7.80 ± 0.31	4.38 ± 0.25	14.54 ± 0.13

41.76 cm on the WALK and MULTI splits, respectively.

Impact of View Selection: To investigate how the use of features from horizontal and vertical views affects performance, we evaluate the performance of 3D pose estimation by configuring the view selection mask $\mathbf{M}_{\texttt{attn}} \in \mathbb{R}^{N_{\text{offset}} \times 2}$, integrated into pseudo-3D deformable attention according to several predefined patterns. Specifically, given that the first column of the view selection mask $\mathbf{M}_{\texttt{attn}}[:, 0]$ and the second column $\mathbf{M}_{\texttt{attn}}[:, 1]$ correspond to horizontal and vertical views, respectively, we set the values in the mask as

- **Both Views**: all values in $\mathbf{M}_{\texttt{attn}}$ to 1 so that the attention weight matrix is used as is;
- **Horizontal View Only**: $\mathbf{M}_{\texttt{attn}}[:, 0] = \mathbf{1}$ and $\mathbf{M}_{\texttt{attn}}[:, 1] = \mathbf{0}$ so that the features sampled from the horizontal feature map are aggregated and those of the vertical feature are omitted;
- **Vertical View Only**: The reverse of "horizontal only": $\mathbf{M}_{\texttt{attn}}[:, 0] = \mathbf{0}$ and $\mathbf{M}_{\texttt{attn}}[:, 1] = \mathbf{1}$;
- **Random Mask**: The values in the mask are randomly assigned for each mini-batch step in the training process;
- **Adaptive View Selection** (ours): The mask values are adaptively determined by the corresponding queries.

Table 4.7 shows the performance comparison in MPJPE. When using only a single view as input, restricting the input to either the horizontal or vertical view results in a noticeable increase in MPJPE, averaging 20.55 cm and 23.78 cm, respectively. In contrast, the adaptive view selection scheme achieves the lowest average MPJPE of 18.99 cm, outperforming both the "random mask" and "both views" settings.

4.1.6 Conclusion and Contribution

This chapter presented RAPTR, a transformer-based framework for 3D human pose estimation from multi-view radar heatmaps under weak supervision. RAPTR introduces a two-stage decoding architecture that separately refines global pose structures and local joint details, combined with structured losses incorporating geometric and physical priors such as 3D gravity and 3D template losses. Through experiments on the HIBER and MMVR datasets, RAPTR demonstrated high accuracy even under weak supervision with only 2D keypoints and 3D BBoxes. This result highlights the model's robustness and the effectiveness of its structured learning design in extracting reliable 3D body structures from limited supervision.

From a thesis-wide perspective, this chapter addresses the methodological challenge of realizing scalable 3D human perception under limited annotation availability. By bridging fully supervised and weakly supervised paradigms within radar-based perception, RAPTR establishes a practical framework that balances annotation cost, physical interpretability, and generalization to real-world indoor environments. This contribution complements the dataset-oriented foundations of Chapter 2 and the signal-based modeling approaches of Chapters 3 forming the methodological pillar toward reproducible radar-based perception.

Chapter 5

Propagation-aware Dynamic Multiband Fusion for Sub-7GHz and 60GHz

5.1 Multi-band Neural Fusion

Building upon the methodological foundations established in Chapter 3 and 4, this chapter extends the discussion from single-band accessibility to the modeling of temporal dynamics of multi-band and asynchronous wireless signals. While the previous chapters demonstrated how standard-compliant single-band measurements such as CBRs can be utilized for accessible and interpretable sensing, real-world Wi-Fi systems often operate across heterogeneous frequency bands and asynchronous links. To bridge this gap, we present a Multi-Band Neural Dynamic Fusion (NDF) framework that unifies asynchronous measurements from sub-7 GHz CSI and 60 GHz beam SNRs through continuous-time latent modeling. This chapter focuses not on the particular signal types employed, but on the general principle of asynchronous sequence fusion, representing the next methodological step from accessibility toward dynamic and propagation-aware representation learning for continuous-time perception.

5.1.1 Introduction

Wi-Fi sensing, including device localization and device-free human sensing, has attracted substantial interest over the past decade in both academic and industrial communities. This growing interest is exemplified by the formation of the IEEE 802.11bf WLAN Sensing task group in 2020, which aims to extend WLAN functionality beyond data transmission to support wireless sensing applications [84, 209–211].

Most existing Wi-Fi sensing methods are built upon coarse RSSI or fine-grained CSI measurements operating in sub-7 GHz frequency bands [25, 28, 96, 212–214]. With sufficiently high temporal resolution, CSI characterizes detailed channel behavior via channel frequency responses

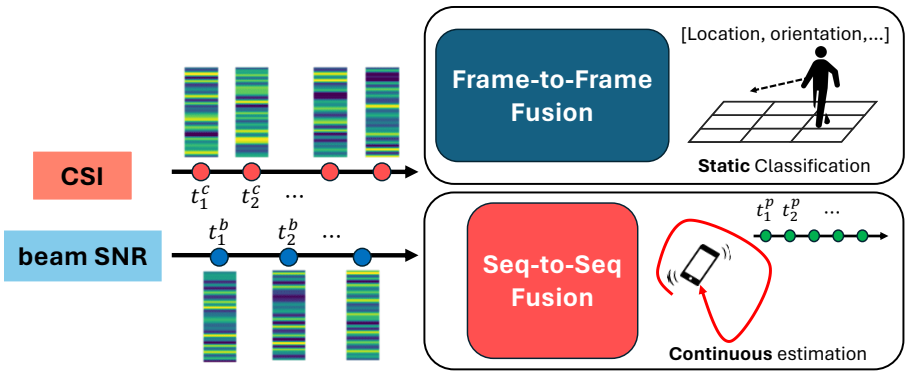


Figure 5.1: Multi-band Wi-Fi fusion, extending a *frame-to-frame* formulation [216] for classification (top) to an asynchronous *sequence-to-sequence* formulation for continuous-time regression (bottom).

across subcarriers (delay) and across multiple transmit–receive antenna pairs (angular information). At the same time, such fine granularity makes CSI sensitive to even slight environmental changes, often leading to channel instability. In contrast, intermediate-resolution measurements used in 60 GHz mmWave systems, such as beam SNR, have been reported to exhibit greater temporal stability [72–75, 215]. These beam SNR values are acquired during sector-level directional beam training, a mandatory procedure in mmWave Wi-Fi to overcome severe path loss and establish reliable links between access points and devices. Nevertheless, the reliance on beam training and subsequent association processes results in sparse measurements with low effective frame rates and nonuniform sampling intervals.

Fusion-based strategies have been widely studied as a means to enhance sensing robustness and performance. At the modality level, heterogeneous fusion has been explored by combining Wi-Fi with complementary sensors, including vision [39] and UWB [216]. Within Wi-Fi measurements themselves, different channel descriptors can also be fused: for example, CSI and RSSI can be concatenated to support joint feature learning [217], and the phase and amplitude components of fine-grained CSI can be jointly exploited for localization [218]. For multi-band Wi-Fi sensing, CSI splicing has emerged as a representative channel fusion technique, in which phase measurements across multiple frequencies are stitched together. By compensating for hardware-induced phase distortions, this approach enables adjacent channels to be coherently aligned, effectively reconstructing CSI over a much wider bandwidth and thereby improving AoA and ToF resolution [31, 149, 219, 220]. These methods, however, typically rely on narrowband CSIs collected from multiple channels under the same wireless standard, where consistent CSI formats and tight temporal synchronization are ensured through rapid channel hopping. In contrast, we consider a more challenging fusion scenario that combines sub-7 GHz and mmWave measurements. In this setting, channel observations across the two bands are generally asynchronous and differ significantly

in their measurement characteristics. Only a small number of prior studies have addressed such cross-band integration, and most existing approaches treat sub-7 GHz CSI mainly as an auxiliary signal to assist mmWave sensing, for instance by narrowing beam search spaces or selecting candidate mmWave access points based on coarse localization [73, 221]. Our earlier work [215] focused on relatively simple sensing tasks, including pose classification with a limited set of static poses, seat occupancy detection, and grid-based localization. Although the sub-7 GHz CSI and mmWave measurements were sampled at different time instants, they could still be fused using a frame-to-frame formulation because each asynchronous sample corresponded to the same stationary label (e.g., pose, occupancy, or location), making precise temporal alignment unnecessary; see the top panel of Figure 5.1, where the sampling times t_n^c and t_n^b are not synchronized. More recently, [222] investigated sequence-to-sequence fusion across Wi-Fi bands by jointly using fine-grained CSI from sub-7 GHz and 60 GHz systems with the aid of a reconfigurable intelligent surface. Assuming comparable frame rates for both CSIs, their approach leverages attention mechanisms and transformer architectures to fuse multi-band channel measurements on simulated datasets.

For more challenging problems, such as continuous-time object trajectory estimation using asynchronous multi-band Wi-Fi measurements, several fundamental issues arise. First, channel observations collected from different frequency bands must be temporally reconciled in order to infer object motion. As depicted in the lower panel of Figure 5.1, CSI samples at times t_n^c are not synchronized with beam SNR measurements at times t_n^b. Moreover, additional temporal offsets may exist between the sensing inputs at t_n^c or t_n^b and the target trajectory states at t_n^p that the model aims to estimate. Second, mmWave beam SNR measurements are acquired at substantially lower rates than CSI. Even under ideal conditions, beam SNRs are typically available at around 10 Hz, corresponding to a beacon interval of 100 ms. In practice, this rate further degrades when multiple users compete for channel access during uplink beam training within each beacon interval. Third, because beam training relies on contention-based channel access, the resulting beam SNR measurements for a given user are often irregularly sampled at the access point, introducing additional temporal nonuniformity that complicates trajectory estimation.

To overcome these challenges, we introduce a multi-band Wi-Fi Neural Dynamic Fusion (NDF) framework. The proposed approach generalizes the stationary *frame-to-frame* formulation in [215] (upper panel of Figure 5.1) to a dynamic, asynchronous *sequence-to-sequence* setting (lower panel of Figure 5.1). This extension enables more demanding downstream tasks, including continuous-space regression and continuous-time object trajectory estimation. Building upon our earlier beam-SNR-only framework in [72], the proposed multi-band NDF introduces a unified neural architecture with multiple encoders, latent dynamic learning components, a post-ODE fusion module, and multiple decoders, as illustrated in Figure 5.2. The main contributions are summarized below:

1. We present, to the best of our knowledge, the first study that tackles asynchronous multi-

band fusion of sub-7 GHz CSI and 60 GHz beam SNR for trajectory estimation of moving objects.

2. We propose a multi-encoder, multi-decoder NDF architecture (Figure 5.2), in which separate encoders estimate initial latent conditions for heterogeneous input sequences, ODE-based models [223, 224] learn and align latent dynamics, and a post-ODE fusion module integrates the aligned latent states.

3. We investigate several post-ODE fusion strategies, including multilayer perceptron (MLP) fusion, pairwise interaction fusion, and weighted importance fusion.

4. We formulate an variational evidence lower bound (ELBO)-based loss that couples prior and approximate posterior distributions of the initial latent states with the likelihoods of multiple decoder outputs, combining unsupervised multi-band reconstruction with supervised coordinate regression.

5. We develop an automated data collection system using commercial off-the-shelf IEEE 802.11ac/ad Wi-Fi routers and a TurtleBot mobile platform, enabling synchronized acquisition of 5 GHz CSI, 60 GHz beam SNR, and ground truth trajectories.

6. We conduct extensive ablation studies on real-world data to analyze trajectory estimation accuracy, generalization behavior, and model interpretability.

5.1.2 Problem Formulation

We formulate trajectory estimation as a continuous regression task based on asynchronous sequences of CSI and beam SNR measurements. As illustrated in Figure 5.2, at each beam SNR measurement time t_n^b, a vector of M_b beam SNR values is collected as $\mathbf{b}_n = [b_{n,1}, b_{n,2}, \cdots, b_{n,M_b}]^\top \in \mathbb{R}^{M_b \times 1}$, where each element corresponds to a specific beam pattern. At a time instant t_n^c, a CSI sample $\mathbf{C}_n \in \mathbb{C}^{N_{\mathrm{Tx}} \times N_{\mathrm{Rx}} \times N_{\mathrm{s}}}$ is obtained, whose entry $C_n(i, j, k)$ denotes the channel frequency response between the i-th transmit antenna, the j-th receive antenna, and the k-th subcarrier. Within a temporal window of duration ΔT_w, we aggregate N_b beam SNR samples and N_c CSI measurements into two asynchronous input sequences. The goal is to infer the object trajectory $\mathbf{p}_n$ at N_p designated prediction times t_n^p inside the same window ΔT_w, namely,

$$\{\mathbf{b}_n, t_n^b\}_{n=0}^{N_b}, \{\mathbf{C}_n, t_n^c\}_{n=0}^{N_c} \rightarrow \{\mathbf{p}_n, t_n^p\}_{n=0}^{N_p}, \tag{5.1}$$

where $\mathbf{p}_n = [x_n, y_n]^\top$ denotes the two-dimensional position at time t_n^p.

We preprocess the raw CSI measurements $\mathbf{C}_n$ using standard calibration techniques to compensate for the lack of synchronization between the Wi-Fi transmitter and receiver [50, 51, 187, 225–227].

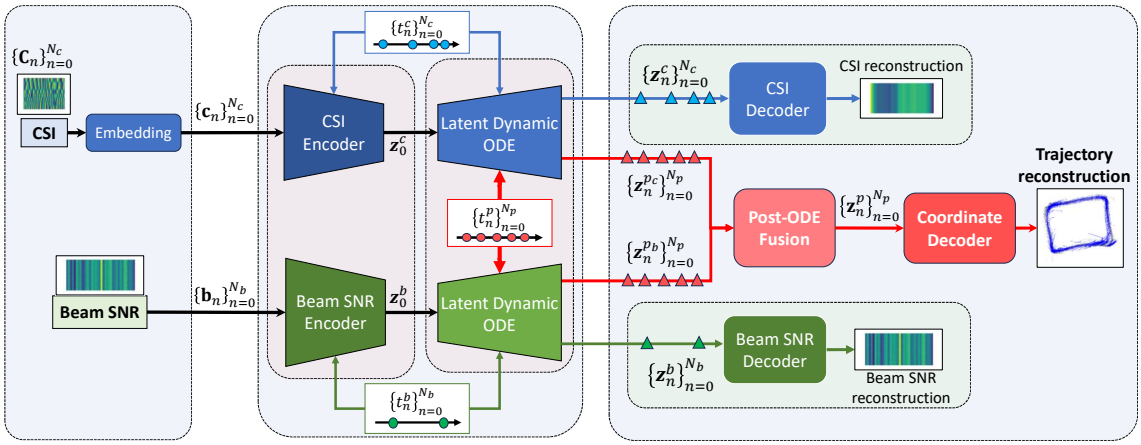


Figure 5.2: Network architecture of the multi-band Wi-Fi neural dynamic fusion (NDF) for continuous-time object trajectory estimation. The architecture consists of multiple encoders, latent dynamic learning modules, a post-ODE fusion module, and multiple decoders.

In particular, we apply the SpotFi method [225] to correct linear phase distortions induced by sampling time offset (STO), and further suppress packet-level phase variations through antenna-wise conjugate multiplication [50]. Following calibration, each CSI sample is projected into a compact representation using a pretrained convolutional autoencoder (CAE), yielding a low-dimensional CSI embedding vector $\mathbf{c}_n \in \mathbb{R}^{M_c \times 1}$, where M_c denotes the embedding dimension.

As depicted in Figure 5.2, the two asynchronous input sequences can therefore be reformulated in terms of the CSI embeddings $\mathbf{c}_n$ and the beam SNR vectors $\mathbf{b}_n$, allowing the overall problem to be rewritten as:

$$\{\mathbf{b}_n, t_n^b\}_{n=0}^{N_b}, \quad \{\mathbf{c}_n, t_n^c\}_{n=0}^{N_c} \rightarrow \{\mathbf{p}_n, t_n^p\}_{n=0}^{N_p}. \tag{5.2}$$

5.1.3 Existing Fusion Strategies

Frame-to-Frame Fusion Approach

Because the CSI samples at t_n^c and the beam SNR measurements at t_n^b are not temporally synchronized, a simple fusion approach is to perform alignment directly at the input stage. In practice, this can be done by resampling one sequence to match the timestamps of the other, for example through linear interpolation or nearest-neighbor matching.

With the linear interpolation (**LinearInt**) approach, for each target time point t_n^p, we first locate the adjacent samples in the CSI and beam SNR sequences. Specifically, indices i_c and i_b are chosen such that $t_{i_c}^c \leq t_n^p \leq t_{i_c+1}^c$ and $t_{i_b}^b \leq t_n^p \leq t_{i_b+1}^b$. The input features at t_n^p are then computed

by linearly interpolating between the two neighboring measurements that bound each interval

$$\begin{aligned}\mathbf{b}_n^{\text{Lin}} &= \mathbf{b}_{i_b} + \frac{t_n^p - t_{i_b}^b}{t_{i_b+1}^b - t_{i_b}^b}(\mathbf{b}_{i_b+1} - \mathbf{b}_{i_b}),\\ \mathbf{c}_n^{\text{Lin}} &= \mathbf{c}_{i_c} + \frac{t_n^p - t_{i_c}^c}{t_{i_c+1}^c - t_{i_c}^c}(\mathbf{c}_{i_c+1} - \mathbf{c}_{i_c}),\end{aligned} \tag{5.3}$$

where n ranges from 0 to N_p. As an alternative, the nearest-neighbor strategy (**NearestInt**) assigns the input by selecting, from each sequence, the measurement whose timestamp is temporally closest to the target output time t_n^p.

$$\begin{aligned}\mathbf{b}_n^{\text{Nea}} &= \begin{cases}\mathbf{b}_{i_b}, & \text{if } t_n^p - t_{i_b}^b \le t_{i_b+1}^b - t_n^p\\ \mathbf{b}_{i_b+1}, & \text{otherwise,}\end{cases}\\ \mathbf{c}_n^{\text{Nea}} &= \begin{cases}\mathbf{c}_{i_c}, & \text{if } t_n^p - t_{i_c}^c \le t_{i_c+1}^c - t_n^p\\ \mathbf{c}_{i_c+1}, & \text{otherwise.}\end{cases}\end{aligned} \tag{5.4}$$

Given the aligned input sequences

$$\{\mathbf{b}_n^{\text{Lin/Nea}}, t_n^p\}_{n=0}^{N_p}, \quad \{\mathbf{c}_n^{\text{Lin/Nea}}, t_n^p\}_{n=0}^{N_p}, \tag{5.5}$$

the aligned sequences can be fused on a frame-to-frame basis, and the resulting representation is used to regress the trajectory coordinates

$$\hat{\mathbf{p}}_n = \mathcal{M}(\mathcal{F}(\mathbf{c}_n^{\text{Lin/Nea}}, \mathbf{b}_n^{\text{Lin/Nea}})), \tag{5.6}$$

where $\mathcal{F}$ denotes the chosen fusion operation, such as concatenation or alternative designs discussed in Section 5.1.4, and $\mathcal{M}$ denotes an MLP network. Other interpolation strategies, including spline-based and piecewise polynomial interpolation, can also be considered.

Sequence-to-Sequence Fusion Approach Unlike frame-to-frame fusion, RNN–based methods can model the temporal evolution of latent features across the full input sequence [96, 228, 229]. By combining the hidden representations learned from the CSI and beam SNR sequences, this approach enables what we term sequence-to-sequence fusion.

Given an input sequence $\{\mathbf{s}_n\}_{n=0}^{N}$, where $\mathbf{s}_n$ denotes either the CSI embedding $\mathbf{c}_n$ or the beam SNR vector $\mathbf{b}_n$, a conventional RNN updates its hidden state from $\mathbf{h}_{n-1}$ at time t_{n-1} to $\mathbf{h}_n$ at time t_n by incorporating the input observation $\mathbf{s}_n$ at time t_n as follows:

$$\mathbf{h}_n = \mathcal{R}(\tilde{\mathbf{h}}_n, \mathbf{s}_n; \boldsymbol{\theta}), \quad \tilde{\mathbf{h}}_n = \mathbf{h}_{n-1}, \tag{5.7}$$

where $\mathcal{R}$ is an Long Short-Term Memory (LSTM) [230] or Gated Recurrent Unit (GRU) [231] unit, and $\tilde{\mathbf{h}}_n$ is an auxiliary vector. In a conventional RNN formulation, the time gaps between consecutive samples are implicitly assumed to be constant, i.e., $\Delta t_n = t_n - t_{n-1}$ is identical for all n. Under this uniform-sampling assumption, the auxiliary state used for updating the network reduces to the hidden state from the previous step, namely $\tilde{\mathbf{h}}_n = \mathbf{h}_{n-1}$.

To handle sequences sampled at non-uniform time intervals, where Δt_n varies across steps, we consider several sequence-to-sequence baseline models. A representative example is **RNN-Decay** [224], which modifies the standard update by applying an exponential decay to the previous hidden state according to the elapsed time before passing it to the RNN cell,

$$\mathbf{h}_n = \mathcal{R}(\tilde{\mathbf{h}}_n, \mathbf{s}_n; \boldsymbol{\theta}), \quad \tilde{\mathbf{h}}_n = \mathbf{h}_{n-1} e^{-\Delta t_n}, \tag{5.8}$$

where the auxiliary vector $\tilde{\mathbf{h}}_n$ accounts for the irregular sampling intervals. Another baseline is **RNN-Δ** [224], which explicitly incorporates irregular sampling by appending the time interval information to the input representation,

$$\mathbf{h}_n = \mathcal{R}(\tilde{\mathbf{h}}_n, \tilde{\mathbf{s}}_n; \boldsymbol{\theta}), \quad \tilde{\mathbf{h}}_n = \mathbf{h}_{n-1}, \quad \tilde{\mathbf{s}}_n = [\mathbf{s}_n^\top, \Delta t_n]^\top, \tag{5.9}$$

while keeping the auxiliary vector as the previous hidden state.

For each target output time t_n^p, we first locate the neighboring indices i_c and i_b in the CSI and beam SNR sequences such that $t_{i_c}^c \le t_n^p \le t_{i_c+1}^c$ and $t_{i_b}^b \le t_n^p \le t_{i_b+1}^b$. The corresponding hidden states $\mathbf{h}_{i_c}^c$ and $\mathbf{h}_{i_b}^b$ are then obtained, having been updated beforehand according to either Equation (5.8) or Equation (5.9). The hidden states $\mathbf{h}_{i_c}^c$ and $\mathbf{h}_{i_b}^b$ are then propagated to the output time instance t_n^p as

$$\begin{aligned} \mathbf{h}_n^{p_c} &\triangleq \mathbf{h}_{t_n^p}^c = \mathcal{R}(\mathbf{h}_{i_c}^c e^{-(t_n^p - t_{i_c}^c)}, \mathbf{0}; \boldsymbol{\theta}), \\ \mathbf{h}_n^{p_b} &\triangleq \mathbf{h}_{t_n^p}^b = \mathcal{R}(\mathbf{h}_{i_b}^b e^{-(t_n^p - t_{i_b}^b)}, \mathbf{0}; \boldsymbol{\theta}), \end{aligned} \tag{5.10}$$

for the RNN-Decay update and

$$\begin{aligned} \mathbf{h}_n^{p_c} &\triangleq \mathbf{h}_{t_n^p}^c = \mathcal{R}(\mathbf{h}_{i_c}^c, [\mathbf{0}^\top, (t_n^p - t_{i_c}^c)]^\top; \boldsymbol{\theta}), \\ \mathbf{h}_n^{p_b} &\triangleq \mathbf{h}_{t_n^p}^b = \mathcal{R}(\mathbf{h}_{i_b}^b, [\mathbf{0}^\top, (t_n^p - t_{i_b}^b)]^\top; \boldsymbol{\theta}), \end{aligned} \tag{5.11}$$

for the RNN-Δ update, where $n = 1, \cdots, N_p$, by setting the current input $\mathbf{s}_n = \mathbf{0}$ at t_n^p. The temporally aligned hidden representations $\mathbf{h}_n^{p_c}$ and $\mathbf{h}_n^{p_b}$ at time t_n^p are subsequently combined through a fusion operation, and the fused latent state is used to predict the corresponding trajectory coordinates,

$$\hat{\mathbf{p}}_n = \mathcal{M}(\mathcal{F}(\mathbf{h}_n^{p_c}, \mathbf{h}_n^{p_b})), \tag{5.12}$$

where $\mathcal{F}$ denotes the fusion operation and $\mathcal{M}$ is implemented as an MLP.

5.1.4 Multi-Band Neural Dynamic Fusion

This section details each component of the multi-band NDF architecture shown in Figure 5.2. The framework uses two dedicated encoders to process the CSI embeddings $\mathbf{c}_n$ and beam SNR measurements $\mathbf{b}_n$ together with their timestamps t_n^c and t_n^b, mapping each sequence into a latent

space and inferring the corresponding initial latent states $\mathbf{z}_0^c$ and $\mathbf{z}_0^b$. Within the latent dynamics module, these initial states are evolved through a shared learnable ODE model [223, 224], producing latent trajectories $\mathbf{z}_n^{p_c}$ and $\mathbf{z}_n^{p_b}$ at the target times t_n^p to enable temporal alignment. The aligned latent representations are subsequently fused after ODE propagation and passed to a coordinate decoder to estimate object trajectories. Moreover, the same ODE model is used to infer latent states $\mathbf{z}_n^c$ and $\mathbf{z}_n^b$ at the original input times t_n^c and t_n^b, which can then be decoded by the corresponding CSI or beam SNR decoders for signal reconstruction.

Encoders: An Estimator for Initial Latent Conditions

The role of the encoder is to infer the posterior distribution of the initial latent state associated with a given input sequence. We begin by describing the encoder for the CSI modality.

The CSI embedding sequence $\mathbf{c}_{N_c}, \ldots, \mathbf{c}_0$ is processed in a backward temporal manner, starting from the final timestamp $t_{N_c}^c$ and proceeding toward the initial time t_0. At each step, the input $\mathbf{c}_n$ is transformed into a hidden representation $\mathbf{h}_n^c \in \mathbb{R}^{H_c \times 1}$ using an auxiliary hidden state $\tilde{\mathbf{h}}_n^c$ according to

$$\mathbf{h}_n^c = \mathcal{R}(\tilde{\mathbf{h}}_n^c, \mathbf{c}_n; \boldsymbol{\theta}_g^c), \tag{5.13}$$

where $\mathcal{R}$ denotes a recurrent unit, realized using either a GRU or LSTM with learnable parameters $\boldsymbol{\theta}_g^c$.

To handle non-uniform temporal sampling in the input sequence, where the interval $\Delta t_n^c = t_n^c - t_{n-1}^c$ varies over time, we incorporate a numerical ODE solver such as Euler or Runge–Kutta methods to evolve the hidden state backward in time. Specifically, the hidden vector $\mathbf{h}_{n+1}^c$ at timestamp t_{n+1}^c is propagated to an auxiliary state $\tilde{\mathbf{h}}_n^c$ at t_n^c, as illustrated in Figure 5.3 [232–235]:

$$\begin{aligned} \tilde{\mathbf{h}}_n^c &= \mathcal{S}(\mathcal{O}_e, \mathbf{h}_{n+1}^c, (t_{n+1}^c, t_n^c); \boldsymbol{\theta}_e^c) \\ &= \mathbf{h}_{n+1}^c + \int_{\tau=t_{n+1}^c}^{t_n^c} \mathcal{O}_e(\mathbf{h}(\tau), \tau; \boldsymbol{\theta}_e^c) d\tau, \end{aligned} \tag{5.14}$$

where $\mathcal{O}_e$ denotes a learnable neural ODE function that governs the temporal evolution of the hidden state and is parameterized by $\boldsymbol{\theta}_e^c$.

By iteratively applying Equations (5.13) and (5.14), the hidden state is propagated backward from the final time $t_{N_c}^c$ to the initial time t_0, yielding the terminal hidden representation $\mathbf{h}_0^c$. This representation is then used to infer the initial latent state $\mathbf{z}_0^c$ and its posterior distribution, which is approximated as a Gaussian with mean $\boldsymbol{\mu}^c$ and variance $(\boldsymbol{\sigma}^c)^2$

$$q_{\theta_c}(\mathbf{z}_0^c | \mathbf{c}_{N_c}, \cdots, \mathbf{c}_0) = q_{\theta_c}(\mathbf{z}_0^c | \mathbf{h}_0^c) = \mathcal{N}(\boldsymbol{\mu}^c, (\boldsymbol{\sigma}^c)^2), \tag{5.15}$$

Following the variational autoencoder formulation [236], the Gaussian parameters $\boldsymbol{\mu}^c$ and $\boldsymbol{\sigma}^c$ are predicted from $\mathbf{h}_0^c$ via

$$\boldsymbol{\mu}^c, \boldsymbol{\sigma}^c = \mathcal{M}(\mathbf{h}_0^c; \boldsymbol{\theta}_z^c), \tag{5.16}$$

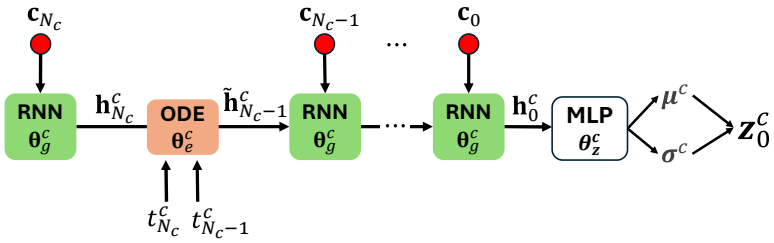


Figure 5.3: CSI encoder module used to estimate the initial latent condition associated with the input sequence. The beam SNR encoder follows an analogous design.

where $\mathcal{M}$ denotes an MLP parameterized by $\boldsymbol{\theta}_z^c$. Since the initial latent variable $\mathbf{z}_0^c$ is treated probabilistically, it is obtained using the reparameterization trick:

$$\mathbf{z}_0^c = \boldsymbol{\mu}^c + \boldsymbol{\sigma}^c \odot \boldsymbol{\epsilon}^c,\ \boldsymbol{\epsilon}^c \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_{L_c}), \tag{5.17}$$

where $\boldsymbol{\epsilon}^c$ is a standard Gaussian noise vector of dimension L_c, and $\odot$ denotes the Hadamard product.

A analogous procedure is carried out for the beam SNR sequence by applying Equations (5.13) to (5.17). Specifically, the beam SNR inputs are processed to infer the posterior distribution of the corresponding initial latent state, which is modeled as a Gaussian distribution:

$$q_{\theta_b}(\mathbf{z}_0^b | \mathbf{b}_{N_b}, \cdots, \mathbf{b}_0) = \mathcal{N}(\boldsymbol{\mu}^b, (\boldsymbol{\sigma}^b)^2), \tag{5.18}$$

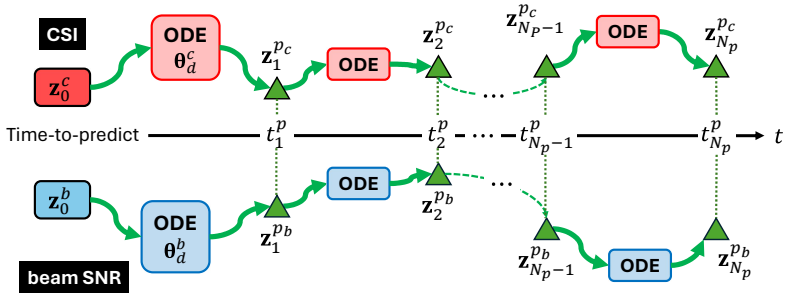
$$\mathbf{z}_0^b = \boldsymbol{\mu}^b + \boldsymbol{\sigma}^b \odot \boldsymbol{\epsilon}^b, \quad \boldsymbol{\epsilon}^b \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_{L_b}), \tag{5.19}$$

where L_b denotes the dimension of $\mathbf{z}_0^b$.

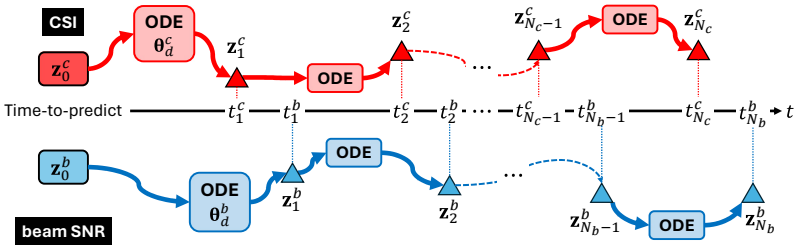
Latent Dynamic Learning Modules: Alignment in Latent Space

Given an initial latent state $\mathbf{z}_0$ associated with an input sequence, we evolve the latent representation over time using a continuous-time neural ODE, denoted by $\mathcal{O}_d$ and parameterized by θ_d. This ODE enables latent states to be queried at arbitrary time instances t_n^q. Depending on the choice of the query time, three different situations arise.

(a) $t_n^q = t_n^p$ for latent state alignment: We first consider the case in which the query time t_n^q matches the coordinate prediction time t_n^p. This setting is particularly important because it allows the latent representations derived from CSI and beam SNR to be aligned at a common temporal reference. Starting from their respective initial latent conditions, the latent trajectories for CSI and beam SNR are propagated forward in time to the same query instant using the shared ODE



(a) Unrolling latent dynamic ODEs for latent state alignment between CSI and beam SNR at t_n^p.



(b) Illustration of latent dynamic ODE unrolling for latent state recovery CSI at t_n^c (top) and beam SNR at t_n^b (bottom).

Figure 5.4: Unrolling latent dynamic ODEs for latent state alignment (a) and recovery (b).

formulation [232–235]:

$$\mathbf{z}_n^{p_c} \triangleq \mathbf{z}_{t_n^p}^c = \mathbf{z}_0^c + \int_{t_0}^{t_n^p} \mathcal{O}_d(\mathbf{z}_t^c, t; \boldsymbol{\theta}_d^c)dt, \tag{5.20}$$

$$\mathbf{z}_n^{p_b} \triangleq \mathbf{z}_{t_n^p}^b = \mathbf{z}_0^b + \int_{t_0}^{t_n^p} \mathcal{O}_d(\mathbf{z}_t^b, t; \boldsymbol{\theta}_d^b)dt, \tag{5.21}$$

where $\mathbf{z}_0^c$ and $\mathbf{z}_0^b$ correspond to the initial latent variables inferred from the CSI and beam SNR sequences, respectively.

In implementation, latent alignment is performed incrementally across successive query times. Once the latent states are aligned at a given time instance, they serve as the starting points for propagating the dynamics to the next query time t_{n+1}^q. This sequential alignment process is illustrated in Figure 5.4(a). For example, at the first prediction time t_1^p, the aligned latent states are obtained as

$$\mathbf{z}_1^{p_c} = \mathbf{z}_0^c + \int_{t_0}^{t_1^p} \mathcal{O}_d(\mathbf{z}_t^c, t; \boldsymbol{\theta}_d^c)dt,$$

$$\mathbf{z}_1^{p_b} = \mathbf{z}_0^b + \int_{t_0}^{t_1^p} \mathcal{O}_d(\mathbf{z}_t^b, t; \boldsymbol{\theta}_d^b)dt,$$

and then

$$\mathbf{z}_2^{p_c} = \mathbf{z}_1^{p_c} + \int_{t_1^p}^{t_2^p} \mathcal{O}_d(\mathbf{z}_t^c, t; \boldsymbol{\theta}_d^c)dt,$$
$$\mathbf{z}_2^{p_b} = \mathbf{z}_1^{p_b} + \int_{t_1^p}^{t_2^p} \mathcal{O}_d(\mathbf{z}_t^b, t; \boldsymbol{\theta}_d^b)dt,$$

where the latent dynamics for CSI and beam SNR are governed by neural ODEs parameterized by $\boldsymbol{\theta}_d^c$ and $\boldsymbol{\theta}_d^b$, respectively.

(b) $t_n^q = t_n^c$ for latent state recovery of CSI: In a manner similar to latent alignment at prediction times, the latent representation associated with each CSI measurement can be recovered at its original sampling time t_n^c. Starting from the initial latent condition inferred from the CSI sequence, the latent trajectory is propagated forward to the corresponding CSI timestamp:

$$\mathbf{z}_n^c \triangleq \mathbf{z}_{t_n^c}^c = \mathbf{z}_0^c + \int_{t_0}^{t_n^c} \mathcal{O}_d(\mathbf{z}_t^c, t; \boldsymbol{\theta}_d^c)dt. \tag{5.22}$$

(c) $t_n^q = t_n^b$ for latent state recovery of beam SNR

Likewise, latent states corresponding to beam SNR measurements are recovered at their native time instances t_n^b by evolving the beam SNR latent dynamics from the initial condition:

$$\mathbf{z}_n^b \triangleq \mathbf{z}_{t_n^b}^b = \mathbf{z}_0^b + \int_{t_0}^{t_n^b} \mathcal{O}_d(\mathbf{z}_t^b, t; \boldsymbol{\theta}_d^b)dt. \tag{5.23}$$

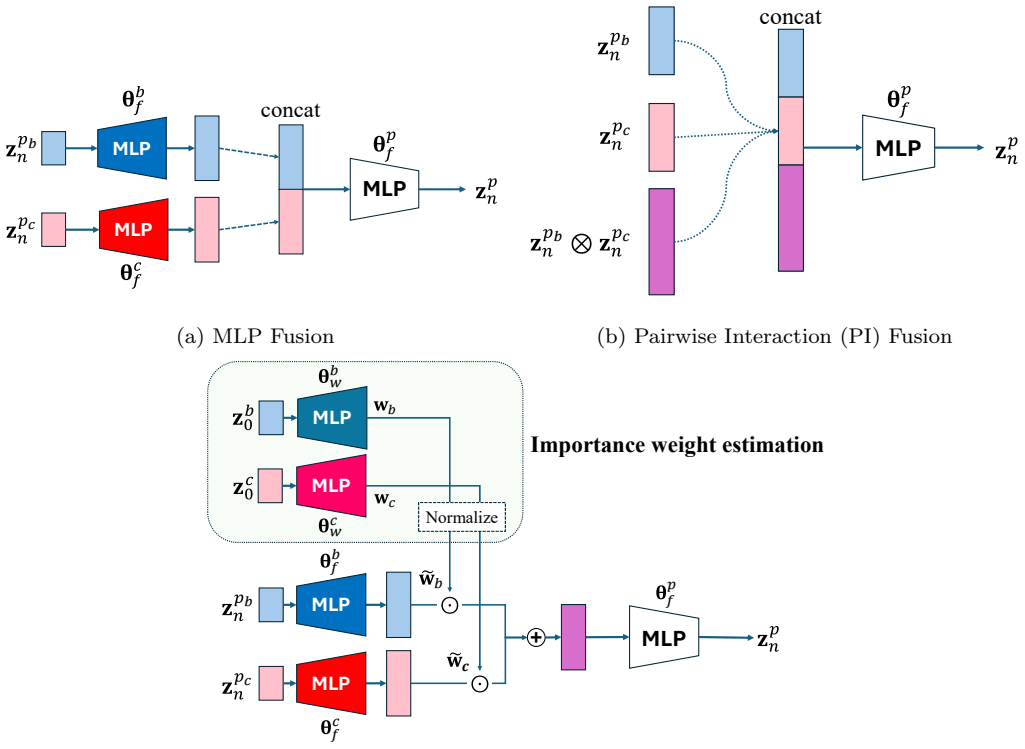
Here, the reference time t_0 is chosen to satisfy $t_0 \leq \min(t_0^c, t_0^b, t_0^p)$, ensuring that the initial latent conditions precede all measurement and prediction times, potentially extending to a point before the beginning of the time window ΔT. Further discussion on the selection of t_0 is provided in Section 5.1.5.

Post-ODE Latent Fusion

Once the CSI and beam-SNR latent states are temporally synchronized at the prediction time points $t_{nn=1}^{p\,N_p}$ via Equations (5.20) and (5.21), they are combined to produce a unified latent representation $\mathbf{z}_n^p \in \mathbb{R}^{L_p}$ with dimensionality L_p, which serves as the input for subsequent coordinate regression. We next introduce three different strategies for performing this multi-band latent fusion.

MLP Fusion: As depicted in Figure 5.5(a), the MLP-based fusion approach first maps the temporally aligned latent vectors $\mathbf{z}_n^{p_c}$ and $\mathbf{z}_n^{p_b}$ into an expanded feature space of dimension L_f using two independent MLPs. The resulting embeddings are then concatenated and passed through another MLP to obtain the fused latent representation $\mathbf{z}_n^p$ as:

$$\mathbf{z}_n^p = \mathcal{M}\left(\mathcal{M}(\mathbf{z}_n^{p_b}; \boldsymbol{\theta}_f^b) \oplus \mathcal{M}(\mathbf{z}_n^{p_c}; \boldsymbol{\theta}_f^c); \boldsymbol{\theta}_f^p\right), \tag{5.24}$$



(a) MLP Fusion

(b) Pairwise Interaction (PI) Fusion

(c) Weighted Importance (WI) Fusion

Figure 5.5: Post-ODE fusion schemes considered in this work.

where $\boldsymbol{\theta}_f^b$, $\boldsymbol{\theta}_f^c$, and $\boldsymbol{\theta}_f^p$ correspond to the trainable parameters of the three MLP modules, $\oplus$ indicates the concatenation operation, and the resulting fused latent vector $\mathbf{z}_n^p$ has dimensionality L_p.

Pairwise Interaction Fusion: The two time-aligned latent vectors, $\mathbf{z}_n^{p_c}$ and $\mathbf{z}_n^{p_b}$, are fused by augmenting them with an explicit interaction term given by their Kronecker product $\mathbf{z}_n^{p_b} \otimes \mathbf{z}_n^{p_c} \in \mathbb{R}^{L_b L_c \times 1}$. These components are concatenated and passed to an MLP to produce the fused representation,

$$\mathbf{z}_n^p = \mathcal{M}\left(\mathbf{z}_n^{p_b} \oplus \mathbf{z}_n^{p_c} \oplus (\mathbf{z}_n^{p_b} \otimes \mathbf{z}_n^{p_c}); \boldsymbol{\theta}_f^p\right), \tag{5.25}$$

where $\otimes$ denotes the Kronecker product and $\boldsymbol{\theta}_f^p$ are the learnable parameters of the fusion MLP. By introducing the Kronecker term, this fusion scheme explicitly captures cross-modal interactions through all pairwise multiplicative combinations of latent dimensions, expanding the representation from L_b or L_c to $L_b L_c$. A schematic overview of this fusion strategy is shown in Figure 5.5 (b).

Weighted Importance Fusion: We also consider a weighted fusion strategy in which the two aligned latent representations are merged according to modality-specific importance scores inferred from their initial latent conditions. An overview of this mechanism is shown in Figure 5.5 (c).

Concretely, the initial latent variables $\mathbf{z}_0^b$ and $\mathbf{z}_0^c$ are first projected into importance vectors of dimension L_f:

$$\mathbf{w}_b = \mathcal{M}(\mathbf{z}_0^b; \boldsymbol{\theta}_w^b) \in \mathbb{R}^{L_f}, \mathbf{w}_c = \mathcal{M}(\mathbf{z}_0^c; \boldsymbol{\theta}_w^c) \in \mathbb{R}^{L_f}, \tag{5.26}$$

where $\boldsymbol{\theta}_w^b$ and $\boldsymbol{\theta}_w^c$ denote learnable parameters. These vectors are then normalized using a row-wise softmax applied to $[\mathbf{w}_b, \mathbf{w}_c] \in \mathbb{R}^{L_f \times 2}$:

$$[\tilde{\mathbf{w}}_b, \tilde{\mathbf{w}}_c] \triangleq \sigma([\mathbf{w}_b, \mathbf{w}_c]) \longrightarrow \tilde{\mathbf{w}}_b + \tilde{\mathbf{w}}_c = \mathbf{1}_{L_f} \tag{5.27}$$

where $\sigma(\cdot)$ denotes the softmax for importance weight normalization and $\mathbf{1}_{L_f}$ is the all-one vector of dimension L_f. In parallel, the aligned latent states $\mathbf{z}_n^{p_b}$ and $\mathbf{z}_n^{p_c}$ are mapped to the same L_f-dimensional space and combined through element-wise weighting:

$$\mathbf{z}_n^p = \mathcal{M}\left([\mathcal{M}(\mathbf{z}_n^{p_b}; \boldsymbol{\theta}_f^b) \odot \tilde{\mathbf{w}}_b] + [\mathcal{M}(\mathbf{z}_n^{p_c}; \boldsymbol{\theta}_f^c) \odot \tilde{\mathbf{w}}_c]; \boldsymbol{\theta}_f^p\right), \tag{5.28}$$

where $\boldsymbol{\theta}_f^b$, $\boldsymbol{\theta}_f^c$, and $\boldsymbol{\theta}_f^p$ are the parameters of the fusion networks.

Decoders for Trajectory Estimation and Input Sequence Reconstruction

As illustrated in Figure 5.2, the NDF architecture incorporates three decoder modules: a coordinate decoder for trajectory regression, along with two additional decoders dedicated to reconstructing the CSI embedding sequence and the beam SNR sequence, respectively.

Trajectory Decoding: Given the fused latent representation $\mathbf{z}_n^p$ at the target time points t_n^p, an MLP decoder $\mathcal{M}$ with parameters $\boldsymbol{\theta}_p$ is used to predict the corresponding trajectory coordinates

$$\hat{\mathbf{p}}_n = \mathcal{M}(\mathbf{z}_n^p; \boldsymbol{\theta}_p), \quad n = 0, \cdots, N_p, \tag{5.29}$$

where $\hat{\mathbf{p}}_n = [\hat{x}_n, \hat{y}_n]^\top$ denotes the estimated position at time t_n^p. The decoder parameters $\boldsymbol{\theta}_p$ are shared across all queried time instances.

By combining the latent dynamic propagation defined in Equations (5.20) and (5.21), the post-ODE fusion mechanism (Equations (5.24), (5.25), or (5.28)), and the coordinate decoding operation above, we obtain an **integrated trajectory decoder** $\mathcal{P}(\cdot)$. This decoder maps the initial latent conditions $\mathbf{z}_0^c$ and $\mathbf{z}_0^b$ directly to the trajectory estimate at time t_n^p:

$$\hat{\mathbf{p}}_n = \mathcal{P}(\mathbf{z}_0^c, \mathbf{z}_0^b, t_0, t_n^p; \boldsymbol{\theta}_{dp}), \tag{5.30}$$

where $\boldsymbol{\theta}_{dp} = \{\boldsymbol{\theta}_d^c, \boldsymbol{\theta}_d^b, \boldsymbol{\theta}_f, \boldsymbol{\theta}_p\}$ and $\boldsymbol{\theta}_f$ collects all learnable parameters of the fusion module. For example, $\boldsymbol{\theta}_f = \{\boldsymbol{\theta}_f^c, \boldsymbol{\theta}_f^b, \boldsymbol{\theta}_f^p\}$ in the case of MLP fusion, while additional weight parameters are included for the weighted importance fusion scheme. As illustrated in Figure 5.6, this integrated formulation establishes a direct connection between the initial latent variables and the predicted coordinates, thereby streamlining the formulation of the ELBO-based objective introduced in the following section. The full set of estimated positions is denoted by $\hat{\mathbf{p}} = \{\hat{\mathbf{p}}_n\}_{n=0}^{N_p}$.

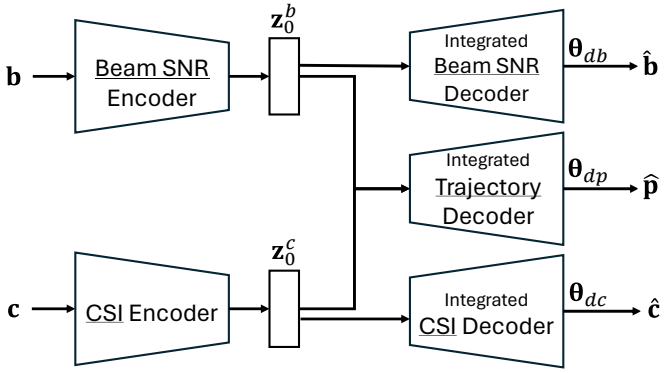


Figure 5.6: An equivalent multi-encoder, multi-decoder NDF structure with multiple integrated decoders directly connecting the two initial latent conditions ($\mathbf{z}_0^c$ and $\mathbf{z}_0^b$) to estimated coordinates $\hat{\mathbf{p}}_n$ and reconstructed input sequences $\hat{\mathbf{b}}_n$ and $\hat{\mathbf{c}}_n$.

CSI Decoding: Given the CSI latent states $\mathbf{z}_n^c$ at their original sampling times t_n^c as defined in Equation (5.22), an MLP decoder parameterized by $\boldsymbol{\theta}_c$ maps each latent state back to the CSI embedding domain:

$$\hat{\mathbf{c}}_n = \mathcal{M}(\mathbf{z}_n^c; \boldsymbol{\theta}_c), \quad n = 0, \cdots, N_c, \tag{5.31}$$

where the decoder parameters $\boldsymbol{\theta}_c$ are shared across all CSI time instances.

In a manner analogous to the integrated trajectory decoder $\mathcal{P}(\cdot)$, we combine the latent dynamic propagation in Equation (5.20) with the CSI embedding decoder above to form an **integrated CSI decoder** $\mathcal{C}(\cdot)$

$$\hat{\mathbf{c}}_n = \mathcal{C}(\mathbf{z}_0^c, t_0, t_n^c; \boldsymbol{\theta}_{dc}), \tag{5.32}$$

where $\boldsymbol{\theta}_{dc} = \{\boldsymbol{\theta}_d^c, \boldsymbol{\theta}_c\}$. This integrated decoder directly reconstructs the CSI embedding sequence at times t_n^c from the initial latent condition $\mathbf{z}_0^c$ associated with the CSI input. The reconstructed CSI embeddings are collectively denoted as $\hat{\mathbf{c}} = \{\hat{\mathbf{c}}_n\}_{n=0}^{N_c}$.

Beam SNR Decoding: The final decoder maps the latent states $\mathbf{z}_n^b$ defined in Equation (5.23) at the beam SNR sampling times t_n^b back to the beam SNR domain

$$\hat{\mathbf{b}}_n = \mathcal{M}(\mathbf{z}_n^b; \boldsymbol{\theta}_b), \quad n = 0, \cdots, N_b, \tag{5.33}$$

where the decoder parameters $\boldsymbol{\theta}_b$ are shared across all beam SNR time instances.

By integrating the latent dynamic propagation in Equation (5.21) with the beam SNR decoder in Equation (5.33), we construct the **integrated beam SNR decoder** $\mathcal{B}(\cdot)$

$$\hat{\mathbf{b}}_n = \mathcal{B}(\mathbf{z}_0^b, t_0, t_n^b; \boldsymbol{\theta}_{db}), \tag{5.34}$$

where $\boldsymbol{\theta}_{db} = \{\boldsymbol{\theta}_d^b, \boldsymbol{\theta}_b\}$. The reconstructed beam SNR sequence is denoted by $\hat{\mathbf{b}} = \{\hat{\mathbf{b}}_n\}_{n=0}^{N_b}$. The inclusion of both CSI and beam SNR reconstruction objectives encourages the latent ODE dynamics to remain tightly coupled to the underlying multi-band wireless propagation behavior. In contrast, when training relies solely on a coordinate regression decoder, the latent dynamics and the coordinate decoder may each converge to parameterizations that are weakly constrained by the physical characteristics of the wireless channels.

Remark: Employing ODEs in both the encoders and the latent dynamic modules offers benefits that extend beyond handling asynchronously sampled signals; it also enables principled alignment of heterogeneous wireless modalities while explicitly modeling their temporal evolution. In the proposed end-to-end framework, the continuous-time dynamics of CSI and beam SNR are embedded into latent distributions via separate ODE-based encoders. Both encoders are constrained to map their respective input sequences to initial latent states that follow a Gaussian distribution, thereby projecting CSI- and beam-SNR-derived representations into a shared latent space. By parameterizing the ODEs with neural networks, the encoders gain sufficient flexibility to satisfy this constraint, as neural ODEs can represent complex state evolution between observations. Moreover, the framework infers aligned initial latent states at a time preceding the first samples of both sequences, allowing the latent dynamic ODEs to propagate trajectories from a common temporal origin. This capability constitutes a key advantage of the proposed NDF framework over conventional RNN-based and RNN-variant approaches, in which hidden states are typically propagated using simple recurrence or decay mechanisms that offer limited expressiveness for capturing continuous-time dynamics and cross-modal alignment.

Complexity Analysis

The computational complexity of the proposed NDF framework can be analyzed by examining the cost of each architectural component. First, the CSI encoder employs recurrent layers whose comlpexity depends on the number of hidden units H_e and the length of the CSI sequence T_c, yielding an approximate complexity of $\mathcal{O}(H_e^2 T_c)$. Likewise, the beam SNR encoder incurs a complexity of $\mathcal{O}(H_e^2 T_b)$, where T_b denotes the number of time steps in the beam SNR sequence. For latent dynamic learning, the CSI-related ODE module propagates latent states over T_c steps for CSI reconstruction and T_p steps for trajectory estimation, resulting in a complexity of $\mathcal{O}(T_c + T_p)$. Similarly, the beam SNR latent ODE contributes a complexity of $\mathcal{O}(T_b + T_p)$. The decoder modules further add to the overall cost. The CSI decoder, beam SNR decoder, and coordinate decoder have respective complexities of $\mathcal{O}(T_c(L+1)C)$, $\mathcal{O}(T_b(L+1)B)$, and $\mathcal{O}(T_p(L+1))$, respectively, where L denotes the latent dimension, and C, B represents the dimension of the reconstructed CSI and beam SNR. By summing the contributions of all components, the overall time complexity of the proposed NDF is given by $\mathcal{O}((H_e^2(T_b + T_c)) + (T_b + T_c + T_p) + ((L+1)(T_c C + T_b B + T_p)))$.

ELBO-Based Loss Function

In the following, we derive an ELBO-based loss function tailored to the multi-encoder, multi-decoder structure of the proposed NDF framework. As illustrated in Figure 5.6, we collectively denote the beam SNR sequence, CSI embedding sequence, and trajectory coordinates as $\mathbf{b} = \{\mathbf{b}_n\}_{n=0}^{N_b}$, $\mathbf{c} = \{\mathbf{c}\}_{n}{}_{n=0}^{N_c}$, and $\mathbf{p} = \{\mathbf{p}\}_{n}{}_{n=0}^{N_p}$, respectively. Based on this notation, a modified ELBO objective can be formulated following the variational autoencoder framework [236]

$$\begin{aligned}
\text{ELBO} =& \mathbb{E}_{q(\mathbf{z}_0^b, \mathbf{z}_0^c|\mathbf{b},\mathbf{c})}[\log p(\mathbf{p}|\mathbf{z}_0^b, \mathbf{z}_0^c)] \\
&+ \lambda_1 \mathbb{E}_{q(\mathbf{z}_0^b|\mathbf{b})}[\log p(\mathbf{b}|\mathbf{z}_0^b)] + \lambda_2 \mathbb{E}_{q(\mathbf{z}_0^c|\mathbf{c})}[\log p(\mathbf{c}|\mathbf{z}_0^c)] \\
&- \lambda_3 D_{\text{KL}}[q(\mathbf{z}_0^b, \mathbf{z}_0^c|\mathbf{b}, \mathbf{c})||p(\mathbf{z}_0^b, \mathbf{z}_0^c)] \\
\overset{(a)}{\approx}& \frac{1}{V}\sum_{v=1}^{V} \log p(\mathbf{p}|\mathbf{z}_0^{b(v)}\mathbf{z}_0^{c(v)}) \\
&+ \frac{\lambda_1}{V}\sum_{v=1}^{V} \log p(\mathbf{b}|\mathbf{z}_0^{b(v)}) + \frac{\lambda_2}{V}\sum_{v=1}^{V} \log p(\mathbf{c}|\mathbf{z}_0^{c(v)}) \\
&- \lambda_3 D_{\text{KL}}[q(\mathbf{z}_0^b, \mathbf{z}_0^c|\mathbf{b}, \mathbf{c})||p(\mathbf{z}_0^b, \mathbf{z}_0^c)],
\end{aligned} \tag{5.35}$$

where $q(\mathbf{z}_0^c|\mathbf{c})$ and $q(\mathbf{z}_0^b|\mathbf{b})$ represent the approximate posterior distributions defined in Equations (5.15) and (5.18), respectively. The joint posterior distribution over the initial variables $\mathbf{z}_0^c$ and $\mathbf{z}_0^b$ can be factorized as

$$q(\mathbf{z}_0^b, \mathbf{z}_0^c|\mathbf{b}, \mathbf{c}) = q(\mathbf{z}_0^c|\mathbf{c})q(\mathbf{z}_0^b|\mathbf{b}), \tag{5.36}$$

under the assumption that the two input sequences are independent and are processed by separate encoders. $\{\lambda_i\}_{i=1}^3$ denote regularization weights, and $p(\mathbf{z}_0^b, \mathbf{z}_0^c)$ are the joint prior distribution of the initial latent variables $\mathbf{z}_0^c$ and $\mathbf{z}_0^b$, which can likewise be factorized as

$$p(\mathbf{z}_0^b, \mathbf{z}_0^c) = p(\mathbf{z}_0^b)p(\mathbf{z}_0^c), \tag{5.37}$$

where $p(\mathbf{z}_0^b) \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_{L_b})$ and $p(\mathbf{z}_0^c) \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_{L_c})$. Furthermore, $p(\mathbf{p}|\mathbf{z}_0^b, \mathbf{z}_0^c)$, $p(\mathbf{c}|\mathbf{z}_0^c)$ and $p(\mathbf{b}|\mathbf{z}_0^b)$ correspond to the output likelihood functions of the three integrated (trajectory/CSI/beam SNR) decoders, illustrated in Figure 5.6. In the above formulation, step (a) is obtained by approximating the posterior expectation with a Monte Carlo sample average over V realizations of the initial latent variables $\mathbf{z}_0^c$ and $\mathbf{z}_0^b$. These samples are generated via the reparameterization procedures in Equations (5.17) and (5.19), using V independent draws of $\boldsymbol{\epsilon}^c$ and $\boldsymbol{\epsilon}^b$, respectively. In practice, we set $V = 1$, since the expectation can be effectively approximated by averaging over independent realizations across minibatch samples.

For the KL divergence term $D_{\text{KL}}[q(\mathbf{z}_0^b, \mathbf{z}_0^c|\mathbf{b}, \mathbf{c})||p(\mathbf{z}_0^b, \mathbf{z}_0^c)]$, we exploit the independence assumptions that (i) the posterior distributions of $\mathbf{z}_0^b$ and $\mathbf{z}_0^c$ factorize given the input sequences $\mathbf{b}$ and

$\mathbf{c}$, and (ii) their prior distributions are also independent. Under these assumptions, the joint KL divergence decomposes as

$$\begin{aligned}
&D_{\mathrm{KL}}[q(\mathbf{z}_0^b, \mathbf{z}_0^c|\mathbf{b}, \mathbf{c})||p(\mathbf{z}_0^b, \mathbf{z}_0^c)] \\
&\overset{(a)}{=} D_{\mathrm{KL}}[q(\mathbf{z}_0^b|\mathbf{b})q(\mathbf{z}_0^c|\mathbf{c})||p(\mathbf{z}_0^b)p(\mathbf{z}_0^c)] \\
&\overset{(b)}{=} D_{\mathrm{KL}}[q(\mathbf{z}_0^b|\mathbf{b})||p(\mathbf{z}_0^b)] + D_{\mathrm{KL}}[q(\mathbf{z}_0^c|\mathbf{c})||p(\mathbf{z}_0^c)],
\end{aligned} \tag{5.38}$$

where step (a) follows from the factorization of the posterior and prior in Equations (5.36) and (5.37). Step (b) is obtained by applying the standard decomposition property of the KL divergence for independent random variables, as

$$\begin{aligned}
&D_{\mathrm{KL}}[q(\mathbf{z}_x, \mathbf{z}_y|\mathbf{x}, \mathbf{y})||p(\mathbf{z}_x, \mathbf{z}_y)] \\
&= \int\int q(\mathbf{z}_x|\mathbf{x})q(\mathbf{z}_y|\mathbf{y}) \left(\log \frac{q(\mathbf{z}_x|\mathbf{x})}{p(\mathbf{z}_x)} + \log \frac{q(\mathbf{z}_y|\mathbf{y})}{p(\mathbf{z}_y)} \right) d\mathbf{z}_x d\mathbf{z}_y \\
&= \underbrace{\int q(\mathbf{z}_y|\mathbf{y}) d\mathbf{z}_y}_{=1} \int q(\mathbf{z}_x|\mathbf{x}) \log \frac{q(\mathbf{z}_x|\mathbf{x})}{p(\mathbf{z}_x)} d\mathbf{z}_x \\
&\quad + \underbrace{\int q(\mathbf{z}_x|\mathbf{x}) d\mathbf{z}_x}_{=1} \int q(\mathbf{z}_y|\mathbf{y}) \log \frac{q(\mathbf{z}_y|\mathbf{y})}{p(\mathbf{z}_y)} d\mathbf{z}_y \\
&= D_{\mathrm{KL}}[q(\mathbf{z}_x|\mathbf{x})||p(\mathbf{z}_x)] + D_{\mathrm{KL}}[q(\mathbf{z}_y|\mathbf{y})||p(\mathbf{z}_y)].
\end{aligned} \tag{5.39}$$

It can be readily derived that

$$\begin{aligned}
D_{\mathrm{KL}}[q(\mathbf{z}_0^c|\mathbf{c})||p(\mathbf{z}_0^c)] &= D_{\mathrm{KL}}[\mathcal{N}(\boldsymbol{\mu}^c, \boldsymbol{\sigma}^c)||\mathcal{N}(\mathbf{0}, \mathbf{I}_{L_c})] \\
&= \frac{1}{2} \sum_{l=1}^{L_c} \left((\mu_l^c)^2 + (\sigma_l^c)^2 - 1 - \log(\sigma_l^c)^2 \right),
\end{aligned} \tag{5.40}$$

$$\begin{aligned}
D_{\mathrm{KL}}[q(\mathbf{z}_0^b|\mathbf{b})||p(\mathbf{z}_0^b)] &= D_{\mathrm{KL}}[\mathcal{N}(\boldsymbol{\mu}^b, \boldsymbol{\sigma}^b)||\mathcal{N}(\mathbf{0}, \mathbf{I}_{L_b})] \\
&= \frac{1}{2} \sum_{l=1}^{L_b} \left((\mu_l^b)^2 + (\sigma_l^b)^2 - 1 - \log(\sigma_l^b)^2 \right),
\end{aligned} \tag{5.41}$$

where $\mu_l^{b/c}$ and $\sigma_l^{b/c}$ are the l-th element of $\boldsymbol{\mu}^{b/c}$ and $\boldsymbol{\sigma}^{b/c}$, respectively.

For the output log-likelihood terms, we first consider the integrated trajectory decoder $\mathcal{P}(\cdot)$, which maps the two initial latent conditions to trajectory coordinates at time instances t_n^p,

$$\begin{aligned}
\log p(\mathbf{p}|\mathbf{z}_0^b, \mathbf{z}_0^c) &= \log p(\mathbf{p}_0, \mathbf{p}_1, \cdots, \mathbf{p}_{N_p}|\mathbf{z}_0^b, \mathbf{z}_0^c) \\
&\overset{(a)}{\approx} \sum_{n=1}^{N_p} \log p(\mathbf{p}_n|\mathbf{z}_0^b, \mathbf{z}_0^c),
\end{aligned} \tag{5.42}$$

where approximation (a) assumes conditional independence among the coordinate outputs at different time instances, enabling a tractable likelihood expression. Each coordinate component of

$\mathbf{p}_n = [x_n, y_n]^\top$ is modeled by a Laplace distribution:

$$p(x_n|\mathbf{z}_0^b, \mathbf{z}_0^c) = \frac{1}{2b_p} \exp\left(-\frac{|x_n - \hat{x}_n|}{b_p}\right),$$
$$p(y_n|\mathbf{z}_0^b, \mathbf{z}_0^c) = \frac{1}{2b_p} \exp\left(-\frac{|y_n - \hat{y}_n|}{b_p}\right), \tag{5.43}$$

where $b_p \in \mathbb{R}$ is a scaling parameter and $\hat{\mathbf{p}}_n = [\hat{x}_n, \hat{y}_n]^\top = \mathcal{P}(\mathbf{z}_0^c, \mathbf{z}_0^b, t_0, t_n^p; \boldsymbol{\theta}_{dp})$ denotes the estimated position. Under this model, the conditional log-likelihood satisfies

$$\log p(\mathbf{p}_n|\mathbf{z}_0^b, \mathbf{z}_0^c) \propto -\frac{\|\mathbf{p}_n - \hat{\mathbf{p}}_n\|_1}{b_p}, \tag{5.44}$$

where $\|\cdot\|_1$ is the ℓ_1 norm. Setting $b_p = 1$ and substituting into Equation (5.42) yields

$$\log p(\mathbf{p}|\mathbf{z}_0^b, \mathbf{z}_0^c) \propto -\sum_{n=1}^{N_p} \|\mathbf{p}_n - \hat{\mathbf{p}}_n\|_1. \tag{5.45}$$

Thus, maximizing this log-likelihood is equivalent to minimizing the mean absolute error (MAE) between the ground truth and estimated trajectories. By applying the same reasoning, the output log-likelihood functions of the integrated beam SNR and CSI decoders can be expressed as,

$$\log p(\mathbf{b}|\mathbf{z}_0^b) \propto -\sum_{n=1}^{N_b} \|\mathbf{b}_n - \hat{\mathbf{b}}_n\|_1, \tag{5.46}$$

$$\log p(\mathbf{c}|\mathbf{z}_0^c) \propto -\sum_{n=1}^{N_c} \|\mathbf{c}_n - \hat{\mathbf{c}}_n\|_1. \tag{5.47}$$

By combining the KL divergence terms with the output log-likelihoods of the three integrated decoders, the modified ELBO in Equation (5.35) can be expressed as the following overall training objective:

$$\begin{aligned}\mathcal{L} = &\sum_{n=1}^{N_p} \|\mathbf{p}_n - \hat{\mathbf{p}}_n\|_1 + \lambda_1 \sum_{n=1}^{N_b} \|\mathbf{b}_n - \hat{\mathbf{b}}_n\|_1 + \lambda_2 \sum_{n=1}^{N_c} \|\mathbf{c}_n - \hat{\mathbf{c}}_n\|_1 \\ &+\lambda_3 \sum_{l=1}^{L_b} \left((\mu_l^b)^2 + (\sigma_l^b)^2 - 1 - \log(\sigma_l^b)^2\right) \\ &+\lambda_4 \sum_{l=1}^{L_c} \left((\mu_l^c)^2 + (\sigma_l^c)^2 - 1 - \log(\sigma_l^c)^2\right)\end{aligned} \tag{5.48}$$

where the first term corresponds to the trajectory estimation error, the second and third terms penalize reconstruction errors of the beam SNR and CSI embeddings, and the last two terms regularize the posterior distributions of the beam SNR and CSI latent variables via their respective KL divergences. Here, the regularization weight originally associated with the joint KL term is decoupled into two independent weights, λ_3 and λ_4, to separately control the contribution of the beam SNR and CSI latent regularization.

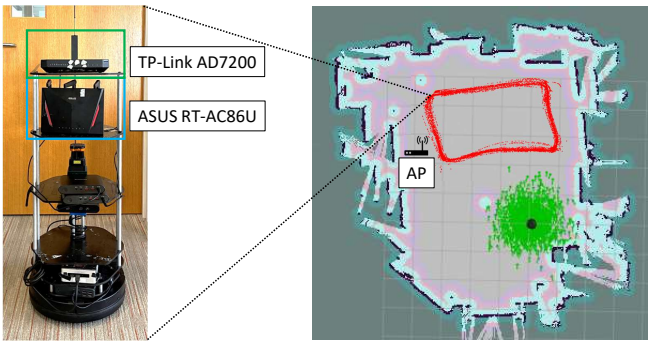


Figure 5.7: In-house multi-band Wi-Fi testbed: a TurtleBot equipped with a TP-Link router for 60 GHz beam SNR and an ASUS router for 5 GHz CSI (left), along with the data collection floorplan showing the AP location and trajectories (right).

5.1.5 Performance Evaluation

In-House Testbed and Data Collection

The in-house testbed introduced in [72] is extended to support simultaneous acquisition of both 5 GHz CSI and 60 GHz beam SNRs. As shown in Figure 5.7, a TurtleBot is equipped with two commercial Wi-Fi routers: an 802.11ac-compliant ASUS RT-AC86U device for collecting 5 GHz over an 80 MHz channel, and an 802.11ad-compliant TP-Link Talon AD7200 router for acquiring 60 GHz beam SNR. The TurtleBot traverses predefined rectangular trajectories (illustrated by red dotted lines in the right panel of Figure 5.7) within a large conference room, while an identical pair of routers placed at the lower-left corner of the trajectory serves as a multi-band AP.

To enable low-level channel measurements on these commercial routers, the original firmware is replaced with open-source alternatives [151, 237], and beam SNR and CSI are extracted following the procedures in [238] and [151]. For the ASUS router, four transmit antennas (three external and one internal) and two receive antennas are utilized, yielding $N_{\mathrm{Tx}} \times N_{\mathrm{Rx}} = 4 \times 2$ spatial streams across $N_{\mathrm{s}} = 234$ non-null subcarriers. Each raw CSI frame $\mathbf{C}_n \in \mathbb{C}^{4\times 2\times 234}$ is calibrated and compressed into a CSI embedding input $\mathbf{c}_n \in \mathbb{R}^{36\times 1}$ with $M_c = 36$. In contrast, the TP-Link router employs a 32-element analog phased array and sequentially probes $M_b = 36$ predefined beam directions, producing beam SNR vectors $\mathbf{b}_n \in \mathbb{R}^{36\times 1}$.

The testbed is further augmented with a LiDAR sensor and wheel encoders, enabling accurate self-localization on a predefined map. The resulting position estimates, recorded at a frame rate of 10 fps, are used as ground truth trajectory labels $\mathbf{p}_n$. To ensure temporal consistency, the system clocks of all devices—including the Wi-Fi routers and the TurtleBot—are synchronized using the Network Time Protocol (NTP), with a central desktop serving as the NTP server. This desktop coordinates and aligns the clocks of all connected devices so that timestamps across the system share a common time reference. Through this setup, we collected 43,237 CSI frames with corre-

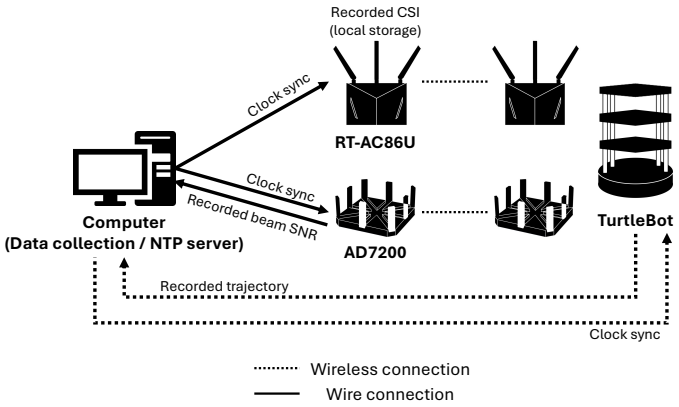


Figure 5.8: The device connection and data flow within the testbed.

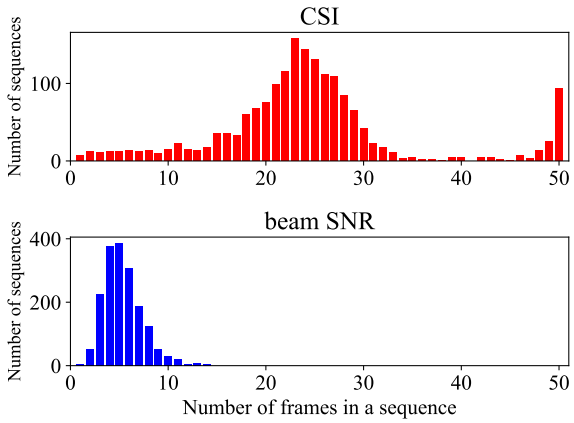


Figure 5.9: The histogram of the number of CSI (top) and beam SNR (bottom) across non-overlapping sequences.

sponding position labels and 9,590 beam SNR frames. A summary of the dataset characteristics and evaluation settings is provided in Table 5.1.

Implementation

We set $\Delta T_w = 5$ seconds to segment the collected CSI frames, beam SNR measurements, and coordinate labels into fixed-length sequences. Figure 5.9 presents histograms of the number of CSI frames (top) and beam SNR frames (bottom) within each non-overlapping 5-second window. Most CSI sequences contain approximately 20–40 frames per window, corresponding to an average sampling rate of 4–8 fps. By comparison, beam SNR sequences include substantially fewer samples within the same duration, yielding an average rate of about 1 fps. For each sequence, all timestamps $\{t_n^b\}_{n=0}^{N_b}$, $\{t_n^c\}_{n=0}^{N_c}$, and $\{t_n^p\}_{n=0}^{N_p}$ are normalized to the interval $[0, 1]$ by scaling relative times with respect to the 5-second window length.

We adopt three data splitting strategies for evaluation:

1. **Random Split**: All frames are grouped into 1,778 non-overlapping 5-second sequences (step

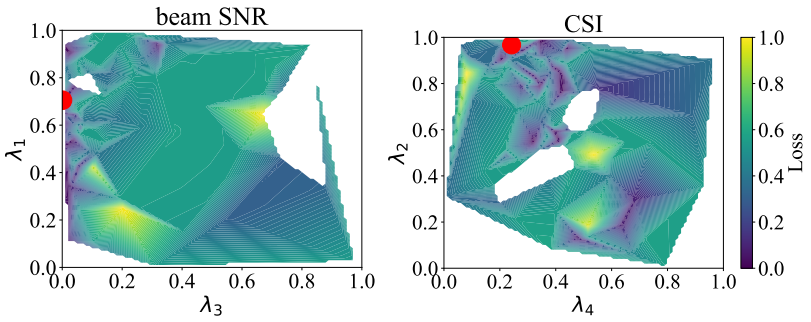
Table 5.1: Summary of experimental parameters.

	Parameter	Notation	Value
Device	**CSI**		
	- Wireless Standard	-	IEEE 802.11ac
	- Center Frequency	-	5 GHz
	- Bandwidth	-	80 MHz
	- Tx antenna	N_{Tx}	4
	- Rx antenna	N_{Rx}	2
	- Subcarriers	N_{s}	234
	beam SNR		
	- Wireless Standard	-	IEEE 802.11ad
	- Center Frequency	-	60 GHz
	- Bandwidth	-	2.16 GHz
	- Tx antenna	-	32
	- Rx antenna	-	32
	- Beam sectors	N_b	36
Dataset	**Total path length**	-	2178 m
	# of samples		
	- CSI	-	43,237
	- beam SNR	-	9,590
	Sampling rate		
	- CSI	-	7.44 Hz
	- beam SNR	-	1.71 Hz
Training	**CSI embedding dimension**	M_c	36
	GRU hidden dimension	H_b, H_c	20
	Latent dimension	L_b, L_c, L_p	20
	Fusion latent dimension	L_f	128
	ODE solver		
	- Encoder	$\mathcal{O}_e$	Euler solver
	- Latent dynamic ODE	$\mathcal{O}_d$	Dopri5 solver
	Loss coefficients	$\lambda_1, \lambda_2, \lambda_3, \lambda_4$	0.7, 1.0, 0.0010, 0.25
	Epochs	-	250
	Batch size	-	32
	Optimizer	-	Adamax
	Learning rate	-	4e-3
	LR scheduler	-	OneCycle

size of 5 seconds), which are randomly partitioned into training, validation, and test sets with a ratio of 80:10:10. Unless otherwise stated, results in subsequent sections are reported under this split.

2. **Temporal Split**: Frames are divided into training and test sets strictly based on temporal order. The first s% of the data are used for training, while the remaining frames form the test set, ensuring that evaluation is performed on future, unseen data. This split is used to assess temporal generalization in the ablation study for different values of s. With a 5-second window, training and test sequences are constructed using step sizes of 1 second and 5 seconds, respectively.

3. **Coordinate Split**: Frames are partitioned according to their ground truth spatial locations.



(a) beam SNR (b) CSI

Figure 5.10: Hyperparameter search for regularization parameters in the ELBO-based loss function of Equation (5.48).

Data collected from a specified region (e.g., a corner area) are held out exclusively for testing and are not observed during training. This split evaluates spatial generalization to unseen coordinates in the ablation analysis.

We pretrain the CSI embedding vector $\mathbf{c}_n$ using an autoencoder composed of three one-dimensional convolutional layers followed by three MLP layers. Both the beam SNR sequences and the CSI embedding sequences are normalized to the range $[0, 1]$ prior to training. For the encoders, GRUs are adopted for both modalities, with hidden dimensions set to $H_b = H_c = 20$ for the beam SNR and CSI sequences, respectively. The dimensions of the initial latent conditions and the unrolled latent states are fixed to $L_b = L_c = 20$, yielding latent variables $\mathbf{z}_n^b$ and $\mathbf{z}_n^c$ for $n = 0, 1, \ldots, N_b/N_c$. After temporal alignment, the latent states are first projected to a higher-dimensional space of size $L_f = 128$ and then mapped back to a fused latent representation of dimension $L_p = 20$. We employ the Euler solver for the ODE-based encoding process and the Dopri5 solver for the latent dynamic ODE. The trajectory, CSI, and beam SNR decoders (Equations 5.29, 5.31, and 5.33) share the same three-layer MLP architecture.

The regularization coefficients in the loss function are determined through hyperparameter optimization using Optuna [239], searching over the interval $(0, 1]$. The search objective is guided by the validation loss of the coordinate estimation task, with 100 trials conducted for up to 125 training epochs each. Figure 5.10 visualizes the loss landscape with respect to the regularization parameters, where red markers indicate parameter combinations that achieve the lowest validation loss at 125 epochs or the lowest intermediate loss when early stopping occurs. Based on this procedure, the regularization parameters in Equation (5.48) are set to $\lambda_1 = 0.7$, $\lambda_2 = 1.0$, $\lambda_3 = 0.0010$, and $\lambda_4 = 0.25$. For training the NDF model, we use a minibatch size of 32 and train for up to 250 epochs, retaining the model checkpoint that attains the best validation loss. Optimization is performed using the Adamax optimizer with a maximum learning rate of 4×10^{-3}, together with a OneCycle learning rate schedule to accelerate convergence [240].

Evaluation Metrics

As a localization error metric, we adopt the Euclidean distance between the estimated coordinate $\hat{\mathbf{p}} = [\hat{x}, \hat{y}]$ and the ground truth $\mathbf{p} = [x, y]$. This distance is computed for each estimation and then averaged over all N estimations, yielding

$$L_{\text{error}} = \frac{1}{N} \sum_{i=1}^{N} \|\mathbf{p} - \hat{\mathbf{p}}\|. \tag{5.49}$$

In addition, we evaluate the distribution of localization errors by computing the CDF across different methods. The CDF of the localization error, denoted by $F_e(t)$, is defined as

$$F_e(t) = \frac{1}{N} \sum_{i=1}^{N} \mathbb{I}(e_i \leq t), \tag{5.50}$$

where t is an error threshold, $e_i = \|\mathbf{p}_i - \hat{\mathbf{p}}_i\|$ represents the localization error at the i-th estimate, and $\mathbb{I}(\cdot)$ is the indicator function given by

$$\mathbb{I}(e_i \leq t) = \begin{cases} 1, & e_i \leq t, \\ 0, & \text{otherwise.} \end{cases} \tag{5.51}$$

Comparison to Baseline Methods

For performance evaluation, we consider a comprehensive set of baseline methods:

- **Single-band (CSI)** methods: Trajectory regression formulated as $\{\mathbf{c}_n, t_n^c\}_{n=0}^{N_c} \longrightarrow \{\mathbf{p}_n, t_n^p\}_{n=0}^{N_p}$ using 1) Linear interpolation (LinearInt) of Equation (5.3); 2) Nearest interpolation (NearestInt) of Equation (5.4); 3) RNN-Decay of Equation (5.8); 4) RNN-Δ of Equation (5.9); 5) DDND of [72].

- **Single-band (beam SNR)** methods: Trajectory regression formulated as $\{\mathbf{b}_n, t_n^b\}_{n=0}^{N_b} \longrightarrow \{\mathbf{p}_n, t_n^p\}_{n=0}^{N_p}$ using 1) Linear interpolation (LinearInt) of Equation (5.3); 2) Nearest interpolation (NearestInt) of Equation (5.4); 3) RNN-Decay of Equation (5.8); 4) RNN-Δ of Equation (5.9); 5) DDND of [72].

- **Frame-to-Frame Fusion** methods: 1) LinearInt fusion of Equations (5.3) and (5.6); 2) NearestInt fusion of Equations (5.4) and (5.6).

- **Sequence-to-Sequence Fusion** methods: 1) RNN-Decay fusion of Equations (5.8), (5.10) and (5.12); 2) RNN-Δ fusion of Equations (5.9), (5.11) and (5.12).

Table 5.2 reports the trajectory estimation performance of all baseline methods and the proposed NDF under the random sequence split. From the mean, median, and the 90th percentile localization error measured in meters, for each method (e.g., the linear interpolation or the RNN-Decay), fusing multi-band measurements consistently outperforms using CSI-only or the beam

Table 5.2: Localization errors (meter) for all evaluated single-band and multi-band baselines and the proposed NDF methods.

	Single-band						**Multi-band**		
	CSI			**beam SNR**					
	Mean	Median	CDF@0.9	Mean	Median	CDF@0.9	Mean	Median	CDF@0.9
LinearInt	1.79	1.89	2.91	0.932	0.726	1.77	0.764	0.506	1.76
NearestInt	1.82	1.92	2.87	1.00	0.715	2.09	0.839	0.553	2.01
RNN-Decay	1.02	0.739	2.27	1.03	0.545	2.64	0.685	0.432	1.69
RNN-Δ	1.03	0.733	2.39	0.994	0.499	2.70	0.506	0.215	1.42
DDND	0.975	0.588	2.45	0.390	0.191	0.859	-	-	-
NDF (ours)	-	-	-	-	-	-	**0.263**	**0.148**	**0.611**

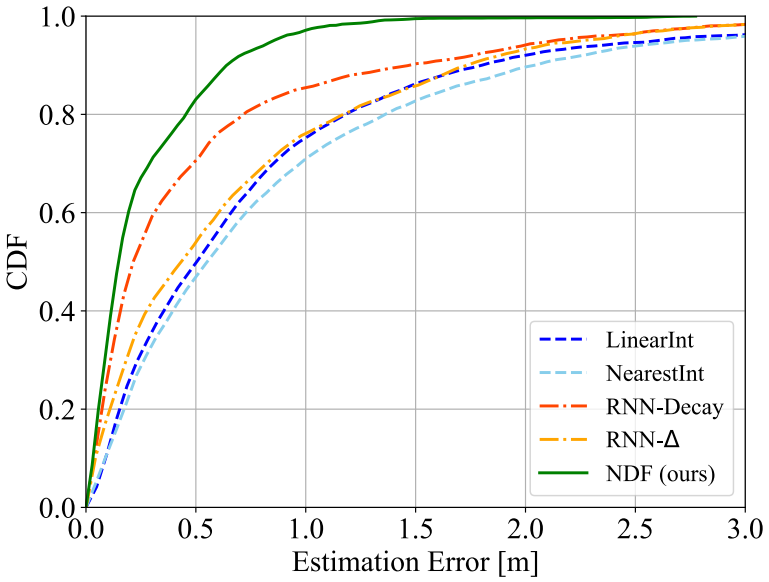
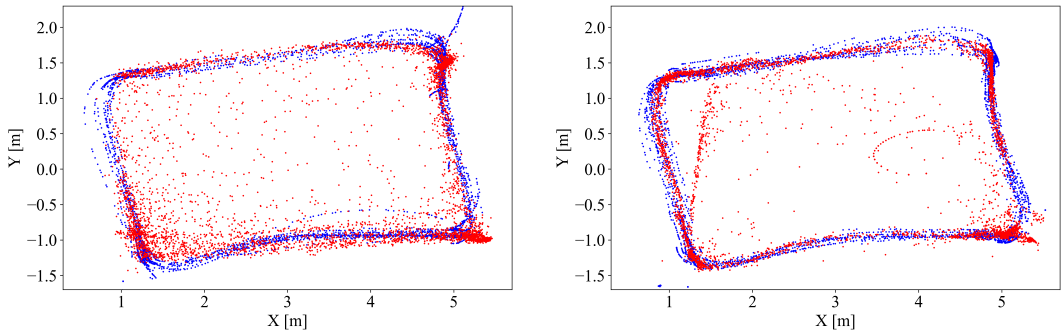


Figure 5.11: Cumulative distribution functions of localization errors for the baseline methods and NDF.

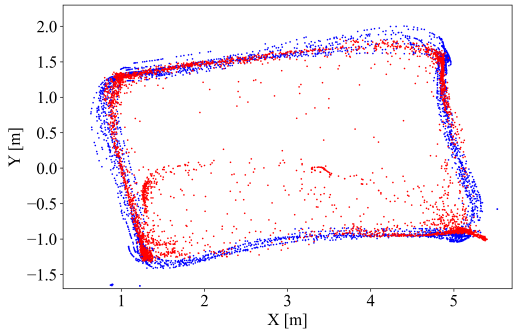
SNR-only inputs. Comparing interpolation-based approaches (linear and nearest) and RNN-based methods (RNN-Decay and RNN-Δ) further shows that temporal modeling via RNNs substantially enhances CSI-only performance and contributes to improved accuracy when combining CSI and beam SNR. As highlighted in the last column of Table 5.2, NDF achieves an additional reduction in localization error by explicitly aligning latent states through latent dynamic ODEs, improving upon the best multi-band baseline (RNN-Decay) to reach a mean error of 14.8 cm. Figure 5.11 illustrates the corresponding CDFs of the localization error for the multi-band baselines and NDF.

To qualitatively compare the baseline methods with the proposed approach, Figure 5.12 overlays the estimated trajectories (red dots) and the ground truth coordinates (dark blue dots) for representative multi-band fusion baselines—nearest interpolation, RNN-Decay, and RNN-Δ—as well as the proposed NDF. A clear progression is observed when moving from frame-to-frame fusion (nearest interpolation) to sequence-to-sequence fusion (RNN-Decay and RNN-Δ), with no-

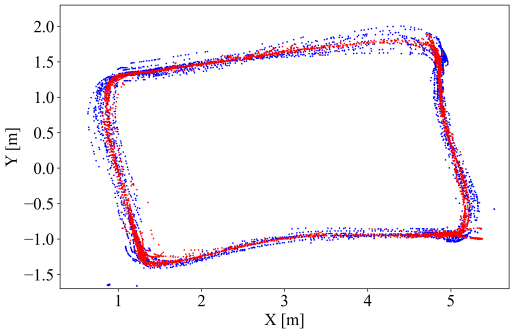


(a) NearestInt

(b) RNN-Decay



(c) RNN-Δ



(d) NDF

Figure 5.12: Estimated trajectories (red) overlaid with ground truth (blue) for representative multi-band baseline methods and the proposed NDF.

ticeably fewer localization errors near the center of the rectangular area. The remaining outliers in the baseline results stem from their limited ability to model temporal continuity, causing abrupt jumps between successive coordinate estimates. In contrast, NDF mitigates this issue through its architecture, which incorporates paired ODE modules to jointly capture the temporal evolution of wireless measurements and object motion. By conditioning trajectory estimates on learned continuous-time dynamics, NDF produces smoother and more physically plausible transitions between locations. As shown in Figure 5.12, NDF achieves the most consistent trajectories, markedly reducing outliers and closely following the rectangular path.

Impact of Sequence Length ΔT_w

We investigate the impact of the sequence length ΔT_w on trajectory estimation accuracy. Because CSI and beam SNR are sampled at approximately 5 Hz and 1 Hz, respectively, increasing ΔT_w effectively increases the number of observations available for modeling latent dynamics. Following the random split protocol, we segment the data into non-overlapping sequences of length $\Delta T_w \in 2, 8$ seconds.

Table 5.3 reports the mean, median, and 90th percentile localization errors for three choices of

Table 5.3: Impact of sequence length ΔT_w.

Sequence length	Mean	Median	CDF@0.9
2 seconds	0.481	0.233	1.20
5 seconds	**0.263**	0.148	0.611
8 seconds	0.270	**0.136**	**0.547**

Table 5.4: The impact of the three fusion schemes in Section 5.1.4 on the estimation error (m).

Fusion scheme	Mean	Median	CDF@0.9
MLP	**0.263**	**0.148**	**0.611**
Pairwise Interaction	0.397	0.230	0.903
Weighted Importance	0.287	0.164	0.618

ΔT_w. A consistent trend emerges: longer sequences yield better trajectory estimation performance. With $\Delta T_w = 2$ seconds, the sparse beam SNR measurements provide insufficient temporal context, leading to degraded accuracy. In contrast, $\Delta T_w = 8$ seconds achieves lower median and CDF@0.9 errors while maintaining a mean error comparable to that of $\Delta T_w = 5$ seconds, suggesting a favorable trade-off between temporal context and estimation robustness.

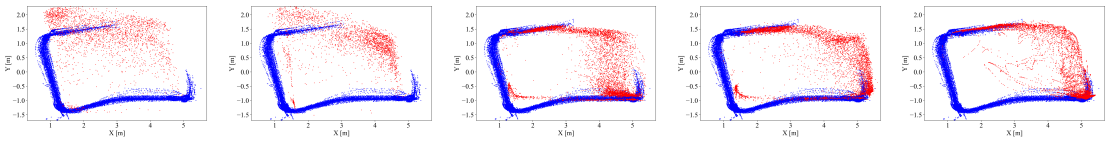
Impact of Fusion Scheme

We next examine the impact of the three fusion strategies introduced in Section 5.1.4 on trajectory estimation accuracy, using the random split protocol with non-overlapping 5-second input sequences. As summarized in Table 5.4, the MLP fusion achieves the best performance across the mean, median, and 90th percentile CDF metrics, while the weighted importance fusion attains nearly identical results. All three fusion strategies consistently outperform the multi-band baseline methods reported in Table 5.2. Overall, these results indicate that once the latent states of beam SNR and CSI are temporally aligned, the specific fusion mechanism has a limited effect on final localization accuracy.

During training, the signal reconstruction terms in Equation 5.48 act as regularizers for the latent dynamic ODE, encouraging it to capture the temporal characteristics of wireless propagation in both frequency bands. Starting from the initial latent states $\mathbf{z}_0^b$ and $\mathbf{z}_0^c$ for beam SNR and CSI, respectively, the latent dynamic ODE regresses corresponding sequences of latent variables at the queried timestamps. We posit that a well-regularized latent dynamic ODE can generate latent trajectories that are primarily conditioned on each band's initial state while implicitly encoding cross-band information. As a result, the choice of fusion scheme becomes less critical, which explains the comparable performance observed even with relatively simple fusion strategies such as pairwise interaction.

Table 5.5: Localization error (m) under temporal split as a function of training data ratio $s\%$ for multi-band cases.

	20%			**40%**			**60%**			**80%**		
	Mean	Median	CDF@0.9	Mean	Median	CDF@0.9	Mean	Median	CDF@0.9	Mean	Median	CDF@0.9
LinearInt	1.36	0.933	3.42	1.40	0.856	4.44	0.968	0.749	1.99	1.00	0.759	2.13
NearestInt	1.39	0.989	3.24	1.40	0.881	4.37	0.981	0.763	2.01	1.01	0.760	2.19
RNN-Decay	0.743	0.266	2.38	0.578	0.232	**1.65**	0.562	0.214	1.62	0.639	0.238	1.99
RNN-Δ	0.671	0.316	**1.88**	0.738	0.284	2.27	0.580	0.233	1.72	0.734	0.332	2.07
NDF (ours)	**0.647**	**0.250**	2.07	0.749	**0.183**	3.81	**0.454**	**0.169**	**0.981**	**0.501**	**0.183**	**1.17**



(a) LinearInt fusion (b) NearestInt fusion (c) RNN-Decay fusion (d) RNN-Δ fusion (e) NDF

Figure 5.13: Estimated trajectories in an unseen corner region are overlaid with the training trajectories. The degree to which the estimates cover the previously unobserved region of the rectangular layout indicates the model's capacity to extrapolate to unseen locations.

Generalization under Temporal Split

To assess generalization to future, temporally disjoint measurements, we adopt a **temporal split** protocol. This setting enables a direct evaluation of performance on Wi-Fi data collected after the training period. The proportion of training data s is varied as $s\% \in \{20\%, 40\%, 60\%, 80\%\}$, with the remaining data reserved for testing. As the available training data is limited, we improve data utilization by segmenting the training data into overlapping 5-second sequences with a stride of 1 second.

As summarized in Table 5.5, localization accuracy strongly depends on the amount of training data. Among the evaluated settings, a temporal split of $60 : 40$ yields the best performance across all metrics. When increasing the training ratio to $80 : 20$, NDF exhibits degraded accuracy compared to the random split of $80 : 10 : 10$ in Table 5.2. For instance, the mean localization error increases from 26.3 cm under the random split to 50.1 cm under the temporal split. This degradation is expected, as temporal splits expose the model to distribution shifts caused by time-varying channel conditions. Nevertheless, across all temporal split configurations, NDF consistently outperforms the baseline methods, demonstrating stronger robustness to temporal domain shifts.

Generalization under Coordinate Split

We further evaluate spatial generalization under the **coordinate split** setting, where trajectories originate from locations that are entirely unseen during training. In this evaluation, we adopt Fréchet distance as the evaluation metric to quantify how well the baseline methods and the proposed NDF reproduce both the shape and spatial placement of trajectories in the unseen corner

Table 5.6: Quantitative comparison of trajectory similarity at unseen locations using the Fréchet distance.

	Mean	Median	CDF@0.9
LinearInt	3.14	3.23	3.90
NearestInt	3.12	3.29	3.73
RNN-Decay	3.60	3.58	3.96
RNN-Δ	3.36	3.36	3.74
NDF (ours)	**2.55**	**2.57**	**2.97**

Table 5.7: Fréchet distance between curves with every 100 points in ground truth and estimations.

	Mean	Median	CDF@0.9
NearestInt	2.07	2.05	**2.67**
RNN-Decay	2.20	2.16	2.93
NDF (ours)	**1.92**	**1.90**	2.72

region. The Fréchet distance measures the similarity between two curves while accounting for both point-wise proximity and ordering along the curves. Formally, for two normalized curves $A = \{a_t\}_{t=0}^{1}$ and $B = \{b_t\}_{t=0}^{1}$, Fréchet distance $F(A, B)$ is defined as

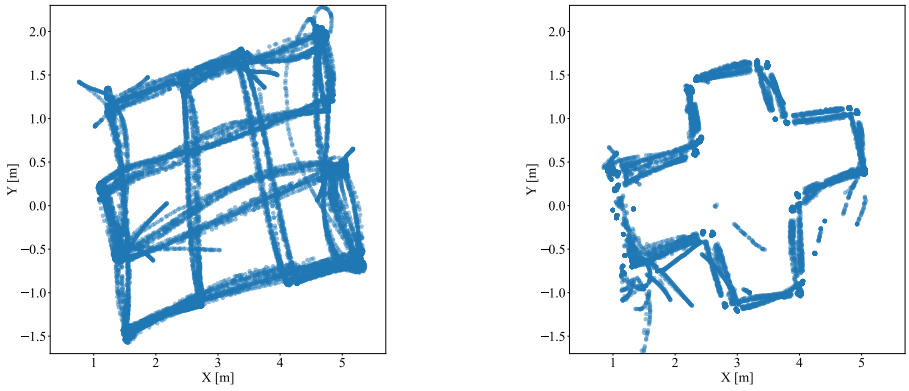
$$F(A, B) = \inf_{f:[0,1]\to[0,1]} \max_{t\in[0,1]} \|A(t) - B(f(t))\|, \tag{5.52}$$

where f is an arbitrary reparametrization function. Table 5.6 reports the localization errors for the unseen region, and Figure 5.13 visualizes the estimated trajectories (red dots) together with the ground truth trajectories in the upper-right corner, while the training trajectories are shown in blue. As illustrated in Figure 5.13 (a) and (b), frame-to-frame fusion baselines (LinearInt and NearestInt) fail to exploit temporal dynamics, leading to fragmented and spatially inconsistent trajectory estimates. Sequence-to-sequence baselines, namely RNN-Decay and RNN-Δ fusion shown in Figure 5.13 (c) and (d), yield better alignment with the training trajectories by capturing temporal dependencies. Figure 5.13 (e) demonstrates the performance of NDF, where the estimated trajectories effectively fill in the unseen corner and closely follow the rectangular structure of the ground truth, indicating substantially improved spatial extrapolation capability.

Training on a more complex trajectory

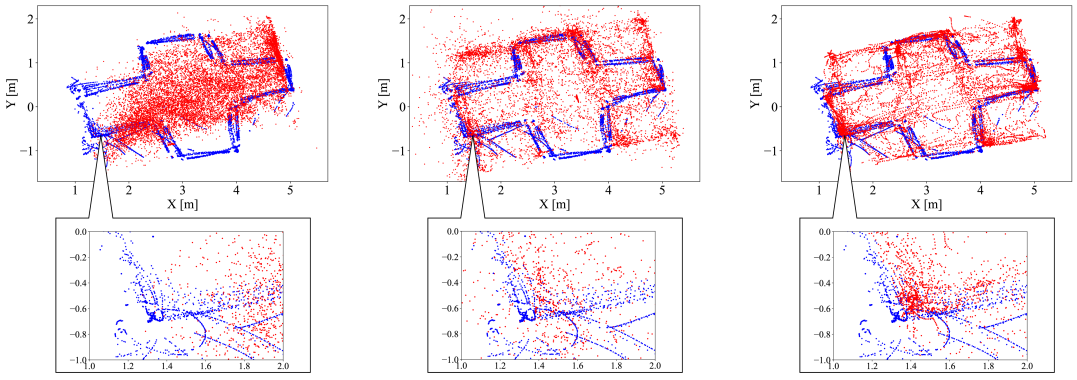
In this evaluation, the baseline methods and NDF are trained using multi-band Wi-Fi data collected along the checkerboard trajectory shown in Figure 5.14(a), and their localization performance is evaluated on multi-band Wi-Fi data collected along the cross trajectory in Figure 5.14(b).

The results, illustrated in Figure 5.15, highlight clear differences in extrapolation behavior beyond the training trajectory. Baseline methods such as NearestInt and RNN-Decay tend to produce estimates clustered around regions observed during training. NearestInt exhibits a strong



(a) Checkerboard trajectory (b) Cross trajectory

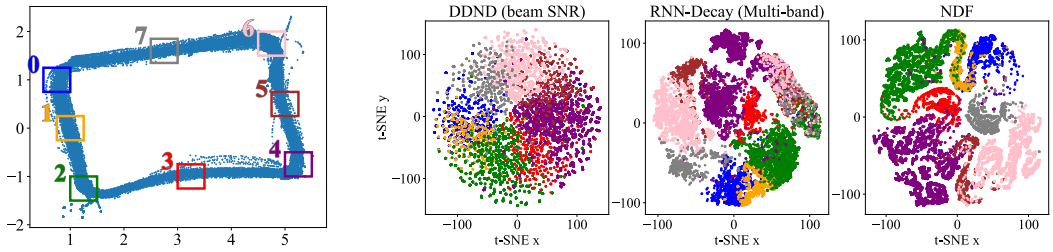
Figure 5.14: Evaluation trajectories with more complex spatial patterns, including (a) a checkerboard trajectory and (b) a cross-shaped trajectory, designed to assess performance beyond the training distribution.



(a) NearestInt (b) RNN-Decay (c) NDF (ours)

Figure 5.15: Cross-trajectory localization results after training on a checkerboard trajectory. The blue dots indicate the ground truth location of the cross-shaped trajectory, while the red dots show the estimated locations produced by two baseline methods (a) and (b) and the proposed NDF (c), using multi-band Wi-Fi.

regression toward densely sampled areas, resulting in over-smoothed estimates near the trajectory center. RNN-Decay partially mitigates this effect by leveraging temporal context, but still struggles to recover the full unseen structure. In contrast, NDF generates spatially coherent predictions that closely follow the unseen cross-shaped trajectory, preserving its global structure, especially at corners and turning points. Trajectory similarity is further quantified using the Fr'echet distance between the estimated and ground truth curves, computed with every 100 points, as reported in Table 5.7. These results confirm that NDF more accurately reproduces the overall shape of unseen trajectories compared with the baseline approaches.



(a) Selected local regions for latent space inspection.

(b) Latent states from selected regions projected on 2D plane by t-SNE.

Figure 5.16: Two-dimensional projections of latent representations corresponding to different spatial regions, with matching colors between the local regions and their corresponding latent states.

Latent Space Visualization

To better understand how spatial information is encoded across different methods, we examine the structure of the learned latent representations by partitioning the trajectory into eight local regions, as illustrated in Figure 5.16(a). All data frames whose ground-truth locations fall within the same region are collected and mapped into latent states using selected baseline methods as well as the proposed NDF. These high-dimensional latent representations are then projected into two dimensions using t-distributed Stochastic Neighbor Embedding (t-SNE) for visualization.

Figure 5.16 shows the t-SNE results for the single-band baseline based on beam SNR (DDND), the multi-band RNN-Decay fusion baseline, and the proposed NDF. Compared to the single-band DDND, the RNN-Decay fusion baseline yields noticeably improved separation among regions, underscoring the benefit of incorporating multi-band Wi-Fi measurements. However, overlap remains between the latent clusters corresponding to regions 2 and 6, particularly in the upper-right area of the embedding.

By contrast, as shown in Figure 5.16(b), NDF produces a more compact and distinctly separated latent representation, with tighter clusters associated with each region. These results indicate that the learned low-dimensional latent space in NDF more effectively preserves the underlying trajectory geometry. Moreover, adjacent clusters are smoothly connected at their boundaries, forming a continuous latent manifold that reflects gradual spatial transitions along the trajectory.

5.1.6 Conclusion and Contribution

This chapter examined the problem of continuous trajectory estimation under heterogeneous and asynchronous wireless observations, and introduced Neural Differential Fusion (NDF), a latent-space modeling approach that explicitly accounts for temporal misalignment across frequency bands. NDF employs a multi-encoder, multi-decoder architecture that aligns latent representations across different frequency bands through a learnable ODE model, thereby realizing principled fusion in the latent space. Experiments with real-world multi-band Wi-Fi data demonstrated

that NDF substantially improves localization accuracy compared to single-band and naive fusion baselines, reducing the average error to approximately 26 cm, which is better than baselines with errors exceeding 50 cm.

From a thesis-wide perspective, this chapter addresses the advanced methodological challenge outlined in Chapter 1 that of developing learning frameworks capable of modeling heterogeneous and asynchronous wireless observations. By integrating neural ODEs with multi-band representation learning, NDF provides a generalizable approach for continuous-time reasoning in wireless perception, bridging the gap between discrete packet measurements and dynamic human motion. This contribution positions NDF as a methodological culmination that advances the expressiveness and temporal coherence of wireless sensing models. Beyond localization, the proposed framework offers a foundation for future research in continuous activity recognition, gait-based identification, and other behavioral sensing tasks that require temporally consistent modeling of wireless signals.

Chapter 6

Discussion

Throughout this dissertation, we have addressed the central objective of advancing wireless sensing toward practical, real-world applicability—by improving data accessibility, developing standardized and scalable acquisition platforms, and designing advanced learning frameworks. Although these efforts collectively demonstrate the feasibility of deploying wireless sensing with commodity devices and realistic environments, they also reveal several methodological and practical limitations that remain unresolved. Building on these findings, this chapter reflects on the insights gained across Wi-Fi and radar modalities, analyzes their respective strengths and weaknesses, and explores future research directions toward more generalizable, interpretable, and integrated wireless perception.

6.1 Remaining Limitations

6.1.1 Methodological Limitations

Although this dissertation has advanced wireless perception across Wi-Fi and mmWave radar modalities, several methodological limitations remain in achieving **robust and generalizable sensing** under real-world variability. Across all proposed systems, the central challenge lies not merely in accuracy but in ensuring consistent performance across unseen domains that inevitably deviate from the conditions represented in the training data.

This limitation became particularly evident through the MMVR benchmark introduced in Chapter 2. Although MMVR provided one of the most diverse radar datasets to date, benchmarks revealed that existing radar-based perception models exhibit poor cross-scene generalization, even when trained on large-scale data (see Table 2.4). These findings underscore that dataset scale alone does not guarantee robustness, and that model design and inductive bias remain critical for handling unseen propagation conditions. Subsequent chapters partially addressed this issue from different perspectives. In Chapter 5, the proposed NDF framework demonstrated improved adaptability to previously unseen trajectories by modeling multi-band Wi-Fi measurements as continuous latent dynamics via neural ODEs. Although this approach outperformed the single-

band and naive fusion baselines (see Figure 5.13, Figure 5.15, and Table 5.6), its performance still degraded under more complex trajectory patterns, suggesting that achieving reliable generalization to arbitrary dynamic scenes remains an open problem. Other works in this dissertation, such as CBR-ACE, CSI2Image, and RAPTR, primarily focused on methodological innovation at the application-level rather than explicit cross-domain robustness. Nevertheless, the findings of these chapters provide the necessary methodological groundwork for future exploration of generalization.

Beyond generalization, another limitation arises from **the trade-off between model sophistication and interpretability**. To achieve high accuracy and sophisticated applications, our methods often rely on increasingly complex neural architectures that, while powerful, obscure the underlying physical relationships between propagation and perception. In this dissertation, several efforts were made to mitigate this gap by embedding physical structure directly into the learning process. For instance, the proposed NDF framework in Chapter 5 models the propagation of Wi-Fi signals as continuous-time dynamics governed by ordinary differential equations, while RAPTR in Chapter 4 introduces pseudo-3D deformable attention that explicitly reasons about geometric relationships in 3D space. These designs provide partial interpretability by revealing what aspects of spatial and temporal variation the model attends to, thereby bridging physical signal behavior and learned representations. However, fully tracing what is learned inside deep neural networks remains difficult, especially as models grow in complexity. This challenge becomes particularly pronounced in implicit generative frameworks such as CSI2Image, where spatial reconstruction is guided by adversarial objectives rather than explicit propagation constraints.

Supervision dependence and computational scalability are also critical concerns. Despite the growing sophistication of neural approaches, most current systems still rely heavily on large-scale, labeled datasets for effective training. This dependence imposes both practical and computational constraints: annotation is labor-intensive, dataset curation requires controlled environments, and scaling to diverse scenes or subjects quickly becomes infeasible. Even when such datasets are available, as with MMVR, fully leveraging them often necessitates increasingly complex and parameter-heavy models, which in turn lead to higher training costs, longer optimization cycles, and demanding hardware requirements at inference time. Although this dissertation explored strategies such as weak label supervision as demonstrated by RAPTR in Chapter 4 and the use of readily available standard-compliant signals by CBR-ACE in Chapter 3 to mitigate these barriers, achieving fully scalable wireless sensing will require a paradigm shift towards self-supervised and lightweight architectures that can generalize from limited or unlabeled data.

Beyond the limitations discussed above, it is worth noting that ongoing standardization efforts in wireless sensing, such as IEEE 802.11bf [211], are expected to improve data accessibility in the near future. By defining unified protocols for precise ToF and/or AoA estimation and standardized sensing feedback within Wi-Fi hardware, these updates will make it possible to obtain propagation-

level measurements natively, without requiring specialized firmware or proprietary modifications. Although such developments may alleviate part of the accessibility barrier identified in this dissertation, the higher-level challenges of generalization, interpretability, and scalable learning will remain critical to realizing practical and human-centric wireless perception.

6.1.2 Community-Level Challenges

Over the past decade, the wireless sensing community has repeatedly highlighted several long-standing goals, such as robust multi-person perception, reliable generalization across diverse environments, and deployability on commodity devices. However, these challenges remain largely unresolved. Despite consistent recognition of their importance, progress toward these objectives has been gradual, while publications continue to emerge that focus on incremental methodological variations or niche sensing tasks. Such works are often able to demonstrate strong performance in narrow or controlled settings and thus still succeed within the current academic evaluation paradigm.

Importantly, this situation does not arise from a lack of technical sophistication or effort. Rather, it reflects a deeper structural constraint: even if researchers seek to develop truly generalizable models, the foundations required to do so are underdeveloped. In fields such as computer vision and natural language processing, breakthrough in robustness and transferability were catalyzed by the availability of shared data formats, large-scale public corpora, and standardized evaluation pipelines. Wireless perception, by contrast, lacks a unified abstraction for radio measurements, and datasets are fragmented along axes of hardware platform, frequency band, signal representation, and preprocessing pipeline. The intrinsic dependence of radio signals on hardware front-ends, antenna geometry, firmware behaviors, and propagation environments makes such standardization technically challenging, but it is precisely this challenge that constrains the field's scientific maturity.

Without a common representational layer, researchers must repeatedly solve bespoke engineering problems before pursuing algorithmic contributions, and accumulated knowledge remains "siloed" in mutually incompatible experimental settings. Under these conditions, claims of progress may reflect success within isolated domains rather than movement toward broad generalization. Thus, the central issue is not merely that generalization has not yet been achieved, but that the community currently lacks the shared data infrastructure and interoperable formats necessary to develop and evaluate methods that could, in principle, generalize.

6.2 Future Prospects for Real-World Deployment

The trajectory of wireless sensing is entering a new stage — one that demands not only higher accuracy, but also adaptability, scalability, and meaningful integration into the fabric of every-

day environments. As wireless technologies continue to converge with advances in deep learning, embedded computing, and ubiquitous networking, the question is no longer whether wireless perception is possible, but how it can evolve into a dependable and interpretable paradigm. This shift requires rethinking the boundaries between communication and perception, between sensing hardware and learning approaches, and between single-modality inference and collective multi-modal intelligence. The following sections discuss these directions, examining how Wi-Fi and radar complement each other, how integration across modalities can yield more robust perception, and how future systems can transition from research prototypes to deployable, everyday technologies.

6.2.1 Methodological Advancements for Generalization

Future progress in wireless sensing will largely depend on overcoming the methodological limitations identified in the previous section, particularly those concerning generalization and scalability. A primary research direction involves improving **domain generalization and adaptation**. As shown by the MMVR benchmark and the NDF analysis, model performance degrades significantly under domain shifts such as unseen trajectories or propagation environments. Existing studies have proposed domain adaptation [54, 241–247] and meta-learning frameworks [248–254] that can transfer representations across different devices, scenes, or users. Another line of advances in few-shot and zero-shot learning [255–262], and self-supervised learning [38, 263–269] also offer promising directions for achieving this adaptability by enabling models to infer novel activities or spatial configurations from limited supervision.

Such adaptation and generalization techniques are not only critical for ensuring robustness under limited real-world data but also play a key role in bridging the gap when leveraging **simulation-generated datasets** [270–273]. A common strategy to mitigate data scarcity is to construct generalized indoor environments in simulation and generate synthetic wireless propagation data across diverse scenarios for model pretraining. However, discrepancies inevitably arise between simulated propagation characteristics (e.g., simplified multipath, ideal materials, or absence of dynamic clutter) and the complex radio conditions encountered in real indoor environments. In this context, adaptation methods become essential tools for aligning synthetic and real-world representations, allowing models trained in simulation to perform reliably under real conditions. Therefore, simulation-based data generation and adaptive learning should be seen as complementary approaches: simulation expands data coverage, while adaptation ensures transferability and robustness in deployment.

While current efforts demonstrate the potential of domain-independent learning in controlled settings, their generalization to more complex or fine-grained perception tasks remains largely unverified. Expanding realistic datasets, as initiated in this dissertation through WiPiCap and MMVR, and further exploring the potential of simulation for modeling indoor wireless propagation

will enable a more systematic evaluation of how these adaptive learning strategies perform under real-world variability.

6.2.2 Toward Coexisting and Task-Adaptive Modalities

Throughout this dissertation, Wi-Fi and mmWave radar have been explored as two representative yet contrasting modalities for indoor wireless perception. Wi-Fi sensing leverages existing communication infrastructure, offering broad accessibility and low deployment cost, whereas radar provides fine spatial resolution and robustness under cluttered or visually occluded conditions. Rather than competing technologies, these modalities should be viewed as **coexisting components** within a unified sensing ecosystem, each suited to different spatial scales, scene conditions, and task requirements.

The key insight emerging from this dissertation is that practical wireless perception will require both **dynamic modality adaptation** and **multi-modal integration**. Instead of committing to a single sensing pipeline, future systems should dynamically allocate and coordinate sensing resources according to situational demands. For instance, Wi-Fi sensing can be activated for coarse environmental awareness or presence detection, while radar can be employed for detailed motion or pose analysis, or both can be jointly leveraged when higher reliability or robustness is required. Such adaptive cooperation allows systems to optimize trade-offs among spatial resolution, energy consumption, latency, and privacy.

Beyond dynamic switching, the next frontier lies in **multi-modal fusion**, where heterogeneous sensing modalities collaborate to form a unified representation of the environment. Integrating complementary signals from Wi-Fi, radar, and potentially other modalities such as visual or acoustic sensors can improve both the robustness and semantic richness of perception. However, effective multi-modal fusion remains an open challenge due to differences in signal domains, spatial resolution, and temporal synchronization. Recent studies on multi-modal integration [274–282] and ISAC [283–287] have made steady progress. These efforts remain largely confined to controlled settings and signal-level integration. In this sense, future work should move toward a vision of context-aware, semantic-level integration that aligns features across different signal domains and provides intelligent understanding of complex indoor scenes.

6.2.3 Establishing Community-Unified RF Data Format

A recurring theme throughout this dissertation is that progress in wireless sensing is fundamentally constrained by the absence of **a common representational layer for RF measurements for sensing.** Unlike modalities such as images or natural language, in which the relationship between observations and their semantic structure is relatively stable across varying environments, wireless signals exhibit a fundamentally different property. Radio measurements are not direct

observations of objects or scenes, but superpositions of multiple propagation paths arising from line-of-sight transmission, specular and diffuse reflections, and scattering effects. Each observed signal component is shaped by the surrounding environment, as well as by hardware configurations, antenna geometry, signal bandwidth, and timing. As a result, the correspondence between radio observations and the underlying physical state is highly environment-dependent, and radio signals cannot be meaningfully interpreted without explicit contextual information about the sensing setup and propagation conditions.

This inherent ambiguity makes shared interpretative frameworks and data abstractions a prerequisite for cumulative and comparable research in wireless sensing. Importantly, the lack of shared data abstractions does not merely slow technical progress; it also shapes the epistemic dynamics of the field. In the absence of standardized formats, large-scale open corpora, and consistent evaluation pipelines, research prototypes inevitably rely on isolated hardware platforms, idiosyncratic preprocessing, and bespoke calibration procedures. Under such fragmented conditions, methods are often validated only within narrow, device-specific settings. As a result, approaches that do not generalize beyond controlled environments may still demonstrate competitive performance within isolated benchmarks and thus succeed under current publication practices.

It is important to emphasize that establishing a unified RF data foundation does not guarantee the resolution of longstanding challenges in wireless sensing, such as strong generalization or robustness to environmental variation. Simply increasing the volume of available data does not, by itself, ensure that these problems will be overcome. Nevertheless, the current situation presents a more fundamental limitation: without shared data representations and accessible datasets, results collected by different groups cannot be placed on a common footing for comparison, accumulation, or critical evaluation. In this sense, many claims regarding generalization or robustness remain difficult to assess beyond the specific environments in which the data were obtained.

Crucially, the goal of RF data standardization is not to reproduce identical physical environments or to exhaustively encode all aspects of the surrounding space. Attempting to enumerate every environmental factor influencing radio propagation would not be feasible. Rather than seeking completeness, effective RF data representations must rely on abstraction, selecting and formalizing only those environmental and system attributes that are most relevant for interpretation, comparison, and modeling. Shared formats and interoperable metadata—including antenna geometry, frequency band, calibration frames, timing, and coarse descriptions of the physical context—serve as proxies for dominant sensing conditions. By making such assumptions and contextual factors explicit, differences in performance can be more reliably attributed to modeling choices rather than uncontrolled experimental variation. In this sense, environment descriptions are not exhaustive records of the physical world, but structured variables that shape radio observations and should be treated as part of the sensing data itself.

More broadly, a healthy RF data ecosystem functions not only as technical infrastructure, but also as a self-correcting mechanism for scientific quality. When common evaluation resources exist, methods that overfit to narrow conditions are more readily exposed, while approaches that demonstrate stability across environments and systems gain visibility and legitimacy. Open and accessible RF data infrastructures therefore act as enabling conditions for cumulative knowledge building, even though they do not guarantee specific algorithmic breakthroughs. The contributions presented in this dissertation suggest that movement toward such an ecosystem is both feasible and worthwhile. However, no individual dataset, toolkit, or research group can establish this foundation alone. Realizing a unified RF data abstraction—and with it, the ability to rigorously examine generalizability in wireless sensing—will require sustained, community-wide coordination across academia, industry, and standards bodies. The trajectory outlined here is not a prescriptive endpoint, but a call for collective effort to transition wireless sensing from device-specific experimentation toward a reproducible, scalable, and societally robust technology stack.

6.2.4 Grounded Foundation Models for Wireless Sensing

Beyond modality coexistence, the next frontier lies in the **semantic grounding of wireless perception**. Recent advances in vision–language foundation models, including CLIP [288], GLIP [289], and Grounding DINO [290], have demonstrated the remarkable power of large-scale pretraining across heterogeneous modalities. Their success stems not merely from model capacity, but from the ability to learn correspondences between fundamentally different representations, such as visual and linguistic domains.

Wireless signals, by contrast, encode information that is rich yet highly abstract. When considered in isolation, they rarely provide direct semantic cues about human actions or environmental context. This observation suggests that further progress in wireless perception may arise from cross-modal grounding—learning to associate radio-frequency observations with more interpretable modalities, including images, depth maps, or textual descriptions of human activities and spatial configurations. In current research, cross-modal grounding efforts in wireless sensing have predominantly focused on associating RF observations with language-based representations, often through large language models (LLMs) or vision–language models (VLMs). For example, mmCLIP [258] aligns mmWave and text embeddings via contrastive learning on synthetic signal-text pairs and employs lightweight fine-tuning to achieve zero-shot human activity recognition. More recent Talk2Radar [291] introduces the first 4D mmWave radar–text dataset for 3D referring expression comprehension in autonomous driving, pairing radar point clouds with natural-language prompts. In a different direction, WiTalk [292] incorporates hierarchical text prompts, including label names and detailed action descriptions, into existing wireless sensing models, using textual information to guide feature learning without explicitly enforcing a shared embedding space between RF signals

and text. Wi-Chat [293] demonstrates that large language models can directly interpret raw Wi-Fi CSI by incorporating physical models of wireless propagation into natural-language prompts. Likewise, GLSDA [294] proposes a large-model-aware semantic distillation framework for Wi-Fi gesture recognition, where semantic representations from a pretrained VLM are used to supervise the training of a Wi-Fi CSI encoder. Similarly, MMSense [295] explores another VLM-based foundation model for wireless sensing, in which visual representations form the core backbone and wireless measurements such as mmWave radar are integrated as additional modalities within a unified representation.

While language-based grounding provides a technically viable and increasingly effective pathway for bridging RF signals and semantic understanding, it represents only one point in a broader design space. A natural extension lies in integrating wireless signals with a wider range of modalities, such as images, videos, and audio, which can provide complementary spatial, temporal, and contextual cues. By fusing RF observations with these more perceptually grounded modalities, future systems may assign more robust and interpretable semantic meaning to the rich but abstract spatial information inherently encoded in wireless signals.

6.2.5 Ethical and Privacy Considerations

Although wireless sensing has long been regarded as a privacy-preserving alternative to cameras or wearables, its growing perceptual capabilities raise new ethical and societal challenges. As learning-based methods become increasingly sophisticated, the boundary between anonymous environmental perception and identifiable personal inference is becoming less distinct. In addition, the inherent transparency and penetrability of radio waves make wireless signals susceptible to intercepting from outside the target space. This physical property implies that research on identification and re-identification, though technically valuable, could be misused if applied without ethical safeguards, as wireless signals can be captured and analyzed remotely. In addition, the use of communication signals in Wi-Fi sensing introduces unique security vulnerabilities. Li et al. [296] demonstrated that intentionally modifying the pilot symbols of IEEE 802.11 packets can alter the measured CSI, thereby misleading activity recognition systems. Several works have demonstrated the perturbation attack, whether injected as adversarial RF interference, spoofed chirps, or targeted jamming. Such perturbations can compromise learned models, causing misclassification or degraded sensing performance [297–300]. Such findings illustrate that both passive misuse and active manipulation of wireless signals must be considered in the broader ethical design of wireless perception technologies.

Currently, wireless perception research has focused primarily on improving accuracy, resolution, or robustness, while discussions about responsible data use and system transparency remain limited. To ensure that wireless sensing technologies evolve in socially acceptable directions, fu-

ture work must incorporate privacy-aware design principles at every layer of the sensing pipeline: from anonymized data collection and federated learning to adversarially robust model design and on-device inference. In addition, interdisciplinary collaboration with fields such as law, ethics, and human–computer interaction will be essential to establish regulatory and social frameworks for responsible deployment.

Ultimately, as wireless perception moves from controlled laboratories to everyday environments, its success will depend not only on technical feasibility, but also on public trust. Ensuring that these systems remain explainable, controllable, and aligned with human values will be critical to realizing the promise of human-centric and privacy-safe wireless intelligence.

Chapter 7

Conclusion

This dissertation examined how wireless sensing can advance toward practical deployment in real-world indoor environments. Although Wi-Fi and mmWave radar sensing offer unique advantages as privacy-preserving and environment-robust alternatives to cameras and wearable devices, their real-world applicability has remained limited. Two key challenges persist: the restricted accessibility and diversity of sensing data and the limited validation of learning methodologies under realistic conditions. To address these challenges, this work pursued both directions—developing open, reproducible data acquisition pipelines and designing propagation-aware frameworks that can operate reliably with realistic signal and environmental variability.

Chapter 1 outlined the motivation and scope of this dissertation, positioning wireless sensing as a privacy-preserving and environment-robust alternative to conventional perception modalities. It identified two central challenges that hinder practical deployment: the limited accessibility of wireless sensing data and the lack of methodologies capable of generalizing across real-world environments. By framing these issues as interdependent pillars—data accessibility and methodological advancement—the chapter established the foundation for the subsequent studies, which collectively aim to bridge the gap between experimental research and deployable wireless perception systems.

Chapter 2 addressed the accessibility and representativeness of sensing data, introducing two complementary frameworks that enable open and reproducible data use for Wi-Fi and radar sensing. WiPiCap demonstrated that the standardized CBRs defined in IEEE 802.11ac/ax can be captured on commodity devices, eliminating device dependencies and lowering the barrier to realistic acquisition of Wi-Fi signals. In parallel, MMVR provided a large-scale multi-view radar dataset with synchronized annotations for detection, pose estimation, and segmentation, substantially extending the diversity and scale of available radar benchmarks. Both frameworks have been released as open-source resources, contributing to the community by facilitating reproducible research and accelerating progress in wireless sensing.

Chapter 3 explored two complementary directions for advancing Wi-Fi sensing using standardized signal formats. CBR-ACE implemented a lightweight signal-processing framework that

reconstructs and analyzes CBR time-series to estimate repetitive human exercises on commodity IEEE 802.11ac/ax devices, demonstrating a fully standard-compliant sensing pipeline. In contrast, CSI2Image introduced a novel application concept for Wi-Fi sensing — translating wireless measurements into spatially meaningful images through generative adversarial learning — thereby extending the scope of wireless perception beyond conventional classification or localization tasks. Together, these studies established both a practical and an exploratory foundation for signal-level and representation-level innovation in standardized Wi-Fi sensing.

Chapter 4 presented RAPTR, a transformer-based framework for 3D human pose estimation using mmWave radar. RAPTR introduced a two-stage decoding process and structured losses that progressively reconstruct 3D pose geometry from weak 2D and 3D bounding-box labels, while its pseudo-3D deformable attention explicitly models spatial relationships within radar reflections. Through this design, RAPTR achieved high-fidelity 3D pose inference without requiring costly 3D pose annotations, providing a scalable and label-efficient approach to radar-based human perception. This contribution extends the focus of the dissertation from data accessibility to practical 3D understanding, reinforcing radar sensing as a viable modality with scalability.

Chapter 5 advanced the methodological scope to multi-band sensing by proposing Neural Dynamic Fusion (NDF), which integrates sub-7 GHz CSI and 60 GHz SNR measurements defined in recent IEEE 802.11 standards. NDF formulates signal propagation as a continuous-time dynamic process and aligns asynchronous multi-band observations via neural ordinary differential equations, enabling unified representation learning across heterogeneous frequencies. While demonstrated through continuous trajectory estimation, the framework offers a general formulation applicable to a wide range of dynamic perception tasks.

Finally, Chapter 6 synthesized the key insights of this dissertation and outlined the remaining challenges and future research directions. It highlighted several paths to advance wireless perception beyond current methodological and infrastructural limits, including: (i) methodological progress and development of grounded foundation models for scalable generalization ; (ii) task-adaptive coexistence and context-aware perception with multimodal integration; (iii) establishment of standardized RF data formats as the community effort; and (iv) development of ethically responsible system design to ensure public trust.

In summary, this dissertation has taken steps toward advancing wireless sensing from controlled, device-dependent experiments to a more standardized, learning-driven paradigm. By improving data accessibility through open acquisition frameworks and incorporating physics-informed principles into neural architectures, it has contributed to building a more reproducible and scalable foundation for wireless perception in realistic indoor settings. While many challenges remain in achieving fully adaptive and universally deployable sensing, the approaches established here are expected to inform the next generation of adaptive, trustworthy, and integrated wireless sensing.

Acknowledgement

This dissertation would not have been possible without the invaluable assistance, encouragement, and support of numerous individuals who have guided, inspired, and accompanied me throughout my doctoral journey. I am deeply grateful to each and every one of them for their contributions, both direct and indirect, to this work and to my personal and academic growth.

First, I would like to extend my deepest appreciation to my supervisor, Associate Professor Shunsuke Saruwatari, for his continued guidance and insightful advice throughout the course of my doctoral studies. I am also profoundly grateful to Professor Takashi Watanabe, who supervised me during the early years of my doctoral program and provided invaluable mentorship and encouragement until his retirement. Both professors have always encouraged me to think independently, pursue my curiosity, and remain persistent in the face of challenges. Their mentorship has profoundly influenced how I approach research and professional growth, and I am sincerely thankful for the trust and opportunities they have provided throughout my studies. Without their devoted supervision and inspiration, this dissertation would never have been possible.

I am profoundly grateful to the members of my thesis committee — Professor Hirozumi Yamaguchi of the Graduate School of Information Science and Technology, Professor Atsuo Inomata of the D3 Center, and Professor Hideyuki Shimonishi of the Cyber Media Center, the University of Osaka — for their thorough reviews and perceptive feedback. Their engagement and expertise during the evaluation process were truly invaluable.

I would also like to express my sincere gratitude to Assistant Professor Takuya Fujihashi, whose guidance and collaboration have significantly broadened my research scope. Working with him across diverse technical areas, including interdisciplinary topics that extended beyond my initial expertise, was an intellectually enriching experience. His analytical rigor, practical insight, and enthusiasm for collaborative exploration have deeply influenced my approach to both research and teamwork.

Deep appreciation is also due to Dr. Pu (Perry) Wang, Senior Principal Research Scientist at Mitsubishi Electric Research Laboratories, who kindly hosted my research stay in the United States. Immersing myself in the intellectually rigorous and innovative environment of his group was a transformative experience that deeply influenced my academic trajectory. The stimulating

discussions with him and his colleagues not only strengthened my research abilities and critical thinking but also exposed me to international research practices and perspectives. I am sincerely thankful for his warm hospitality, invaluable feedback, and continuous encouragement even after my return to Japan.

I am profoundly grateful to Ms. Ami Takahashi, Ms. Sayaka Nishimura, and Ms. Megumi Fukui for their exceptional administrative support and assistance in managing laboratory affairs throughout my doctoral journey. Their meticulous work and constant kindness ensured that the research environment remained smooth and well-organized, allowing me to focus entirely on my studies and experiments.

I would also like to extend my heartfelt thanks to all students and associates of our laboratory. Having the opportunity to interact with students working on diverse topics beyond wireless communications has greatly broadened my research horizons. The discussions, informal seminars, and occasional collaborations we shared offered new ideas and perspectives that continuously refreshed my motivation. I truly believe that being in such an inspiring environment — one that encouraged open dialogue, mutual learning, and creative exploration — made the seven years I spent here an invaluable and fulfilling experience that I will cherish for life.

Finally, I am deeply grateful to my family for their unwavering support, patience, and encouragement throughout my doctoral journey. Their constant understanding, sacrifices, and belief in me have been an immeasurable source of strength that sustained me through every challenge and moment of doubt. This dissertation stands as much a testament to their love and trust as to my academic effort. I owe them more than words can express, and I dedicate this achievement to them with heartfelt gratitude.

List of All Publications

Journal Articles

1. Yusuke Ikemura, **Sorachi Kato**, Takeru Fukushima, Takuya Fujihashi, Tomoki Murakami, Takashi Watanabe, Shunsuke Saruwatari, "Recursive Multi-Branch Neural Network for Object Property Estimation Using Wi-Fi CSI," *Journal of Information Processing*, 2025 (accepted).

2. Takeru Fukushima, Takuma Matsukawa, Yusuke Ikemura, Hayato Terasoma, **Sorachi Kato**, Takuya Fujihashi, Tomoki Murakami, Yusuke Asai, Shunsuke Saruwatari, "Wi-Fi Meets Sound: A Deep Fusion Leap Beyond Single-Modality for Device Localization," *IEEE Access*, vol. 13, pp. 170291–170306, 2025.

3. Akihiro Kuwabara, **Sorachi Kato**, Toshiaki Koike-Akino, and Takuya Fujihashi, "JSCC-Aided INR for High-Frequency Detail Preservation in LiDAR," *IEEE Open Journal of the Communications Society*, Vol. 6, pp. 6352–6367, 2025.

4. **Sorachi Kato**, Pu (Perry) Wang, Toshiaki Koike-Akino, Takuya Fujihashi, Hassan Mansour, Petros Boufounos, "Multi-Band Wi-Fi Neural Dynamic Fusion," *IEEE Transactions of Wireless Communications*, vol. 24, no. 5, pp. 3698–3714, 2025.

5. **Sorachi Kato**, Tomoki Murakami, Takuya Fujihashi, Takashi Watanabe, Shunsuke Saruwatari, "CBR-ACE: Counting Human Exercise using Wi-Fi Beamforming Reports," *Journal of Information Processing*, vol. 30, pp. 66–74, 2022.

6. **Sorachi Kato**, Takeru Fukushima, Tomoki Murakami, Hirantha Abeysekera, Yusuke Iwasaki, Takuya Fujihashi, Takashi Watanabe, Shunsuke Saruwatari, "CSI2Image: Image Reconstruction from Channel State Information Using Generative Adversarial Networks," *IEEE Access*, vol. 9, pp. 47154–47168, 2021.

International Conference Papers

1. Yusuke Ikemura, **Sorachi Kato**, Tomoki Murakami, Takeru Fukushima, Takuya Fujihashi, and Shunsuke Saruwatari, "Demo: WiPiCap: Real-time IEEE 802.11ac/ax Compressed

Beamforming Report Decoding System," in *IEEE Consumer Communications and Networking Conference (CCNC)*, pp. 1–2, 2026 (accepted).

2. Akihiro Kuwabara, Yutaro Osako, **Sorachi Kato**, Toshiaki Koike-Akino, and Takuya Fujihashi, "Global-Local Knowledge-Assisted Joint Source-Channel Coding," in *IEEE Consumer Communications and Networking Conference (CCNC)*, pp. 1–6, 2026 (accepted).

3. Tomoki Kamise, **Sorachi Kato**, Takumasa Ishioka, Tatsuya Fukui, Tsubasa Okamoto, Toshihito Fujiwara, Satoshi Narikawa, Shunsuke Saruwatari, and Takuya Fujihashi, "INR-Assisted Speculative Spatial Audio Transmission for Zero-Delay Cloud Gaming," in *IEEE Consumer Communications and Networking Conference (CCNC)*, pp. 1–6, 2026 (accepted).

4. Sorachi Kato, Ryoma Yataka, Pu (Perry) Wang, Pedro Miraldo, Takuya Fujihashi, and Petros Boufounos, "RAPTR: Radar-based 3D Pose Estimation using Transformer," in *the 39th Annual Conference on Neural Information Processing Systems (NeurIPS)*, pp. 1–26, 2025.

5. Akihiro Kuwabara, Hinata Kirihara, **Sorachi Kato**, Toshiaki Koike-Akino, and Takuya Fujihashi, "L-GGSC: Learnable Graph-based Gaussian Splatting Compression," in *IEEE/CVF International Conference on Computer Vision Workshop on ECLR*, pp. 3057–3063, 2025.

6. Takuya Fujihashi, **Sorachi Kato**, and Toshiaki Koike-Akino, "FV-NeRV: Neural Compression for Free Viewpoint Videos," in *Neural Information Processing Systems Workshop on Machine Learning and Compression*, pp. 1–10, 2024.

7. M. Mahbubur Rahman, Ryoma Yataka, **Sorachi Kato**, Pu (Perry) Wang*, Peizhao Li, Adriano Cardace, and Petros Boufounos, "MMVR: Millimeter-wave Multi-View Radar Dataset and Benchmark for Indoor Perception," in *European Conference on Computer Vision (ECCV)*, pp. 306–322, 2024.

8. Takuya Fujihashi, Sorachi Kato, and Toshiaki Koike-Akino, "Implicit Neural Representation For Low-Overhead Graph-Based Holographic-Type Communications," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 2825–2829, 2024.

9. **Sorachi Kato**, Pu (Perry) Wang, Toshiaki Koike-Akino, Takuya Fujihashi, Hassan Mansour, and Petros Boufounos, "Object Trajectory Estimation with Multi-Band Wi-Fi Neural Dynamic Fusion," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 13261–13265, 2024.

10. Akihiro Kuwabara, Yutaro Osako, **Sorachi Kato**, Takuya Fujihashi, Toshiaki Koike-Akino, and Takashi Watanabe, "Implicit Neural Representation-Based Hybrid Digital-Analog Im-

age Delivery," in *International Conference on Computing, Networking and Communications (CCNC)*, pp. 605–610, 2024.

11. Natsuki Ikuo, **Sorachi Kato**, Takuma Matsukawa, Tomoki Murakami, Takuya Fujihashi, Takashi Watanabe, and Shunsuke Saruwatari, "CSI2PC: 3D Point Cloud Reconstruction Using CSI," in *IEEE Consumer Communications and Networking Conference (CCNC)*, pp. 254–259, 2024.

12. Sorachi Kato, Takuma Matsukawa, Tomoki Murakami, Takuya Fujihashi, Takashi Watanabe, and Shunsuke Saruwatari, "Fundamental Investigation of Wi-Fi Beamforming Report Properties on Wireless Sensing," in *IEEE 12th Sensor Array and Multichannel Signal Processing Workshop (SAM)*, pp. 341–344, 2022.

Domestic Conference Papers

The following works were presented in Japanese. The titles are translated from the original Japanese.

1. Masaki Nishimura, Yusuke Ikemura, **Sorachi Kato**, Takeru Fukushima, Tomoki Murakami, and Shunsuke Saruwatari, "Fundamental Study on the Degradation of Wi-Fi Sensing Accuracy under Environmental Changes", in *IEICE Society Conference*, no. B-15-05, pp. 1–1, 2025.

2. Takuma Matsukawa, Yusuke Ikemura, Hayato Terasoma, **Sorachi Kato**, Takuya Fujihashi, Tomoki Murakami, Takashi Watanabe, and Shunsuke Saruwatari, "Fundamental Study on Multimodal Localization Using Millimeter-Wave and Acoustic Signals", in *IEICE Technical Committee on Sensor Networks and Mobile Intelligence (SeMI)*, Vol. 124, No. 353, pp. 118–123, 2025.

3. Yusuke Ikemura, Hayato Terasoma, Takuma Matsukawa, **Sorachi Kato**, Takuya Fujihashi, Tomoki Murakami, Takashi Watanabe, and Shunsuke Saruwatari, "Fundamental Study on Non-Invasive Inspection Using Wi-Fi". in *IEICE Technical Committee on Sensor Networks and Mobile Intelligence (SeMI)*, Vol. 124, No. 353, pp. 7–12, 2025.

4. Hayato Terasoma, Yusuke Ikemura, Takuma Matsukawa, **Sorachi Kato**, Takuya Fujihashi, Tomoki Murakami, Takashi Watanabe, and Shunsuke Saruwatari, "Fundamental Study on Sugar Content Sensing Using Wi-Fi", in *IEICE Technical Committee on Sensor Networks and Mobile Intelligence (SeMI)*, Vol. 124, No. 353 pp. 1–6, 2025.

5. Yusuke Ikemura, Takuma Matsukawa, **Sorachi Kato**, Tomoki Murakami, Takuya Fujihashi, Takashi Watanabe, and Shunsuke Saruwatari, "A Study on Wi-Fi Sensing Using a Recursive

Multi-Stage Network", in *IPSJ Multimedia, Distributed, Cooperative, and Mobile (DICOMO) Symposium*, No. 4C-1, pp. 702–708, 2024.

6. Akihiro Kuwabara, Yutaro Osako, **Sorachi Kato**, Takuya Fujihashi, Toshiaki Koike-Akino, and Takashi Watanabe, "A Study on LiDAR Point Clouds Using Neural Implicit Representation", in *IPSJ Multimedia, Distributed, Cooperative, and Mobile (DICOMO) Symposium*, No. 6E-1, pp. 1222–1230, 2024.

7. Yusuke Ikemura, Natsuki Ikuo, Takuma Matsukawa, **Sorachi Kato**, Tomoki Murakami, Takuya Fujihashi, Shunsuke Saruwatari, and Takashi Watanabe, "Fundamental Study on Multitask Wi-Fi Sensing for 3D Structures", in *National Convention of IPSJ*, No. 2Y-02, pp. 1–2, 2024.

8. Takuma Matsukawa, **Sorachi Kato**, Takuya Fujihashi, Tomoki Murakami, Takashi Watanabe, and Shunsuke Saruwatari, "Experimental Study on Angle-of-Arrival Estimation Assuming IEEE 802.11bf", in *IPSJ Multimedia, Distributed, Cooperative, and Mobile (DICOMO) Symposium*, No. 5D-2, pp. 787–794, 2023.

9. Yusuke Ikemura, Takuma Matsukawa, Natsuki Ikuo, **Sorachi Kato**, Takuya Fujihashi, Tomoki Murakami, Shunsuke Saruwatari, and Takashi Watanabe, "Fundamental Study on Density Estimation of 3D Structures Using Wi-Fi", in *IEICE Society Conference*, No. B-15-30, pp. 1–1, 2023.

10. **Sorachi Kato**, Tomoki Murakami, Takuya Fujihashi, Takashi Watanabe, and Shunsuke Saruwatari, "A Study on Improving Environmental Robustness of Wi-Fi Sensing", in *IEICE Technical Committee on Sensor Networks and Mobile Intelligence (SeMI)*, No. 13, pp. 1–2, 2023.

11. Takuma Matsukawa, **Sorachi Kato**, Takuya Fujihashi, Tomoki Murakami, Takashi Watanabe, and Shunsuke Saruwatari, "A Study on Angle-of-Arrival Estimation Using Wi-Fi Access Points", in *National Convention of IPSJ*, No. 6ZA-05, pp. 1–2, 2023.

12. Natsuki Ikuo, **Sorachi Kato**, Hiroaki Shinmiya, Takuya Fujihashi, Tomoki Murakami, Takashi Watanabe, and Shunsuke Saruwatari, "A Study on 3D Structure Reconstruction Using CSI", in *IEICE Technical Committee on Sensor Networks and Mobile Intelligence (SeMI)*, No. 21, pp. 1–6, 2022.

13. Natsuki Ikuo, **Sorachi Kato**, Hiroaki Shinmiya, Takuya Fujihashi, Tomoki Murakami, Takashi Watanabe, and Shunsuke Saruwatari, "Fundamental Study on 3D Structure Reconstruction Using CSI", in *IEICE Society Conference*, pp. 1–1, 2022.

14. Takuma Matsukawa, **Sorachi Kato**, Takuya Fujihashi, Takashi Watanabe, and Shunsuke Saruwatari, "Fundamental Study on Channel Feedback Errors Caused by Compression in IEEE 802.11ac", in *IEICE General Conference*, No. B-15-39, pp. 477–477, 2022.

15. **Sorachi Kato**, Tomoki Murakami, Takuya Fujihashi, Takashi Watanabe, and Shunsuke Saruwatari, "A Study on the Development of a Motion Repetition Estimation System Using Compressed Beamforming Reports", in *IPSJ Multimedia, Distributed, Cooperative, and Mobile (DICOMO) Symposium*, no. 2B-1, pp. 243–255, 2021.

16. Shunpei Ohira, Kentaro Hayashi, **Sorachi Kato**, Kazuya Ohara, Tomoki Murakami, Takuya Fujihashi, Takashi Watanabe, and Shunsuke Saruwatari, "Fundamental Study on Enhancing Angle-of-Arrival Estimation Accuracy Utilizing the Wideband IEEE 802.11ac/ax", in *National Convention of IPSJ*, No. 7V-01, pp. 1–2, 2021.

17. **Sorachi Kato**, Tomoki Murakami, Takuya Fujihashi, Shunsuke Saruwatari, and Takashi Watanabe, "Fundamental Study on the Development of a DIY-Style Wireless Sensing System", in *National Convention of IPSJ*, No. 1T-01, pp.1–2, 2021

18. **Sorachi Kato**, Yusuke Iwasaki, Takeru Fukushima, Takuya Fujihashi, Tomoki Murakami, Abeysekera Hirantha, Shunsuke Saruwatari, and Takashi Watnanabe, "Fundamental Study on the Temporal Robustness of CSI2Image", in *National Convention of IPSJ*, No. 5P-02, pp. 1–2, 2020.

Awards

1. Best Presentation Award, "A Study on Improving the Robustness of Wi-Fi Sensing against Environmental Change,", *IEICE Technical Committee on Sensor Networks and Mobile Intelligence (SeMI)*, 2023. (in Japanese)

2. Young Researcher Award, "A Study on Improving the Robustness of Wi-Fi Sensing against Environmental Change,", *IEICE Technical Committee on Sensor Networks and Mobile Intelligence (SeMI)*, 2023. (in Japanese)

3. IPSJ Outstanding Paper Award, "CBR-ACE: Counting Human Exercise using Wi-Fi Beamforming Reports," *Journal of Information Processing*, 2022.

4. Award of the Graduate School of Information Science and Technology of The University of Osaka, The University of Osaka, 2023.

5. Specially Selected Paper, "CBR-ACE: Counting Human Exercise using Wi-Fi Beamforming Reports," *Journal of Information Processing*, 2022.

6. Student Encouragement Award of IPSJ National Convention, "A Fundamental Study on the Construction of a DIY Wireless Sensing System," in *the 83rd National Convention of IPSJ*, 2021. (in Japanese)

7. Student Encouragement Award of IPSJ National Convention, "A Fundamental Study on the Temporal Robustness of CSI2Image," in *the 82nd National Convention of IPSJ*, 2020. (in Japanese)

Scholarship and Financial Supports

1. The Telecommunications Advancement Foundation Overseas Travel Expenses Support, "MMVR: Millimeter-wave Multi-View Radar Dataset and Benchmark for Indoor Perception," in *European Conference on Computer Vision*, 2024.

2. Research Fellow of the Japan Society for the Promotion of Science (JSPS): 2023-2025

Others

1. **Sorachi Kato**, Takuya Fujihashi, Takashi Watanabe, Shunsuke Saruwatari, and Tomoki Murakami, "CSI2Image: CSI-to-Image Conversion using a Generative Model," *Beyond 5G White Paper Supplementary Volume "Sensing Technologies"*, pp. 19–23, 2024.

2. WiPiCap: Wi-Fi Compressed Beamforming Report Decoding System, https://github.com/watalabo/WiPiCap.

3. Millimeter-wave Multi-View Radar (MMVR) Dataset, https://zenodo.org/records/12611978.

4. RAPTR: Radar-based 3D Pose Estimation using Transformer, https://github.com/merlresearch/radar-pose-transformer.

Bibliography

[1] S. Ruan, R. Wang, X. Shen, H. Liu, B. Xiao, J. Shi, K. Zhang, Z. Huang, Y. Liu, E. Chen, and Y. He, "A survey of multi-sensor fusion perception for embodied AI: Background, methods, challenges and prospects," *arXiv preprint arXiv:2506.19769*, 2025.

[2] J. Zhu, B. Sun, L. Jia, and H. Hu, "From sensing technology towards digital twin in applications," *Inventions*, vol. 9, no. 2, p. 43, Apr. 2024.

[3] A. Varol, N. H. Motlagh, M. Leino, S. Tarkoma, and J. Virkki, "Creation of AI-driven smart spaces for enhanced indoor environments – A survey," *arXiv preprint arXiv:2412.14708*, 2024.

[4] D. Bouchabou, S. M. Nguyen, C. Lohr, B. LeDuc, and I. Kanellos, "A survey of human activity recognition in smart homes based on IoT sensors algorithms: Taxonomies, challenges, and opportunities with deep learning," *Sensors (Basel)*, vol. 21, no. 18, p. 6037, Sep. 2021.

[5] F. Kong, Y. Zou, Z. Li, and Y. Deng, "Advances in portable and wearable acoustic sensing devices for human health monitoring," *Sensors*, vol. 24, no. 16, p. 5354, Aug. 2024.

[6] T. Li and H. Yu, "Visual-inertial fusion-based human pose estimation: A review," *IEEE Transactions on Instrumentation and Measurement*, vol. 72, pp. 1–16, 2023.

[7] Q. Lin, S. Peng, Y. Wu, J. Liu, W. Hu, M. Hassan, A. Seneviratne, and C. H. Wang, "E-jacket: Posture detection with loose-fitting garment using a novel strain sensor," in *ACM/IEEE International Conference on Information Processing in Sensor Networks (IPSN)*, Apr. 2020.

[8] Q. Lei, J.-X. Du, H.-B. Zhang, S. Ye, and D.-S. Chen, "A survey of vision-based human action evaluation methods," *Sensors*, vol. 19, no. 19, p. 4129, Sep. 2019.

[9] J. Lian, J. Lou, L. Chen, and X. Yuan, "EchoSpot: Spotting your locations via acoustic sensing," *ACM Interactive, Mobile, Wearable and Ubiquitous Technologies (IMWUT)*, vol. 5, no. 3, pp. 1–21, Sep. 2021.

[10] S. C. Mukhopadhyay, "Wearable sensors for human activity monitoring: A review," *IEEE Sensors Journal*, vol. 15, no. 3, pp. 1321–1330, Mar. 2015.

[11] X. Zhu, J. Liu, L. Lu, T. Zhang, T. Qiu, C. Wang, and Y. Liu, "Enabling intelligent connectivity: A survey of secure ISAC in 6G networks," *IEEE Communications Surveys and Tutorials*, vol. 27, no. 2, pp. 748–781, Apr. 2025.

[12] L. Fan, J. Wang, Y. Chang, Y. Li, Y. Wang, and D. Cao, "4D mmWave radar for autonomous driving perception: A comprehensive survey," *IEEE Transactions on Intelligent Vehicles*, vol. 9, no. 4, pp. 4606–4620, Apr. 2024.

[13] H. Kong, C. Huang, J. Yu, and X. Shen, "A survey of mmWave radar-based sensing in autonomous vehicles, smart homes and industry," *IEEE Communications Surveys and Tutorials*, vol. 27, no. 1, pp. 1–1, 2024.

[14] J. Zhang, R. Xi, Y. He, Y. Sun, X. Guo, W. Wang, X. Na, Y. Liu, Z. Shi, and T. Gu, "A survey of mmWave-based human sensing: Technology, platforms and applications," *IEEE Communications Surveys and Tutorials*, vol. 25, no. 4, pp. 2052–2087, 2023.

[15] X. Chen, C. Li, C. Jiang, W. Meng, and W. Xiao, "WiPhase: A human activity recognition approach by fusing of reconstructed WiFi CSI phase features," *IEEE Transactions on Mobile Computing*, vol. Pp, no. 99, pp. 1–13, 2024.

[16] F. Luo, S. Khan, B. Jiang, and K. Wu, "Vision transformers for human activity recognition using WiFi channel state information," *IEEE Internet of Things Journal*, vol. 11, pp. 28 111–28 122, Sep. 2024.

[17] F. Meneghello, D. Garlisi, N. D. Fabbro, I. Tinnirello, and M. Rossi, "SHARP: Environment and person independent activity recognition with commodity IEEE 802.11 access points," *IEEE Transactions on Mobile Computing*, vol. 22, no. 10, pp. 6160–6175, Oct. 2023.

[18] W. Meng, Z. Liu, B. Li, W. Cui, J. T. Zhou, and L. Zhang, "GrapHAR: A Lightweight Human Activity Recognition Model by Exploring the Sub-carrier Correlations," *IEEE Transactions on Wireless Communications*, pp. 1–1, 2023.

[19] D. Wang, J. Yang, W. Cui, L. Xie, and S. Sun, "AirFi: Empowering WiFi-based passive human gesture recognition to unseen environment via domain generalization," *IEEE Transactions on Mobile Computing*, vol. 23, no. 2, pp. 1–12, 2022.

[20] Y. Hu, F. Zhang, C. Wu, B. Wang, and K. J. Ray Liu, "DeFall: Environment-Independent Passive Fall Detection Using WiFi," *IEEE Internet of Things Journal*, vol. 9, no. 11, pp. 8515–8530, Jun. 2022.

[21] Y. Zhang, Y. Zheng, K. Qian, G. Zhang, Y. Liu, C. Wu, and Z. Yang, "Widar3.0: Zero-Effort Cross-Domain Gesture Recognition With Wi-Fi," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 11, pp. 8671–8688, Nov. 2022.

[22] H. Abdelnasser, K. Harras, and M. Youssef, "A Ubiquitous WiFi-Based Fine-Grained Gesture Recognition System," *IEEE Transactions on Mobile Computing*, vol. 18, no. 11, pp. 2474–2487, Nov. 2019.

[23] J. Zhang, Z. Tang, M. Li, D. Fang, P. Nurmi, and Z. Wang, "CrossSense: Towards Cross-Site and Large-Scale WiFi Sensing," in *International Conference on Mobile Computing and Networking (MobiCom)*, Oct. 2018, pp. 305–320.

[24] S. Kato, P. Wang, T. Koike-Akino, T. Fujihashi, H. Mansour, and P. T. Boufounos, "Multi-Band Wi-Fi Neural Dynamic Fusion," *IEEE Transactions on Wireless Communications*, vol. 24, no. 5, pp. 3698–3714, 2025.

[25] Y. Li, J. Yang, S.-L. Shih, W.-T. Shih, C.-K. Wen, and S. Jin, "Efficient IoT devices localization through Wi-Fi CSI feature fusion and anomaly detection," *IEEE Internet of Things Journal*, vol. 11, no. 24, pp. 39 306–39 322, 2024.

[26] F. Xiao, S. Zhang, S. Tang, S. Shen, H. Dong, and Y. Zhong, "WiSion: Bolstering MAV 3D Indoor State Estimation by Embracing Multipath of WiFi," *IEEE Transactions on Vehicular Technology*, vol. 72, no. 1, pp. 253–266, 2023.

[27] G. Chi, Z. Yang, J. Xu, C. Wu, J. Zhang, J. Liang, and Y. Liu, "Wi-drone: wi-fi-based 6-DoF tracking for indoor drone flight control," in *International Conference on Mobile Systems, Applications and Services (MobiSys)*, 2022, pp. 56–68.

[28] R. Ayyalasomayajula, A. Arun, C. Wu, S. Sharma, A. R. Sethi, D. Vasisht, and D. Bharadia, "Deep learning based wireless localization for indoor navigation," in *International Conference on Mobile Computing and Networking (MobiCom)*, 2020.

[29] W. Gong and J. Liu, "RoArray: Towards More Robust Indoor Localization Using Sparse Recovery with Commodity WiFi," *IEEE Transactions on Mobile Computing*, vol. 18, no. 6, pp. 1380–1392, Jun. 2019.

[30] L. Zhang and H. Wang, "Device-Free Tracking via Joint Velocity and AOA Estimation With Commodity WiFi," *IEEE Sensors Journal*, vol. 19, no. 22, pp. 10 662–10 673, Nov. 2019.

[31] D. Vasisht, S. Kumar, and D. Katabi, "Decimeter-level localization with a single WiFi access point," in *USENIX Conference on Networked Systems Design and Implementation (NSDI)*, 2016, pp. 165–178.

[32] K. Yan, F. Wang, B. Qian, H. Ding, J. Han, and X. Wei, "Person-in-WiFi 3D: End-to-End Multi-Person 3D Pose Estimation with Wi-Fi," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024, pp. 969–978.

[33] Y. Zhou, A. Zhu, C. Xu, F. Hu, and Y. Li, "PerUnet: Deep Signal Channel Attention in Unet for WiFi-Based Human Pose Estimation," *IEEE Sensors Journal*, vol. 22, no. 20, pp. 19 750–19 760, Oct. 2022.

[34] Y. Ren, Z. Wang, Y. Wang, S. Tan, Y. Chen, and J. Yang, "GoPose: 3D Human Pose Estimation Using WiFi," *ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies (IMWUT)*, vol. 6, no. 2, pp. 1–25, Jul. 2022.

[35] Y. Ren, Z. Wang, S. Tan, Y. Chen, and J. Yang, "Winect: 3D Human Pose Tracking for Free-form Activity Using Commodity WiFi," *ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies (IMWUT)*, vol. 5, no. 4, Dec. 2022.

[36] W. Jiang, H. Xue, C. Miao, S. Wang, S. Lin, C. Tian, S. Murali, H. Hu, Z. Sun, and L. Su, "Towards 3D human pose construction using wifi," in *International Conference on Mobile Computing and Networking (MobiCom)*, 2020.

[37] F. Wang, S. Zhou, S. Panev, J. Han, and D. Huang, "Person-in-WiFi: Fine-grained person perception using WiFi," in *IEEE/CVF International Conference on Computer Vision (ICCV)*, Oct. 2019, pp. 5452–5461.

[38] H. Rizk and A. M. Elmogy, "Self-supervised WiFi-based identity recognition in multi-user smart environments," *Sensors*, vol. 25, May 2025.

[39] L. Deng, J. Yang, S. Yuan, H. Zou, C. X. Lu, and L. Xie, "GaitFi: Robust Device-Free Human Identification via WiFi and Vision Multimodal Learning," *IEEE Internet of Things Journal*, vol. 10, no. 1, pp. 625–636, Jan. 2023.

[40] R. Ou, Y. Chen, and Y. Deng, "WiWalk: Gait-Based Dual-User Identification Using WiFi Device," *IEEE Internet of Things Journal*, vol. 10, no. 6, pp. 5321–5334, Mar. 2023.

[41] Y. Zhang, Y. Zheng, G. Zhang, K. Qian, C. Qian, and Z. Yang, "GaitSense: Towards Ubiquitous Gait-Based Human Identification with Wi-Fi," *ACM Transactions on Sensor Networks*, vol. 18, no. 1, Oct. 2021.

[42] D. Wang, Z. Zhou, X. Yu, and Y. Cao, "CSIID: WiFi-based human identification via deep learning," in *International Conference on Computer Science and Education (ICCSE)*, 2019, pp. 326–330.

[43] T. Xin, B. Guo, Z. Wang, M. Li, Z. Yu, and X. Zhou, "FreeSense: Indoor Human Identification with Wi-Fi Signals," in *IEEE Global Communications Conference (GLOBECOM)*, 2016, pp. 1–7.

[44] J. Zhang, B. Wei, W. Hu, and S. S. Kanhere, "WiFi-ID: Human Identification Using WiFi Signal," in *International Conference on Distributed Computing in Sensor Systems (DCOSS)*, 2016, pp. 75–82.

[45] Z. Guo, W. Yuan, L. Gui, B. Sheng, and F. Xiao, "BreatheBand: A Fine-grained and Robust Respiration Monitor System Using WiFi Signals," *ACM Transactions on Sensor Networks*, vol. 19, no. 4, May 2023.

[46] Y. Li, D. Wu, J. Zhang, X. Xu, Y. Xie, T. Gu, and D. Zhang, "DiverSense: Maximizing Wi-Fi Sensing Range Leveraging Signal Diversity," *ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies (IMWUT)*, vol. 6, no. 2, pp. 1–28, Jul. 2022.

[47] J. Hu, J. Yang, J.-B. Ong, D. Wang, and L. Xie, "ResFi: WiFi-Enabled Device-Free Respiration Detection Based on Deep Learning," in *IEEE International Conference on Control & Automation (ICCA)*, Jun. 2022, pp. 510–515.

[48] Y. Zeng, D. Wu, J. Xiong, J. Liu, Z. Liu, and D. Zhang, "MultiSense: Enabling Multi-person Respiration Sensing with Commodity WiFi," *ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies (IMWUT)*, vol. 4, no. 3, pp. 1–29, Sep. 2020.

[49] Y. Zeng, D. Wu, J. Xiong, E. Yi, R. Gao, and D. Zhang, "FarSense: Pushing the Range Limit of WiFi-based Respiration Sensing with CSI Ratio of Two Antennas," *ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies (IMWUT)*, vol. 3, no. 3, pp. 1–26, Sep. 2019.

[50] Y. Zeng, D. Wu, R. Gao, T. Gu, and D. Zhang, "FullBreathe: Full Human Respiration Detection Exploiting Complementarity of CSI Phase and Amplitude of WiFi Signals," *ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies (IMWUT)*, vol. 2, no. 3, pp. 1–19, Sep. 2018.

[51] X. Wang, C. Yang, and S. Mao, "PhaseBeat: Exploiting CSI Phase Data for Vital Sign Monitoring with Commodity WiFi Devices," in *IEEE International Conference on Distributed Computing Systems (ICDCS)*, Jun. 2017, pp. 1230–1239.

[52] X. Wang, K. Niu, A. Yu, J. Xiong, Z. Yao, J. Wang, W. Li, and D. Zhang, "WiMeasure: Millimeter-level Object Size Measurement with Commodity WiFi Devices," *ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies (IMWUT)*, vol. 7, no. 2, pp. 1–26, Jun. 2023.

[53] Y. Gu, Y. Zhu, J. Li, and Y. Ji, "WiMate: Location-independent material identification based on commercial WiFi devices," in *IEEE Global Communications Conference (GLOBECOM)*, Dec. 2021, pp. 01–06.

[54] C. Shi, T. Zhao, Y. Xie, T. Zhang, Y. Wang, X. Guo, and Y. Chen, "Environment-independent In-baggage Object Identification Using WiFi Signals," in *IEEE International Conference on Mobile Ad Hoc and Smart Systems (MASS)*, Oct. 2021, pp. 71–79.

[55] C. Wang, J. Liu, Y. Chen, H. Liu, and Y. Wang, "Towards in-baggage suspicious object detection using commodity WiFi," in *IEEE Conference on Communications and Network Security (CNS)*, May 2018, pp. 1–9.

[56] S. Yousefi, H. Narui, S. Dayal, S. Ermon, and S. Valaee, "A Survey on Behavior Recognition Using WiFi Channel State Information," *IEEE Communications Magazine*, vol. 55, no. 10, pp. 98–104, Oct. 2017.

[57] E. Ramesh and T. Nishio, "High-resolution efficient image generation from WiFi CSI using a pretrained latent diffusion model," *arXiv preprint arXiv:2506.10605*, 2025.

[58] G. Cao, T. Maekawa, K. Ohara, and Y. Kishino, "Enhancing environmental independency for CSI-based depth image reconstruction," in *IEEE International Conference on Pervasive Computing and Communications Workshops and other Affiliated Events*, Mar. 2025, pp. 537–542.

[59] C. Chen, S. Ohta, T. Nishio, M. Bennis, J. Park, and M. Wahib, "Enabling visual scene recovery from Wi-fi CSI for occlusion-free surveillance," *IEEE Internet of Things Journal*, vol. 12, no. 11, pp. 15 040–15 056, Jun. 2025.

[60] C. Chen, T. Nishio, M. Bennis, and J. Park, "RF-Inpainter: Multimodal image inpainting based on vision and radio signals," *IEEE Access*, vol. 10, pp. 110 689–110 700, 2022.

[61] M. H. Kefayati, V. Pourahmadi, and H. Aghaeinia, "Multi-view WiFi imaging," *Signal Processing*, vol. 197, p. 108552, Aug. 2022.

[62] D. Avola, M. Cascio, L. Cinque, A. Fagioli, and G. L. Foresti, "Human Silhouette and Skeleton Video Synthesis Through Wi-Fi Signals," *International Journal of Neural Systems*, vol. 32, no. 05, p. 2250015, 2022.

[63] M. H. Kefayati, V. Pourahmadi, and H. Aghaeinia, "Wi2Vi: Generating Video Frames From WiFi CSI Samples," *IEEE Sensors Journal*, vol. 20, no. 19, pp. 11 463–11 473, Oct. 2020.

[64] G. Zhu, Y. Hu, W. Gao, W.-H. Wang, B. Wang, and K. J. R. Liu, "CSI-Bench: A large-scale in-the-wild dataset for multi-task WiFi sensing," *arXiv preprint arXiv:2505.21866*, May 2025.

[65] S. Huang, K. Li, D. You, Y. Chen, A. Lin, S. Liu, X. Li, and J. A. McCann, "WiMANS: A Benchmark Dataset for WiFi-Based Multi-user Activity Sensing," in *European Conference on Computer Vision (ECCV)*, 2024, pp. 72–91.

[66] F. Wang, Y. Lv, M. Zhu, H. Ding, and J. Han, "XRF55: A radio frequency dataset for human indoor action analysis," *ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies (IMWUT)*, vol. 8, no. 1, pp. 1–34, Mar. 2024.

[67] F. Meneghello, C. Chen, C. Cordeiro, and F. Restuccia, "Toward integrated sensing and communications in IEEE 802.11bf WI-fi networks," *IEEE Communications Magazine*, vol. 61, no. 7, pp. 128–133, Jul. 2023.

[68] F. Meneghello, N. D. Fabbro, D. Garlisi, I. Tinnirello, and M. Rossi, "A CSI Dataset for Wireless Human Sensing on 80 MHz Wi-Fi Channels," *IEEE Communications Magazine*, vol. 61, no. 9, pp. 146–152, 2023.

[69] J. Yang, H. Huang, Y. Zhou, X. Chen, Y. Xu, S. Yuan, H. Zou, C. X. Lu, and L. Xie, "MM-Fi: multi-modal non-intrusive 4D human dataset for versatile wireless sensing," *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

[70] L. Guo, S. Guo, L. Wang, C. Lin, J. Liu, B. Lu, J. Fang, Z. Liu, Z. Shan, and J. Yang, "Wiar: A Public Dataset for Wifi-Based Activity Recognition," *IEEE Access*, vol. 7, pp. 154 935–154 945, 2019.

[71] Y. Ma, G. Zhou, S. Wang, H. Zhao, and W. Jung, "SignFi: Sign Language Recognition Using WiFi," *ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies (IMWUT)*, vol. 2, no. 1, pp. 1–21, Mar. 2018.

[72] C. J. Vaca-Rubio, P. Wang, T. Koike-Akino, Y. Wang, P. Boufounos, and P. Popovski, "Object Trajectory Estimation With Continuous-Time Neural Dynamic Learning of Millimeter-Wave Wi-Fi," *IEEE Journal of Selected Topics in Signal Processing*, vol. 18, no. 5, pp. 796–811, 2024.

[73] A. Blanco, P. J. Mateo, F. Gringoli, and J. Widmer, "Augmenting mmWave localization accuracy through sub-6 GHz on off-the-shelf devices," in *International Conference on Mobile Systems, Applications and Services (MobiSys)*, Jun. 2022, pp. 477–490.

[74] T. Koike-Akino, P. Wang, M. Pajovic, H. Sun, and P. V. Orlik, "Fingerprinting-Based Indoor Localization With Commercial MMWave WiFi: A Deep Learning Approach," *IEEE Access*, vol. 8, pp. 84 879–84 892, 2020.

[75] D. Garcia, J. O. Lacruz, P. Jiménez Mateo, and J. Widmer, "POLAR: Passive object localization with IEEE 802.11ad using phased antenna arrays," in *IEEE International Conference on Computer Communications (INFOCOM)*, Jul. 2020, pp. 1838–1847.

[76] A. Vashist, D. R. Bhanushali, R. Relyea, C. Hochgraf, A. Ganguly, P. D. Sai Manoj, R. Ptucha, A. Kwasinski, and M. E. Kuhl, "Indoor wireless localization using consumer-grade 60 GHz equipment with machine learning for intelligent material handling," in *IEEE International Conference on Consumer Electronics (ICCE)*, Jan. 2020, pp. 1–6.

[77] J. Yu, P. Wang, T. Koike-Akino, Y. Wang, P. V. Orlik, and H. Sun, "Human Pose and Seat Occupancy Classification with Commercial MMWave WiFi," in *IEEE Globecom Workshops*, Dec. 2020, pp. 1–6.

[78] P. Wang, M. Pajovic, T. Koike-Akino, H. Sun, and P. V. Orlik, "Fingerprinting-based indoor localization with commercial mmWave WiFi - part II: Spatial beam SNRs," in *IEEE Global Communications Conference (GLOBECOM)*, Dec. 2019, pp. 1–6.

[79] N. N. Bhat, R. Berkvens, and J. Famaey, "Gesture Recognition with mmWave Wi-Fi Access Points: Lessons Learned," in *IEEE International Symposium on a World of Wireless, Mobile and Multimedia Networks (WoWMoM)*, 2023, pp. 127–136.

[80] ——, "CSI4Free: GAN-Augmented mmWave CSI for Improved Pose Classification," in *IEEE International Symposium on Joint Communications and Sensing (JC&S)*, vol. 27, Mar. 2024, pp. 1–6.

[81] T. Koike-Akino, P. Wang, and Y. Wang, "Quantum Transfer Learning for Wi-Fi Sensing," in *IEEE International Conference on Communications (ICC)*, May 2022, pp. 654–659.

[82] S. Kato, T. Murakami, T. Fujihashi, T. Watanabe, and S. Saruwatari, "CBR-ACE: Counting Human Exercise using Wi-Fi Beamforming Reports," *Journal of Information Processing*, vol. 30, pp. 66–74, 2022.

[83] R. Xiao, X. Chen, Y. He, J. Han, and J. Han, "Lend me your beam: Privacy implications of plaintext beamforming feedback in WiFi," *Symposium on Network and Distributed System Security (NDSS)*, 2025.

[84] K. F. Haque, F. Meneghello, and F. Restuccia, "Wi-BFI: Extracting the IEEE 802.11 beamforming feedback information from commercial WI-fi devices," in *ACM Workshop on Wire-*

less Network Testbeds, Experimental evaluation & Characterization (WiNTECH), Oct. 2023, pp. 104–111.

[85] F. Tian, Y. Jiang, C. Bao, B. Peng, F. Hong, and Z. Guo, "CBR-HAR: Real-time human activity recognition based on CBR," *IEEE Wireless Communications Letters*, vol. 14, no. 7, pp. 2179–2183, Jul. 2025.

[86] S. Chen, H. Jiang, J. Xiong, H. Jingyang, P. Wang, C. Liu, Z. Xiao, and B. Li, "BeamCount: Indoor Crowd Counting Using Wi-Fi Beamforming Feedback Information," in *International Symposium on Theory, Algorithmic Foundations, and Protocol Design for Mobile Networks and Mobile Computing (MobiHoc)*, 2024, pp. 1–10.

[87] N. Suraweera, A. Winter, J. Sorensen, S. Li, M. Johnson, I. B. Collings, S. V. Hanly, W. Ni, and M. Hedley, "Passive Through-Wall Counting of People Walking Using WiFi Beamforming Reports," *IEEE Systems Journal*, pp. 1–7, 2020.

[88] N. Suraweera, A. Winter, J. Sorensen, M. Johnson, S. Li, I. B. Collings, S. V. Hanly, W. Ni, and M. Hedley, "Stand-off Detection of Human Presence and Movement Using IEEE 802.11ac Beamforming Reports," in *International Conference on Signal Processing and Communication Systems (ICSPCS)*, Dec. 2019, pp. 1–7.

[89] E. Yi, D. Wu, J. Xiong, F. Zhang, K. Niu, W. Li, and D. Zhang, "BFMSense: WiFi Sensing Using Beamforming Feedback Matrix," in *USENIX Symposium on Networked Systems Design and Implementation (NSDI)*, Apr. 2024, pp. 1697–1712.

[90] C. Wu, X. Huang, J. Huang, and G. Xing, "Enabling Ubiquitous WiFi Sensing with Beamforming Reports," in *ACM Special Interest Group on Data Communication (SIGCOMM)*, 2023, pp. 20–32.

[91] S. Kondo, S. Itahara, K. Yamashita, K. Yamamoto, Y. Koda, T. Nishio, and A. Taya, "Bi-directional beamforming feedback-based firmware-agnostic WiFi sensing: An empirical study," *IEEE Access*, vol. 10, pp. 36 924–36 934, 2022.

[92] T. Kanda, S. Kondo, H. Shimomura, T. Sato, H. Awano, and K. Yamamoto, "Beamforming feedback-based respiration and heart rate estimation toward firmware-agnostic WiFi sensing," *IEEE Access*, vol. 13, pp. 146 008–146 019, 2025.

[93] X. Wang, J. Wang, K. Niu, J. Xiong, F. Zhang, E. Yi, A. Yu, Z. Yao, and D. Zhang, "Wi2DMeasure: WiFi-based 2D Object Size Measurement," in *ACM Conference on Embedded Networked Sensor Systems (SenSys)*, Nov. 2024, pp. 253–266.

[94] S. Kato, T. Fukushima, T. Murakami, H. Abeysekera, Y. Iwasaki, T. Fujihashi, T. Watanabe, and S. Saruwatari, "CSI2Image: Image Reconstruction From Channel State Information Using Generative Adversarial Networks," *IEEE Access*, vol. 9, pp. 47 154–47 168, 2021.

[95] M. Xue, W. Sun, H. Yu, H. Tang, A. Lin, X. Zhang, and R. Zimmermann, "Locate the mobile device by enhancing the WiFi-based indoor localization model," *IEEE Internet of Things Journal*, vol. 6, no. 5, pp. 8792–8803, Oct. 2019.

[96] M. T. Hoang, B. Yuen, X. Dong, T. Lu, R. Westendorp, and K. Reddy, "Recurrent Neural Networks for Accurate RSSI Indoor Localization," *IEEE Internet of Things Journal*, vol. 6, no. 6, pp. 10 639–10 651, Dec. 2019.

[97] S. Kianoush, S. Savazzi, F. Vicentini, V. Rampa, and M. Giussani, "Device-free RF human body fall detection and localization in industrial workplaces," *IEEE Internet of Things Journal*, vol. 4, no. 2, pp. 351–362, Apr. 2017.

[98] D. Halperin, W. Hu, A. Sheth, and D. Wetherall, "Tool release: gathering 802.11n traces with channel state information," *ACM SIGCOMM Computer Communication Review*, vol. 41, no. 1, p. 53, Jan. 2011.

[99] X. Shi, M. Bouazizi, and T. Ohtsuki, "MmPoint-attention: A unified attention framework for human pose and activity recognition from mmWave radar point clouds," *IEEE Sensors Journal*, vol. 25, pp. 26 794–26 805, Jul. 2025.

[100] F. Ding, Z. Luo, P. Zhao, and C. X. Lu, "milliFlow: Scene Flow Estimation on mmWave Radar Point Cloud for Human Motion Sensing," in *European Conference on Computer Vision (ECCV)*, 2025, pp. 202–221.

[101] A. Liu, Y.-T. Lin, and K. Sundaresan, "View-agnostic human exercise cataloging with single MmWave radar," *ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies (IMWUT)*, vol. 8, no. 3, pp. 1–23, Aug. 2024.

[102] Z. Wu, Z. Cao, X. Yu, J. Zhu, C. Song, and Z. Xu, "A novel multiperson activity recognition algorithm based on point clouds measured by millimeter-wave MIMO radar," *IEEE Sensors Journal*, vol. 23, pp. 19 509–19 523, Sep. 2023.

[103] F. Jin, A. Sengupta, and S. Cao, "mmFall: Fall Detection Using 4-D mmWave Radar and a Hybrid Variational RNN AutoEncoder," *IEEE Transactions on Automation Sceience and Engineering*, vol. 19, no. 2, pp. 1245–1257, Apr. 2022.

[104] C. Yu, Z. Xu, K. Yan, Y.-R. Chien, S.-H. Fang, and H.-C. Wu, "Noninvasive human activity recognition using millimeter-wave radar," *IEEE Systems Journal*, vol. 16, no. 2, pp. 3036–3047, Jun. 2022.

[105] A. D. Singh, S. S. Sandha, L. Garcia, and M. Srivastava, "RadHAR: Human Activity Recognition from Point Clouds Generated through a Millimeter-wave Radar," in *ACM Workshop on Millimeter-Wave Networks and Sensing Systems (mmNets)*, 2019, pp. 51–56.

[106] P. Song, L. Mei, and H. Cheng, "Human semantic segmentation using millimeter-wave radar sparse point clouds," in *International Conference on Computer Supported Cooperative Work in Design (CSCWD)*, May 2023, pp. 1275–1280.

[107] P. Zhao, C. X. Lu, J. Wang, C. Chen, W. Wang, N. Trigoni, and A. Markham, "Human tracking and identification through a millimeter wave radar," *Ad Hoc Networks*, vol. 116, no. 102475, p. 102475, May 2021.

[108] J. Fan, J. Yang, Y. Xu, and L. Xie, "Diffusion Model Is a Good Pose Estimator from 3D RF-Vision," in *European Conference on Computer Vision (ECCV)*, 2025, pp. 1–18.

[109] J. Zhu, X. Huang, Z. Deng, and Y. Qiu, "MRadHPRS: Human pose recognition system from point clouds generated through a millimeter-wave radar," *IEEE Transactions on Aerospace and Electronic Systems*, vol. 61, no. 2, pp. 3227–3242, Apr. 2025.

[110] W. Li, R. Liu, S. Wang, D. Cao, and W. Jiang, "Egocentric human pose estimation using head-mounted mmWave radar," in *ACM Conference on Embedded Networked Sensor Systems (SenSys)*, Nov. 2023, pp. 431–444.

[111] H. Xue, Q. Cao, Y. Ju, H. Hu, H. Wang, A. Zhang, and L. Su, "M ^{4} esh: MmWave-based 3D human mesh construction for multiple subjects," in *ACM Conference on Embedded Networked Sensor Systems (SenSys)*, Nov. 2022, pp. 391–406.

[112] H. Xue, Y. Ju, C. Miao, Y. Wang, S. Wang, A. Zhang, and L. Su, "mmMesh: towards 3D real-time dynamic human mesh construction using millimeter-wave," in *International Conference on Mobile Systems, Applications, and Services (MobiSys)*, 2021, pp. 269–282.

[113] A. Sengupta, F. Jin, R. Zhang, and S. Cao, "mm-Pose: Real-Time Human Skeletal Posture Estimation Using mmWave Radars and CNNs," *IEEE Sensors Journal*, vol. 20, pp. 10 032–10 044, Nov. 2019.

[114] H. Liu, X. Liu, X. Xie, X. Tong, and K. Li, "PmTrack: Enabling Personalized mmWave-based human tracking," *ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies (IMWUT)*, vol. 7, no. 4, pp. 1–30, Dec. 2023.

[115] B. Amjad, Q. Z. Ahmed, P. I. Lazaridis, F. A. Khan, M. Hafeez, and Z. D. Zaharis, "Deep learning approach for optimal localization using an mm-wave sensor," *IEEE Transactions on Instrumentation and Measurement*, vol. 72, pp. 1–15, 2023.

[116] J. Pegoraro and M. Rossi, "Real-time people tracking and identification from sparse mm-wave radar point-clouds," *IEEE Access*, vol. 9, pp. 78 504–78 520, May 2021.

[117] H. Wang, J. Ma, Z. Huang, and K. Liu, "Multitarget vital signs estimation based on millimeter-wave radar in complex scenes," *IEEE Sensors Journal*, vol. 24, pp. 39 432–39 442, Dec. 2024.

[118] L. Engel, J. Mueller, E. J. F. Rendon, E. Dorschky, D. Krauss, I. Ullmann, B. M. Eskofier, and M. Vossiek, "Advanced millimeter wave radar-based human pose estimation enabled by a deep learning neural network trained with optical motion capture ground truth data," *IEEE Journal of Microwaves*, vol. 5, no. 2, pp. 373–387, Mar. 2025.

[119] W. Brescia, P. Gomes, L. Toni, S. Mascolo, and L. De Cicco, "MilliNoise: A millimeter-wave radar sparse point cloud dataset in indoor scenarios," in *ACM Multimedia Systems Conference (MMSys)*, Apr. 2024, pp. 422–428.

[120] H. Cui, S. Zhong, J. Wu, Z. Shen, N. Dahnoun, and Y. Zhao, "MiliPoint: A Point Cloud Dataset for mmWave Radar," *Advences in Neural Information Processing Systems (NeurIPS)*, pp. 62 713–62 726, Sep. 2023.

[121] A. Chen, X. Wang, S. Zhu, Y. Li, J. Chen, and Q. Ye, "mmBody Benchmark: 3D Body Reconstruction Dataset and Analysis for Millimeter Wave Radar," in *ACM International Conference on Multimedia (ACM-MM)*, 2022, pp. 3501–3510.

[122] S. An, Y. Li, and U. Ogras, "mRI: multi-modal 3D human Pose estimation dataset using mmWave, RGB-D, and inertial sensors," *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

[123] S. An and U. Y. Ogras, "MARS: mmWave-based assistive rehabilitation system for smart healthcare," *ACM Transactions on Embedded Computing Systems*, vol. 20, no. 5s, pp. 1–22, Oct. 2021.

[124] C. X. Lu, S. Rosa, P. Zhao, B. Wang, C. Chen, J. A. Stankovic, N. Trigoni, and A. Markham, "See through smoke: robust indoor mapping with low-cost mmWave radar," in *International Conference on Mobile Systems, Applications, and Services (MobiSys)*, 2020, pp. 14–27.

[125] Y. Liu, J. Zhang, Y. Chen, W. Wang, S. Yang, X. Na, Y. Sun, and Y. He, "Real-time continuous activity recognition with a commercial mmWave radar," *IEEE Transactions on Mobile Computing*, vol. 24, no. 3, pp. 1684–1698, Mar. 2025.

[126] B. Jin, X. Ma, B. Hu, Z. Zhang, Z. Lian, and B. Wang, "Gesture-mmWAVE: Compact and accurate millimeter-wave radar-based dynamic gesture recognition for embedded devices," *IEEE Transactions on Human-Machine Systems*, vol. 54, no. 3, pp. 337–347, Jun. 2024.

[127] P. Zhao, C. X. Lu, B. Wang, N. Trigoni, and A. Markham, "CubeLearn: End-to-end learning for human motion recognition from raw mmWave radar signals," *IEEE Internet of Things Journal*, vol. 10, no. 12, pp. 10 236–10 249, Jun. 2023.

[128] E. Kurtoglu, A. C. Gurbuz, E. A. Malaia, D. Griffin, C. Crawford, and S. Z. Gurbuz, "ASL trigger recognition in mixed activity/signing sequences for RF sensor-based user interfaces," *IEEE Transactions on Human-Machine Systems*, vol. 52, no. 4, pp. 699–712, Aug. 2022.

[129] A. Ali, P. Parida, V. Va, S. Ni, K. N. Nguyen, B. L. Ng, and J. C. Zhang, "End-to-end dynamic gesture recognition using MmWave radar," *IEEE Access*, vol. 10, pp. 88 692–88 706, 2022.

[130] R. Yataka, A. Cardace, P. Wang, P. Boufounos, and R. Takahashi, "RETR: Multi-view Radar dEtection TRansformer for indoor perception," *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 19 839–19 869, Nov. 2024.

[131] Z. Wu, D. Zhang, C. Xie, C. Yu, J. Chen, Y. Hu, and Y. Chen, "RFMask: A Simple Baseline for Human Silhouette Segmentation With Radio Signals," *IEEE Transactions on Multimedia*, vol. 25, pp. 4730–4741, 2023.

[132] B. Vandersmissen, N. Knudde, A. Jalalvand, I. Couckuyt, A. Bourdoux, W. De Neve, and T. Dhaene, "Indoor person identification using a low-power FMCW radar," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 56, no. 7, pp. 3941–3952, Jul. 2018.

[133] Y. Xu, X. Wang, J. J. Kupiainen, J. Säe, J. Boutellier, J. Nurmi, and B. Tan, "Tracking the occluded indoor target with scattered millimeter wave signal," *IEEE Sensors Journal*, vol. 24, no. 22, pp. 38 102–38 112, Nov. 2024.

[134] W. Chen, H. Yang, X. Bi, R. Zheng, F. Zhang, P. Bao, Z. Chang, X. Ma, and D. Zhang, "Environment-aware multi-person tracking in indoor environments with MmWave radars," *ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies (IMWUT)*, vol. 7, no. 3, pp. 1–29, Sep. 2023.

[135] J. Pegoraro, F. Meneghello, and M. Rossi, "Multiperson continuous tracking and identification from mm-wave micro-Doppler signatures," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 59, no. 4, pp. 2994–3009, Apr. 2021.

[136] S. Kato, R. Yataka, P. P. Wang, P. Miraldo, T. Fujihashi, and P. Boufounos, "RAPTR: Radar-based 3D Pose Estimation using Transformer," *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 1–13, 2025.

[137] B. Zhu, Z. He, W. Xiong, G. Ding, T. Huang, and W. Xiang, "ProbRadarM3F: mmWave radar-based human skeletal pose estimation with probability map guided multi-format feature fusion," *IEEE Transactions on Aerospace and Electronic Systems*, pp. 1–11, 2025.

[138] H. Wan, R. Song, C. Xie, Z. Lu, Q. Chen, Z. Wu, D. Zhang, Y. Hu, and Y. Chen, "QRF-Pose: Query-Based 3D Pose Estimation using Radio Signals," in *International Conference on Wireless Communications and Signal Processing (WCSP)*, 2024, pp. 1271–1277.

[139] C. Xie, D. Zhang, Z. Wu, C. Yu, Y. Hu, and Y. Chen, "RPM 2.0: RF-Based Pose Machines for Multi-Person 3D Pose Estimation," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 34, no. 1, pp. 490–503, Jan. 2024.

[140] Y. H. Ho, J. H. Cheng, S. Y. Kuan, Z. Jiang, W. Chai, H. W. Huang, C. L. Lin, and J. N. Hwang, "RT-Pose: A 4D Radar Tensor-Based 3D Human Pose Estimation and Localization Benchmark," in *European Conference on Computer Vision (ECCV)*, 2024, pp. 107–125.

[141] S. P. Lee, N. P. Kini, W. H. Peng, C. W. Ma, and J. N. Hwang, "HuPR: A benchmark for human pose estimation using millimeter wave radar," in *IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, Jan. 2023, pp. 5715–5724.

[142] H. Kong, X. Xu, J. Yu, Q. Chen, C. Ma, Y. Chen, Y.-C. Chen, and L. Kong, "m3Track: mmWave-based Multi-user 3D posture tracking," in *International Conference on Mobile Systems, Applications and Services (MobiSys)*, 2022, pp. 491–503.

[143] M. Zhao, Y. Tian, H. Zhao, M. A. Alsheikh, T. Li, R. Hristov, Z. Kabelac, D. Katabi, and A. Torralba, "RF-based 3D skeletons," in *ACM Special Interest Group on Data Communication (SIGCOMM)*, Aug. 2018, pp. 267–281.

[144] Z. Chang, F. Zhang, X. Ma, P. Wang, W. Chen, D. Zhang, B. Jouaber, and D. Zhang, "MmECare: Enabling fine-grained vital sign monitoring for emergency care with handheld MmWave radars," *ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies (IMWUT)*, vol. 8, no. 4, pp. 1–24, Nov. 2024.

[145] Q. Hu, Q. Zhang, H. Lu, S. Wu, Y. Zhou, Q. Huang, H. Chen, Y.-C. Chen, and N. Zhao, "Contactless arterial blood pressure waveform monitoring with mmWave radar," *ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies (IMWUT)*, vol. 8, no. 4, pp. 1–29, Nov. 2024.

[146] J. Chen, D. Zhang, Z. Wu, F. Zhou, Q. Sun, and Y. Chen, "Contactless electrocardiogram monitoring with millimeter wave radar," *IEEE Transactions on Mobile Computing*, vol. 23, no. 1, pp. 270–285, Jan. 2024.

[147] M. Alizadeh, G. Shaker, J. C. M. D. Almeida, P. P. Morita, and S. Safavi-Naeini, "Remote monitoring of human vital signs using mm-wave FMCW radar," *IEEE Access*, vol. 7, pp. 54 958–54 968, 2019.

[148] M. M. Rahman, R. Yataka, S. Kato, P. Wang, P. Li, A. Cardace, and P. Boufounos, "MMVR: Millimeter-wave multi-view radar dataset and benchmark for indoor perception," in *European Conference on Computer Vision (ECCV)*, 2024, pp. 306–322.

[149] Y. Xie, Z. Li, and M. Li, "Precise Power Delay Profiling with Commodity WiFi," in *International Conference on Mobile Computing and Networking (MobiCom)*, 2015, pp. 53–64.

[150] S. M. Hernandez and E. Bulut, "Lightweight and standalone IoT based WiFi sensing for active repositioning and mobility," in *IEEE International Symposium on a World of Wireless, Mobile and Multimedia Networks (WoWMoM)*, Aug. 2020, pp. 277–286.

[151] F. Gringoli, M. Schulz, J. Link, and M. Hollick, "Free Your CSI: A Channel State Information Extraction Platform For Modern Wi-Fi Chipsets," in *International Workshop on Wireless Network Testbeds, Experimental Evaluation and Characterization (WiNTECH)*, 2019, pp. 21–28.

[152] Z. Jiang, T. H. Luan, X. Ren, D. Lv, H. Hao, J. Wang, K. Zhao, W. Xi, Y. Xu, and R. Li, "Eliminating the Barriers: Demystifying Wi-Fi Baseband Design and Introducing the PicoScenes Wi-Fi Sensing Platform," *IEEE Internet of Things Journal*, vol. 9, no. 6, pp. 4476–4496, Mar. 2022.

[153] A. Zubow, P. Gawłowicz, and F. Dressler, "On Phase Offsets of 802.11 ac Commodity WiFi," in *Wireless On-demand Network Systems and Services Conference (WONS)*, 2021, pp. 1–4.

[154] S. Aditya, A. F. Molisch, and H. M. Behairy, "A survey on the impact of multipath on wideband time-of-arrival based localization," *Proceedings of IEEE*, vol. 106, no. 7, pp. 1183–1203, Jul. 2018.

[155] Z. Xia, X. Pan, S. Song, L. E. Li, and G. Huang, "Vision Transformer with Deformable Attention," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 4784–4793.

[156] C. Qi, H. Su, K. Mo, and L. Guibas, "PointNet: Deep Learning on Point Sets for 3D Classification and Segmentation," in *IEEE Conference on Computer Vision adn Pattern Recognition (CVPR)*, Dec. 2016, pp. 77–85.

[157] F.-B. Ueng, H.-F. Wang, R. Chang, and L.-D. Jeng, "Zero Forcing and Minimum Mean Square Error Equalization for OFDM-CDMA Multiuser Detection in Multipath Fading Channels," in *International Symposium on Intelligent Signal Processing and Communications*, 2006, pp. 505–509.

[158] *IEEE Standard for Information Technology–Telecommunications and Information Exchange between Systems Local and Metropolitan Area Networks–Specific Requirements–Part 11: Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specifications–Amendment 1: Enhancements for High-Efficiency WLAN*, Std. IEEE Std 802.11ax-2021, 2021.

[159] *IEEE Standard for Information technology – Telecommunications and information exchange between systemsLocal and metropolitan area networks – Specific requirements – Part 11: Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specifications – Amendment 4: Enhancements for Very High Throughput for Operation in Bands below 6 GHz*, Std. IEEE Std 802.11ac-2013, 2021.

[160] H. Caesar, V. Bankiti, A. H. Lang, S. Vora, V. E. Liong, Q. Xu, A. Krishnan, Y. Pan, G. Baldan, and O. Beijbom, "nuScenes: A Multimodal Dataset for Autonomous Driving," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 11 618–11 628.

[161] O. Schumann, M. Hahn, N. Scheiner, F. Weishaupt, J. F. Tilly, J. Dickmann, and C. Wöhler, "RadarScenes: A Real-World Radar Point Cloud Data Set for Automotive Applications," *arXiv preprint arXiv:2104.02493*, 2021.

[162] M. Sheeny, E. De Pellegrin, S. Mukherjee, A. Ahrabian, S. Wang, and A. Wallace, "RADIATE: A Radar Dataset for Automotive Perception in Bad Weather," in *IEEE International Conference on Robotics and Automation (ICRA)*, 2021, pp. 1–7.

[163] D. Barnes, M. Gadd, P. Murcutt, P. Newman, and I. Posner, "The Oxford Radar RobotCar Dataset: A Radar Extension to the Oxford RobotCar Dataset," in *IEEE International Conference on Robotics and Automation (ICRA)*, 2020, pp. 6433–6438.

[164] G. Kim, Y. S. Park, Y. Cho, J. Jeong, and A. Kim, "MulRan: Multimodal Range Dataset for Urban Place Recognition," in *IEEE International Conference on Robotics and Automation (ICRA)*, 2020, pp. 6246–6253.

[165] A. Ouaknine, A. Newson, J. Rebut, F. Tupin, and P. Pérez, "CARRADA Dataset: Camera and Automotive Radar with Range- Angle- Doppler Annotations," in *International Conference on Pattern Recognition (ICPR)*, 2021, pp. 5068–5075.

[166] Y. Wang, Z. Jiang, Y. Li, J.-N. Hwang, G. Xing, and H. Liu, "RODNet: A Real-Time Radar Object Detection Network Cross-Supervised by Camera-Radar Fused Object 3D Localization," *IEEE Journal of Selected Topics in Signal Processing*, vol. 15, no. 4, pp. 954–967, 2021.

[167] T.-Y. Lim, S. A. Markowitz, and M. N. Do, "RaDICaL: A synchronized FMCW radar, depth, IMU and RGB camera data dataset with low-level FMCW radar signals," *IEEE J. Sel. Top. Signal Process.*, vol. 15, no. 4, pp. 941–953, Jun. 2021.

[168] J. Rebut, A. Ouaknine, W. Malik, and P. Pérez, "Raw High-Definition Radar for Multi-Task Learning," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 17 000–17 009.

[169] A. Palffy, E. Pool, S. Baratam, J. F. P. Kooij, and D. M. Gavrila, "Multi-Class Road User Detection With 3+1D Radar in the View-of-Delft Dataset," *IEEE Robotics and Automation Letters*, vol. 7, no. 2, pp. 4961–4968, 2022.

[170] D. H. Paek, S. H. Kong, and K. T. Wijaya, "K-Radar: 4D Radar Object Detection for Autonomous Driving in Various Weather Conditions," *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 35, pp. 3819–3829, 2022.

[171] P. Li, P. Wang, K. Berntorp, and H. Liu, "Exploiting Temporal Relations on Radar Perception for Autonomous Driving," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 17 050–17 059.

[172] I. Bilik, O. Longman, S. Villeval, and J. Tabrikian, "The Rise of Radar for Autonomous Vehicles: Signal Processing Solutions and Future Research Directions," *IEEE Signal Processing Magazine*, vol. 36, no. 5, pp. 20–31, 2019.

[173] A. Pandharipande, C.-H. Cheng, J. Dauwels, S. Z. Gurbuz, J. Ibanez-Guzman, G. Li, A. Piazzoni, P. Wang, and A. Santra, "Sensing and Machine Learning for Automotive Perception: A Review," *IEEE Sensors Journal*, vol. 23, no. 11, pp. 11 097–11 115, 2023.

[174] R. Yataka, P. Wang, P. Boufounos, and R. Takahashi, "SIRA: Scalable Inter-frame Relation and Association for Radar Perception," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024, pp. 15 024–15 034.

[175] F. Adib, C.-Y. Hsu, H. Mao, D. Katabi, and F. Durand, "Capturing the human figure through a wall," *ACM Transactions on Graphics*, vol. 34, no. 6, 2015.

[176] L. L. Scharf and C. Demeure, *Statistical signal processing : detection, estimation, and time series analysis*, 1991.

[177] B. Cheng, I. Misra, A. G. Schwing, A. Kirillov, and R. Girdhar, "Masked-attention Mask Transformer for Universal Image Segmentation," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 1280–1289.

[178] K. Sun, B. Xiao, D. Liu, and J. Wang, "Deep High-Resolution Representation Learning for Human Pose Estimation," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 5686–5696.

[179] K. Chen, J. Wang, J. Pang, Y. Cao, Y. Xiong, X. Li, S. Sun, W. Feng, Z. Liu, J. Xu, Z. Zhang, D. Cheng, C. Zhu, T. Cheng, Q. Zhao, B. Li, X. Lu, R. Zhu, Y. Wu, J. Dai, J. Wang, J. Shi, W. Ouyang, C. C. Loy, and D. Lin, "MMDetection: Open MMLab Detection Toolbox and Benchmark," *arXiv preprint arXiv:1906.07155*, Jun. 2019.

[180] M. Contributors, "OpenMMLab Pose Estimation Toolbox and Benchmark," https://github.com/open-mmlab/mmpose, 2020.

[181] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The Pascal Visual Object Classes (VOC) Challenge," *International Journal of Computer Vision*, vol. 88, no. 2, pp. 303–338, 2010.

[182] T. Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft COCO: Common Objects in Context," in *European Conference on Computer Vision (ECCV)*, 2014, pp. 740–755.

[183] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh, "Realtime Multi-person 2D Pose Estimation Using Part Affinity Fields," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 1302–1310.

[184] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-End Object Detection with Transformers," in *European Conference on Computer Vision (ECCV)*, 2020, pp. 213–229.

[185] M. M. Rahman, R. Yataka, P. Wang, S. Kato, P. Li, A. Cardace, and P. Boufounos, "Millimeter-wave Multi-View Radar (MMVR) Dataset," https://zenodo.org/records/12611978, 2024.

[186] T. Zhang, T. Song, D. Chen, T. Zhang, and J. Zhuang, "WiGrus: A Wi-Fi-Based Gesture Recognition System Using Software-Defined Radio," *IEEE Access*, vol. 7, pp. 131 102–131 113, 2019.

[187] X. Guo, B. Liu, C. Shi, H. Liu, Y. Chen, and M. C. Chuah, "WiFi-Enabled Smart Human Dynamics Monitoring," in *ACM Conference on Embedded Network Sensor Systems (SenSys)*, 2017.

[188] B. U. Park and J. S. Marron, "Comparison of Data-Driven Bandwidth Selectors," *Journal of the American Statistical Association*, vol. 85, no. 409, pp. 66–72, Mar 1990.

[189] B. W. Silverman, *Density estimation for statistics and data analysis*, 1986, vol. 26.

[190] G. R. Terrell and D. W. Scott, "Oversmoothed Nonparametric Density Estimates," *Journal of the American Statistical Association*, vol. 80, no. 389, pp. 209–214, 1985.

[191] L. Guo, Z. Lu, X. Wen, S. Zhou, and Z. Han, "From Signal to Image: Capturing Fine-Grained Human Poses with Commodity Wi-Fi," *IEEE Communications Letters*, vol. 24, no. 4, pp. 802–806, April 2020.

[192] I. Goodfellow, J. P. Abadie, M. Mirza, B. Xu, D. W. Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative Adversarial Nets," *Advances in Neural Information Processing Systems (NIPS)*, pp. 2672–2680, December 2014.

[193] Z. Wang, Q. She, and T. E. Ward, "Generative Adversarial Networks in Computer Vision: A Survey and Taxonomy," *ACM Computing Surveys*, vol. 54, no. 2, Feb. 2021.

[194] Z. Pan, W. Yu, X. Yi, A. Khan, F. Yuan, and Y. Zheng, "Recent Progress on Generative Adversarial Networks (GANs): A Survey," *IEEE Access*, vol. 7, pp. 36 322–36 333, March 2019.

[195] A. Radford, L. Metz, and S. Chintala, "Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks," *arXiv preprint arXiv:1511.06434*, Jan. 2016.

[196] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You Only Look Once: Unified, Real-Time Object Detection," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016, pp. 779–788.

[197] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image Quality Assessment: From Error Visibility to Structural Similarity," *IEEE Transactions on Image Processing*, vol. 13, no. 4, pp. 600–612, April 2004.

[198] S. Yao, R. Guan, Z. Peng, C. Xu, Y. Shi, W. Ding, E. G. Lim, Y. Yue, H. Seo, K. L. Man, J. Ma, X. Zhu, and Y. Yue, "Exploring Radar Data Representations in Autonomous Driving: A Comprehensive Review," *arXiv preprint arXiv:2312.04861*, 2024.

[199] M. Skog, O. Kotlyar, V. Kubelka, and M. Magnusson, "Human Detection from 4D Radar Data in Low-Visibility Field Conditions," *arXiv preprint arXiv:2404.05307*, 2024.

[200] A. Santra, P. Wang, G. Shaker, B. S. Mysore, G. Dolmans, Y. Chen, N. Shariati, and A. Pandharipande, "Machine Learning-Powered Radio Frequency Sensing: A Review," *IEEE Sensors Journal*, vol. 25, no. 13, pp. 23 164–23 183, 2025.

[201] J. Wang, K. Sun, T. Cheng, B. Jiang, C. Deng, Y. Zhao, D. Liu, Y. Mu, M. Tan, X. Wang, W. Liu, and B. Xiao, "Deep High-Resolution Representation Learning for Visual Recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 43, no. 10, pp. 3349–3364, 2021.

[202] M. Zhao, T. Li, M. A. Alsheikh, Y. Tian, H. Zhao, A. Torralba, and D. Katabi, "Through-Wall Human Pose Estimation Using Radio Signals," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 7356–7365.

[203] D. Shi, X. Wei, L. Li, Y. Ren, and W. Tan, "End-to-End Multi-Person Pose Estimation with Transformers," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 11 059–11 068.

[204] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778.

[205] X. Zhu, W. Su, L. Lu, B. Li, X. Wang, and J. Dai, "Deformable DETR: Deformable Transformers for End-to-End Object Detection," in *International Conference on Learning Representations*, 2021. [Online]. Available: https://openreview.net/forum?id=gZ9hCDWe6ke

[206] T. Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollar, "Focal Loss for dense object detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 42, no. 2, pp. 318–327, Feb. 2020.

[207] J. Martinez, R. Hossain, J. Romero, and J. J. Little, "A simple yet effective baseline for 3D human pose estimation," in *IEEE International Conference on Computer Vision (ICCV)*, Oct. 2017, pp. 2659–2668.

[208] C. Ionescu, D. Papava, V. Olaru, and C. Sminchisescu, "Human3.6M: Large Scale Datasets and Predictive Methods for 3D Human Sensing in Natural Environments," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 36, no. 7, pp. 1325–1339, jul 2014.

[209] S. Blandino, T. Ropitault, C. R. C. M. da Silva, A. Sahoo, and N. Golmie, "IEEE 802.11bf DMG Sensing: Enabling High-Resolution mmWave Wi-Fi Sensing," *IEEE Open Journal of Vehicular Technology*, vol. 4, pp. 342–355, 2023.

[210] C. Chen, H. Song, Q. Li, F. Meneghello, F. Restuccia, and C. Cordeiro, "Wi-Fi Sensing Based on IEEE 802.11bf," *IEEE Communications Magazine*, vol. 61, no. 1, pp. 121–127, Jan. 2023.

[211] R. Du, H. Hua, H. Xie, X. Song, Z. Lyu, M. Hu, Narengerile, Y. Xin, S. McCann, M. Montemurro, T. X. Han, and J. Xu, "An Overview on IEEE 802.11bf: WLAN Sensing," *IEEE Communications Surveys and Tutorials*, vol. 27, no. 1, pp. 184–217, Feb. 2025.

[212] M. Youssef and A. Agrawala, "The Horus Location Determination System," *Wireless Networks*, vol. 14, no. 3, pp. 357–374, Jun. 2008.

[213] H. Chen, Y. Zhang, W. Li, X. Tao, and P. Zhang, "ConFi: Convolutional Neural Networks Based Indoor Wi-Fi Localization Using Channel State Information," *IEEE Access*, vol. 5, pp. 18 066–18 074, 2017.

[214] J. Ding and Y. Wang, "WiFi CSI-Based Human Activity Recognition Using Deep Recurrent Neural Network," *IEEE Access*, vol. 7, pp. 174 257–174 269, 2019.

[215] J. Yu, P. Wang, T. Koike-Akino, Y. Wang, P. V. Orlik, and R. M. Buehrer, "Multi-Band Wi-Fi Sensing With Matched Feature Granularity," *IEEE Internet of Things Journal*, vol. 9, no. 23, pp. 23 810–23 825, Dec. 2022.

[216] J. Yu, P. Wang, T. Koike-Akino, and P. V. Orlik, "Multi-Modal Recurrent Fusion for Indoor Localization," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 5083–5087.

[217] A. U. Ahmed, R. Arablouei, F. de Hoog, B. Kusy, and R. Jurdak, "Multi-radio Data Fusion for Indoor Localization using Bluetooth and WiFi," in *International Conference on Pervasive and Embedded Computing and Communication Systems (PECCS)*, 2019, pp. 13–24.

[218] X. Dang, X. Si, Z. Hao, and Y. Huang, "A Novel Passive Indoor Localization Method by Fusion CSI Amplitude and Phase Information," *Sensors*, vol. 19, no. 4, Feb. 2019.

[219] J. Xiong, K. Sundaresan, and K. Jamieson, "ToneTrack: Leveraging Frequency-Agile Radios for Time-Based Indoor Wireless Localization," in *International Conference on Mobile Computing and Networking (MobiCom)*, 2015, pp. 537–549.

[220] H. Zhu, Y. Zhuo, Q. Liu, and S. Chang, "π-splicer: Perceiving accurate CSI phases with commodity WiFi devices," *IEEE Transactions on Mobile Computing*, vol. 17, no. 9, pp. 2155–2165, Sep. 2018.

[221] N. Maletic, V. Sark, J. Gutiérrez, and E. Grass, "Device Localization Using mmWave Ranging with Sub-6-Assisted Angle of Arrival Estimation," in *IEEE International Symposium on Broadband Multimedia Systems and Broadcasting (BMSB)*, 2018, pp. 1–6.

[222] J. Hu, D. Niyato, and J. Luo, "Cross-Domain Learning Framework for Tracking Users in RIS-Aided Multi-Band ISAC Systems With Sparse Labeled Data," *IEEE Journal on Selected Areas in Communications*, vol. 42, no. 10, pp. 2754–2768, 2024.

[223] R. T. Q. Chen, Y. Rubanova, J. Bettencourt, and D. Duvenaud, "Neural ordinary differential equations," *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 6572–6583, 2018.

[224] Y. Rubanova, R. T. Q. Chen, and D. Duvenaud, "Latent ODEs for irregularly-sampled time series," *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 5320–5330, 2019.

[225] M. Kotaru, K. Joshi, D. Bharadia, and S. Katti, "SpotFi: Decimeter Level Localization Using WiFi," in *ACM Conference on Special Interest Group on Data Communication (SIGCOMM)*, Aug. 2015, pp. 269–282.

[226] D. Zhang, D. Wu, K. Niu, X. Wang, F. Zhang, J. Yao, D. Jiang, and F. Qin, "Practical Issues and Challenges in CSI-based Integrated Sensing and Communication," in *IEEE International Conference on Communications Workshops*, May 2022, pp. 836–841.

[227] D. Zhang, Y. Hu, Y. Chen, and B. Zeng, "Calibrating Phase Offsets for Commodity WiFi," *IEEE Systems Journal*, vol. 14, no. 1, pp. 661–664, Mar. 2020.

[228] H. Sun, X. Zhu, Y. Liu, and W. Liu, "WiFi Based Fingerprinting Positioning Based on Seq2seq Model," *Sensors*, vol. 20, no. 13, Jul. 2020.

[229] J. Yu, H. M. Saad, and R. Michael Buehrer, "Centimeter-Level Indoor Localization using Channel State Information with Recurrent Neural Networks," in *IEEE/ION Position, Location and Navigation Symposium (PLANS)*, 2020, pp. 1317–1323.

[230] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, Nov. 1997.

[231] K. Cho, B. van Merrienboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation," *arXiv preprint arXiv:1406.1078*, Jun. 2014.

[232] J. Kelly, J. Bettencourt, M. J. Johnson, and D. Duvenaud, "Learning differential equations that are easy to solve," in *Advances in Neural Information Processing Systems (NeurIPS*, 2020.

[233] C. Finlay, J.-H. Jacobsen, L. Nurbekyan, and A. M. Oberman, "How to train your neural ODE: the world of Jacobian and Kinetic regularization," in *International Conference on Machine Learning (ICML)*, vol. 119, no. 296, Jul. 2020, pp. 3154–3164.

[234] A. Zhu, P. Jin, B. Zhu, and Y. Tang, "On numerical integration in neural ordinary differential equations," in *International Conference on Machine Learning (ICML)*, 2022, pp. 27 527–27 547.

[235] H. H. N. Nguyen, T. Nguyen, H. Vo, S. Osher, and T. Vo, "Improving neural ordinary differential equations with Nesterov's accelerated gradient method," in *Advances in Neural Information Processing Systems (NeurIPS)*, S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, Eds., vol. 35, 2022, pp. 7712–7726.

[236] D. P. Kingma and M. Welling, "Auto-Encoding Variational Bayes," *arXiv preprint arXiv:1312.6114*, Dec. 2013.

[237] D. Steinmetzer, D. Wegemer, and M. Hollick, "Talon Tools: The Framework for Practical IEEE 802.11ad Research," in *Available: https://seemoo.de/talon-tools/*, 2018.

[238] G. Bielsa, J. Palacios, A. Loch, D. Steinmetzer, P. Casari, and J. Widmer, "Indoor Localization Using Commercial Off-The-Shelf 60 GHz Access Points," in *IEEE International Conference on Computer Communications (INFOCOM)*, Apr. 2018, pp. 2384–2392.

[239] T. Akiba, S. Sano, T. Yanase, T. Ohta, and M. Koyama, "Optuna: A Next-generation Hyperparameter Optimization Framework," in *ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2019, pp. 2623–2631.

[240] L. N. Smith and N. Topin, "Super-convergence: Very fast training of neural networks using large learning rates," in *Artificial intelligence and machine learning for multi-domain operations applications*, vol. 11006, 2019, pp. 369–386.

[241] Y. Hu, X. Yang, Z. Xia, and F. Xu, "Human activity recognition trained on simulated millimeter-wave radar data with domain adaptation," *IEEE Transactions on Instrumentation and Measurement*, vol. 74, pp. 1–13, 2025.

[242] H. Yan, X. Zhang, J. Huang, Y. Feng, M. Li, A. Wang, W. Ou, H. Wang, and Z. Liu, "Wi-SFDAGR: WiFi-based cross-domain gesture recognition via source-free domain adaptation," *IEEE Internet of Things Journal*, vol. 12, no. 13, pp. 24 159–24 173, Jul. 2025.

[243] W. Jiao, W. Du, C. Zhang, and L. Suo, "A high-efficient WI-Fi-based cross-domain recognition framework using multisource domain adaptation for single-transceiver scenarios," *IEEE Sensors Journal*, vol. 25, pp. 14 196–14 208, Apr. 2025.

[244] E. Pinyoanuntapong, A. Ali, K. Jakkala, P. Wang, M. Lee, Q. Peng, C. Chen, and Z. Sun, "GaitSADA: Self-Aligned Domain Adaptation for mmWave Gait Recognition," in *IEEE International Conference on Mobile Ad Hoc and Smart Systems (MASS)*, 2023, pp. 218–226.

[245] B.-B. Zhang, D. Zhang, Y. Li, Y. Hu, and Y. Chen, "Unsupervised domain adaptation for RF-based gesture recognition," *IEEE Internet of Things Journal*, vol. 10, no. 23, pp. 21 026–21 038, Dec. 2023.

[246] H. Li, X. Chen, J. Wang, D. Wu, and X. Liu, "DAFI: WiFi-based Device-free Indoor Localization via Domain Adaptation," *ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies (IMWUT)*, vol. 5, no. 4, pp. 1–21, Dec. 2022.

[247] X. Chen, H. Li, C. Zhou, X. Liu, D. Wu, and G. Dudek, "Fidora: Robust WiFi-Based Indoor Localization via Unsupervised Domain Adaptation," *IEEE Internet of Things Journal*, vol. 9, no. 12, pp. 9872–9888, Jun. 2022.

[248] P. Dai, J. Zhou, J. Ma, H. Zhang, and X. Wu, "Meta-transfer learning-based cross-domain gesture recognition using WiFi channel state information," *IEEE Transactions on Consumer Electronics*, vol. 71, no. 2, pp. 2530–2543, May 2025.

[249] J. Cha, H. Kim, and J. Choi, "Meta-learning-based people counting and localization models employing CSI from commodity WiFi NICs," *arXiv preprint arXiv:2502.03117*, Feb. 2025.

[250] B. Han, L. Wang, X. Lu, J. Meng, and Z. Zhou, "Cross-modal meta-learning for WiFi-based human activity recognition," in *International Conference on Mobile Computing and Networking (MobiCom)*, Oct. 2023, pp. 1–3.

[251] W. Wei, J. Yan, X. Wu, C. Wang, and G. Zhang, "A meta-learning approach for device-free indoor localization," *IEEE Communications Letters*, vol. 27, no. 3, pp. 846–850, Mar. 2023.

[252] X. Shen, H. Zheng, X. Feng, and J.-J. Hu, "ML-HGR-net: A meta-learning network for FMCW radar based hand gesture recognition," *IEEE Sensors Journal*, vol. 22, pp. 10 808–10 817, Jun. 2022.

[253] J. Zhang, Z. Chen, C. Luo, B. Wei, S. S. Kanhere, and J. Li, "MetaGanFi: Cross-Domain Unseen Individual Identification Using WiFi Signals," *ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies (IMWUT)*, vol. 6, no. 3, pp. 1–21, Sep. 2022.

[254] Z. Fan, H. Zheng, and X. Feng, "A meta-learning-based approach for hand gesture recognition using FMCW radar," in *International Conference on Wireless Communications and Signal Processing (WCSP)*, Oct. 2020, pp. 522–527.

[255] Y. Ye, T. Cui, S. Guo, and G. Cui, "An end-to-end framework for few-shot millimeter-wave radar-based hand gesture recognition," in *International Joint Conference on Neural Networks (IJCNN)*, vol. 27, Jun. 2024, pp. 1–8.

[256] S. Bi, X. Chen, L. Zheng, H. Hou, and X. Lin, "RoGER: Few-shot learning based robust gesture recognition with multi-modal WI-fi CSI measurements," in *IEEE International Conference on Machine Learning for Communication and Networking (ICMLCN)*, May 2024, pp. 517–522.

[257] M. Han, H. Yang, T. Ni, D. Duan, M. Ruan, Y. Chen, J. Zhang, and W. Xu, "mmSign: mmWave-based few-shot online handwritten signature verification," *ACM Transactions on Sensor Networks*, vol. 20, no. 4, pp. 1–31, Jul. 2024.

[258] Q. Cao, H. Xue, T. Liu, X. Wang, H. Wang, X. Zhang, and L. Su, "mmCLIP: Boosting mmWave-based zero-shot HAR via signal-text alignment," in *ACM Conference on Embedded Networked Sensor Systems (SenSys)*, Nov. 2024, pp. 184–197.

[259] B. Yang, H. Wang, L. Hu, H. Zhu, C.-T. Lam, and K. Fang, "Few-shot cross-domain-based WiFi sensing system for online learning in IoT," *IEEE Sensors Journal*, vol. 23, no. 23, pp. 29 892–29 905, Dec. 2023.

[260] S. Gong, H. Shi, X. Yan, Y. Fang, A. Paul, Z. Wu, and W. Long, "Human activity recognition with FMCW radar using few-shot learning," *IEEE Sensors Journal*, vol. 23, no. 15, pp. 17 815–17 824, Aug. 2023.

[261] C. Feng, N. Wang, Y. Jiang, X. Zheng, K. Li, Z. Wang, and X. Chen, "Wi-learner: Towards one-shot learning for cross-domain WI-fi based gesture recognition," *ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies (IMWUT)*, vol. 6, no. 3, pp. 1–27, Sep. 2022.

[262] Y. Zhang, Y. Chen, Y. Wang, Q. Liu, and A. Cheng, "CSI-based human activity recognition with graph few-shot learning," *IEEE Internet of Things Journal*, vol. 9, no. 6, pp. 4139–4151, 2021.

[263] H. Zhang, D. Zhang, R. Song, Z. Wu, J. Chen, L. Fang, Z. Lu, Y. Hu, H. Lin, and Y. Chen, "UMIMO: Universal unsupervised learning for mmwave radar sensing with MIMO array synthesis," *IEEE Transactions on Mobile Computing*, vol. 24, no. 8, pp. 7042–7058, Aug. 2025.

[264] K. Xu, J. Wang, H. Zhu, and D. Zheng, "Evaluating Self-Supervised Learning for WiFi CSI-based Human Activity Recognition," *ACM Transactions on Sensor Networks*, vol. 21, no. 2, pp. 1–38, Mar. 2025.

[265] M. M. Rahman and S. Z. Gurbuz, "Self-supervised contrastive learning for radar-based human activity recognition," in *IEEE Radar Conference (RadarConf)*, May 2023, pp. 1–6.

[266] J. Yang, X. Chen, H. Zou, D. Wang, and L. Xie, "AutoFi: Toward automatic WI-fi human sensing via geometric self-supervised learning," *IEEE Internet of Things Journal*, vol. 10, no. 8, pp. 7416–7425, Apr. 2023.

[267] K. Xu, J. Wang, L. Zhang, H. Zhu, and D. Zheng, "Dual-stream contrastive learning for channel state information based human activity recognition," *IEEE Journal of Biomedical and Health Informatics*, vol. 27, no. 1, pp. 329–338, Jan. 2023.

[268] H. Abedi, P. P. Morita, J. Boger, A. Wong, and G. Shaker, "Unsupervised learning for hallway gait analysis using FMCW radar," in *IEEE International Symposium on Antennas and Propagation and USNC-URSI Radio Science Meeting (AP-S/URSI)*, Jul. 2022, pp. 836–837.

[269] M. J. Bocus, H. S. Lau, R. McConville, R. J. Piechocki, and R. Santos Rodriguez, "Self-supervised WiFi-based activity recognition," in *IEEE Globecom Workshops*, Dec. 2022, pp. 552–557.

[270] Y. Liu, M. Ahmadi, J. Fuchs, M. Alaee-Kerahroodi, and M. R. Bhavani Shankar, "Dynamic Indoor mmWave MIMO Radar Simulation: An Image Rendering-Based Approach," *IEEE Transactions on Antennas and Propagation*, vol. 73, no. 4, pp. 1984–1999, 2025.

[271] T. Huang, H. Ding, W. Sun, C. Zhao, G. Wang, F. Wang, K. Zhao, Z. Wang, and W. Xi, "One Snapshot is All You Need: A Generalized Method for mmWave Signal Generation," in *IEEE International Conference on Computer Communications (INFOCOM)*, 2025, pp. 1–10.

[272] X. Chen and X. Zhang, "RF genesis: Zero-shot generalization of mmWave sensing through simulation-based data synthesis and generative diffusion models," in *ACM Conference on Embedded Networked Sensor Systems (SenSys)*, Nov. 2023, pp. 28–42.

[273] X. Zhang, Z. Li, and J. Zhang, "Synthesized millimeter-waves for human motion sensing," in *Proceedings of the ACM Conference on Embedded Networked Sensor Systems (SenSys)*, Nov. 2022, pp. 377–390.

[274] J. Yi, H. Zou, M. Ling, J. Gao, and Y. L. Murphey, "M2FM: A Multi Modal Fusion Model for Human Action Recognition With Camera and Millimeter-Wave Radar," *IEEE Internet of Things Journal*, 2025.

[275] A. Chen, X. Wang, K. Shi, Y. Huo, J. Chen, and Q. Ye, "Toward Weather-Robust 3D Human Body Reconstruction: Millimeter-Wave Radar-Based Dataset, Benchmark, and Multi-Modal Fusion," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 35, no. 1, pp. 273–286, 2025.

[276] T. Fukushima, T. Matsukawa, Y. Ikemura, H. Terasoma, S. Kato, T. Fujihashi, T. Murakami, Y. Asai, and S. Saruwatari, "Wi-Fi Meets Sound: A Deep Fusion Leap Beyond Single-Modality for Device Localization," *IEEE Access*, vol. 13, pp. 170 291–170 306, 2025.

[277] M. Wei, D. Zhao, L. Zhang, C. Wang, Y. Zhang, Q. Wang, X. Fan, Y. Zhong, and S. Mao, "Wi-fitness: Improving WI-fi sensing with video perception for smart fitness," *IEEE Internet of Things Journal*, vol. 12, no. 3, pp. 2876–2889, Feb. 2025.

[278] L. Hao, B. Huang, B. Jia, and G. Mao, "Heterogeneous dual-attentional network for WiFi and video-fused multi-modal crowd counting," *IEEE Transactions on Mobile Computing*, vol. 23, no. 12, pp. 14 233–14 247, Dec. 2024.

[279] C. Mao, C. Tan, J. Hu, and M. Zheng, "ViFi-ReID: A Two-Stream Vision-WiFi Multimodal Approach for Person Re-Identification," *arXiv preprint arXiv:2410.09875*, Oct. 2024.

[280] K. Cai, Q. Xia, P. Li, J. Stankovic, and C. X. Lu, "Robust Human Detection under Visual Degradation via Thermal and mmWave Radar Fusion," *arXiv preprint arXiv:2307.03623*, Jul. 2023.

[281] A. Chen, X. Wang, K. Shi, S. Zhu, B. Fang, Y. Chen, J. Chen, Y. Huo, and Q. Ye, "Imm-Fusion: Robust mmWave-RGB Fusion for 3D Human Body Reconstruction in All Weather Conditions," in *IEEE International Conference on Robotics and Automation (ICRA)*, 2023, pp. 2752–2758.

[282] Y. Tian, Y. Wang, R. Zheng, X. Liu, X. Tong, and K. Li, "WSTrack: A Wi-Fi and Sound Fusion System for Device-free Human Tracking," in *IEEE International Conference on Computer Communications (INFOCOM)*, May 2023, pp. 1–10.

[283] Y. He, M. Xu, F. Xiao, and J. Luo, "VersaBeam: Versatile beamforming for integrated sensing and communication over commodity WI-fi," in *IEEE International Conference on Computer Communications (INFOCOM)*, May 2025, pp. 1–10.

[284] Y. He, M. Xu, Z. Chen, F. Xiao, and J. Luo, "Beam-Fi: Integrated sensing and communication via MU-MIMO upon commodity Wi-Fi," *ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies (IMWUT)*, vol. 9, no. 3, pp. 1–22, Sep. 2025.

[285] Z. Chen, C. Hu, T. Zheng, H. Cao, Y. Yang, Y. Chu, H. Jiang, and J. Luo, "ISAC-Fi: Enabling Full-Fledged Monostatic Sensing Over Wi-Fi Communication," *IEEE Journal on Selected Areas in Sensors*, vol. 1, pp. 139–153, 2024.

[286] J. Pegoraro, J. O. Lacruz, F. Meneghello, E. Bashirov, M. Rossi, and J. Widmer, "RAPID: Retrofitting IEEE 802.11ay access points for indoor human detection and sensing," *IEEE Transactions on Mobile Computing*, vol. 23, no. 5, pp. 1–18, 2024.

[287] Y. He, J. Liu, M. Li, G. Yu, J. Han, and K. Ren, "SenCom: Integrated Sensing and Communication with Practical WiFi," in *International Conference on Mobile Computing and Networking (MobiCom)*, 2023.

[288] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, "Learning Transferable Visual Models From Natural Language Supervision," *arXiv preprint arXiv:2103.00020*, 2021.

[289] L. H. Li, P. Zhang, H. Zhang, J. Yang, C. Li, Y. Zhong, L. Wang, L. Yuan, L. Zhang, J.-N. Hwang, K.-W. Chang, and J. Gao, "Grounded Language-Image Pre-training," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2022, pp. 10 955–10 965.

[290] S. Liu, Z. Zeng, T. Ren, F. Li, H. Zhang, J. Yang, Q. Jiang, C. Li, J. Yang, H. Su, J. Zhu, and L. Zhang, "Grounding DINO: Marrying DINO with Grounded Pre-training for Open-Set Object Detection," in *European Conference on Computer Vision (ECCV)*, 2024, pp. 38–55.

[291] R. Guan, R. Zhang, N. Ouyang, J. Liu, K. L. Man, X. Cai, M. Xu, J. Smith, E. G. Lim, Y. Yue, and H. Xiong, "Talk2Radar: Bridging Natural Language with 4D mmWave Radar for 3D Referring Expression Comprehension," *arXiv preprint arXiv:2405.12821*, 2025.

[292] Z. Yang, Z. Huang, G. Wang, H. Ding, T. X. Han, and F. Wang, "Talk is not always cheap: Promoting wireless sensing models with text prompts," *arXiv preprint arXiv:2504.14621*, Apr. 2025.

[293] Y. Ren, H. Zhang, H. Yuan, J. Zhang, and Y. Shen, "Wi-chat: Large language model-powered WI-fi-based human activity recognition," in *Proceedings of the International Workshop on Environmental Sensing Systems for Smart Cities*, Jun. 2025, pp. 8–13.

[294] F.-Q. Cui, Y.-T. Guo, T. Zheng, and J. Huang, "Generalizing WiFi Gesture Recognition via Large-Model-Aware Semantic Distillation and Alignment," *arXiv preprint arXiv:2510.13390*, 2025.

[295] Z. Li, X. Luo, X. Ge, L. Zhou, X. Lin, and Y. Liu, "MMSense: Adapting vision-based foundation model for multi-task multi-modal wireless sensing," *arXiv preprint arXiv:2511.12305*, Nov. 2025.

[296] C. Li, M. Xu, Y. Du, L. Liu, C. Shi, Y. Wang, H. Liu, and Y. Chen, "Practical Adversarial Attack on WiFi Sensing Through Unnoticeable Communication Packet Perturbation," in *International Conference on Mobile Computing and Networking (MobiCom)*, 2024, pp. 373–387.

[297] L. Xu, X. Zheng, Y. Zhang, L. Liu, and H. Ma, "WiCAM2.0: Imperceptible and targeted attack on deep learning based WiFi sensing," *ACM Transactions on Sensor Networks*, vol. 20, no. 6, pp. 1–22, Nov. 2024.

[298] Y. Xie, R. Jiang, X. Guo, Y. Wang, J. Cheng, and Y. Chen, "Universal Targeted Adversarial Attacks Against mmWave-based Human Activity Recognition," in *IEEE International Conference on Computer Communications (INFOCOM)*, 2023, pp. 1–10.

[299] F. Xiao, Y. Huang, Y. Zuo, W. Kuang, and W. Wang, "Over-the-air adversarial attacks on deep learning WI-fi fingerprinting," *IEEE Internet of Things Journal*, vol. 10, no. 11, pp. 9823–9835, Jun. 2023.

[300] J. Liu, Y. He, C. Xiao, J. Han, L. Cheng, and K. Ren, "Physical-world attack towards WiFi-based behavior recognition," in *IEEE International Conference on Computer Communications (INFOCOM)*, May 2022, pp. 400–409.